Design and Analysis of Studies Based on Hierarchical Composite Endpoints: Insights from the DARE-19 Trial

Samvel B. Gasparyan1  ·  Joan Buenconsejo2  ·  Elaine K. Kowalewski3  ·  Jan Oscarsson1  ·  Olof F. Bengtsson1  ·  Russell Esterline2  ·  Gary G. Koch3  ·  Otavio Berwanger4  ·  Mikhail N. Kosiborod5

Received: 28 January 2022 / Accepted: 19 May 2022 / Published online: 14 June 2022
© The Drug Information Association, Inc 2022

Abstract
Background/Aim DARE-19 (NCT04350593) was a randomized trial studying the effects of dapagliflozin, an SGLT2 inhibitor, in hospitalized patients with COVID-19 pneumonia and cardiometabolic risk factors. The conduct of DARE-19 offered the opportunity to define an innovative and clinically meaningful endpoint in a new disease that would best reflect the known profile of dapagliflozin, accompanied by the statistical challenges of analysis and interpretation of such a novel endpoint.

Methods Hierarchical composite endpoints (HCEs) are based on clinical outcomes which, unlike traditional composite endpoints incorporate ranking of components according to clinical importance. Design of an HCE requires the clinical considerations specific to the therapeutic area under study and the mechanism of action of the investigational treatment. Statistical aspects for the clinical endpoints include the proper definition of the estimand as suggested by ICH E9(R1) for the precise specification of the treatment effect measured by an HCE.

Results We describe the estimand of the DARE-19 trial, where an HCE was constructed to capture the treatment effect of dapagliflozin in hospitalized patients with COVID-19, and was analyzed using a win odds. Practical aspects of designing new studies based on an HCE are described. These include sample size, power, and minimal detectable effect calculations for an HCE based on the win odds analysis, as well as handling of missing data and the clinical interpretability of the win odds in relation to the estimand.

Conclusions HCEs are flexible endpoints that can be adapted for use in different therapeutic areas, with win odds as the analysis method. DARE-19 is an example of a COVID-19 trial with an HCE as one of the primary endpoints for estimating a clinically interpretable treatment effect in the COVID-19 setting.

Keywords COVID-19  ·  Dapagliflozin  ·  Win odds  ·  SGLT2 inhibitors  ·  Hierarchical composite endpoints

Introduction

One of the challenges in designing clinical trials that can inform clinical practice is the selection of a meaningful primary efficacy endpoint. The ideal endpoint should be modifiable by the intervention, clinically important, and relevant in the management of the disease in the target population. Clinical endpoints are generally focused on individual components of the syndrome assessed at a specific timepoint, all of which have different impacts on a patient’s status. However, in practice, the totality of various components of the syndrome over the entire treatment period represents a patient’s true disease burden. Therefore, the use of composite endpoints which combine various components of the syndrome are potentially better positioned to capture the whole disease burden of patients. This is especially important
when studying acutely ill and hospitalized patients. Therefore, in assessing the total disease burden, a comprehensive approach is to combine both favorable events (i.e., improvements in clinical status such as a shorter hospital stay) and unfavorable events (i.e., deterioration in clinical status such as the occurrence of major clinical events), observed during the treatment period.

The capture of both improvement and deterioration expands the scope of information for the analysis and thereby creates a more sensitive endpoint to detect the benefit (and possibly the overall benefit-risk) of an intervention. This approach may improve the power and lead to smaller and more feasible studies, and therefore gain in efficiency in trial design. Traditional composite endpoints, unlike an ordinal composite endpoint, cannot distinguish favorable outcomes from unfavorable ones. Hierarchical composite endpoints (HCE), which combine events of different clinical importance into an ordinal outcome and prioritize the most severe event, are meant to address this problem [1, 2]. For a given patient, an HCE takes all events experienced during the follow-up into consideration and gives higher priority to the most severe event. Therefore, patients experiencing deterioration either followed by or preceded by an improvement will be assigned to the “unfavorable” outcome in the analysis.

In addition to defining the endpoint, clinical trial guidelines highlight the importance of defining the estimand—i.e., a precise definition of the treatment effect that is to be estimated, which corresponds to the clinical question of interest [3]. For constructing the estimand, assessment should be made for intercurrent events, defined as events occurring after treatment initiation that affect either the measurement of the endpoint or interpretation of the effect of treatment on the endpoint (e.g., discontinuation of assigned treatment, use of additional or alternative treatment, or death); and any plans on how to address them should be specified. In many settings, the occurrence of a major event during the conduct of the trial may represent a meaningful degree of clinical deterioration that cannot be ignored in any analysis of efficacy. These intercurrent events, for example, can be handled by including them in the definition of a hierarchical composite endpoint by assigning the appropriate rank. Hence, HCEs have the flexibility to incorporate potential intercurrent events.

COVID-19 is the disease caused by the novel coronavirus SARS-CoV-2. Severe COVID-19 is characterized by hospitalization with pneumonia with risk for respiratory failure, heart and kidney complications. The therapeutic goal of a COVID-19 treatment needs to be curative; whereas for chronic diseases, such as heart failure, the therapeutic goals are to manage the disease and to prevent worsening. Therefore, an important outcome of a COVID-19 treatment is recovery as defined for a fixed period of time, e.g., 30 days. Recovery in hospitalized patients, in the simplest form, is the outcome of discharge from the hospital. However, prevention of worsening of disease and mortality are also important goals of treatment, and so inclusion of outcomes for both improvement and deterioration (which includes death) in patients’ clinical status will produce a clinically meaningful endpoint in patients hospitalized for COVID-19.

DARE-19 was a randomized clinical trial in hospitalized COVID-19 patients to evaluate the effects of dapagliflozin, an SGLT2 inhibitor which had been previously shown to have cardiorenal organ protective benefits in ambulatory patients with cardiometabolic disease, but has no known antiviral properties [4, 5]. Clinical challenges related to the trial were the definition of an innovative and clinically meaningful endpoint in a new disease that would match the known cardiorenal protective profile of dapagliflozin, together with the statistical challenges of analysis and interpretation of such novel endpoints.

In this study, an HCE of recovery was constructed to combine both deterioration in clinical status compared to baseline and recovery (improvement in clinical status compared to baseline) into a single metric. This HCE was analyzed with the win odds (win ratio with ties) [6–9]; and it was one of the prespecified dual primary endpoints for which the trial was powered to detect a clinically meaningful treatment effect.

We describe the practical aspects of designing new studies based on an HCE. These include sample size, power, and minimal detectable effect calculations for an HCE based on the win odds analysis, as well as handling of missing data and the clinical interpretability of the win odds in relation to the estimand. As an example, we provide the case study of a randomized controlled trial DARE-19, which had an HCE as an endpoint, with the win odds as its main analysis method. We describe also the estimand of DARE-19 associated with this HCE.

Methods

DARE-19

The DARE-19 trial methods and main results were published previously. Briefly, DARE-19 (NCT04350593) was an international, multicenter, randomized, double-blind, placebo-controlled trial to evaluate the effects of treatment with dapagliflozin for 30 days in hospitalized patients with COVID-19 with respiratory failure, and at least one cardiometabolic risk factor: hypertension, type 2 diabetes, atherosclerotic cardiovascular disease, heart failure, or chronic kidney disease. The trial randomized 1250 patients across 95 centers in seven countries. The trial had dual primary endpoints of time to organ dysfunction or death.
and the novel hierarchical composite endpoint including both recovery and deterioration. The endpoints were tested in parallel [4, 5]. The objective of DARE-19 was to investigate if treatment with dapagliflozin could prevent worsening by reducing complications, as well as increase the number of patients that recover/leave the hospital without complications.

HCE Definition

In the COVID-19 setting, during the start of the pandemic, WHO suggested a comprehensive endpoint (see Fig. 1) that accounted for multiple clinical states (including death and cure or recovery) [10]. Although there are several versions of the WHO suggested ordinal scale [11], all of them have the following important characteristics:

1. They always include death and hospital discharge as negative and positive outcomes (usually the ordinal outcomes consist of a 7- or 8-point scale for the vital status, the amount of oxygen support needed, hospital discharge or limitation of physical activities if discharged).
2. They are assessed at a fixed timepoint and, therefore, do not include events that occur between the baseline and the prespecified timepoint (for example, Day 15 or Day 30).
3. They do not prioritize the most severe outcome, but rather use the latest observed outcome in the analysis (except death, if it occurred before the timepoint).

The WHO ordinal endpoints are the simplest types of HCE, but they do not account for the most severe event for the patient during the entire follow-up period if it is not death. Accordingly, these endpoints are designed to capture the treatment effect of drugs that have antiviral effects and are thus expected to reduce the oxygen support of patients, and the time to hospital discharge. In contrast, DARE-19 was studying dapagliflozin, an SGLT2 inhibitor, that has proven benefits on cardio-renal organ protection in the chronic setting [12], but does not have known anti-viral effects. Therefore, an HCE was tailored to include outcomes that are relevant in the COVID-19 setting (oxygen support and hospital discharge) with outcomes that are COVID-19 complications (cardio-renal outcomes and death), to investigate if dapagliflozin has an organ protective effect in the acute illness setting. Hospital discharge, as an intercurrent event (since the ranking only accounts for organ dysfunction events during the index hospitalization), is managed with the composite strategy through its inclusion in the definition of the endpoint.

The key difference between the HCE used in DARE-19 and the ordinal scale endpoint suggested by WHO is that the HCE, in contrast to the WHO endpoint, accounts for in-hospital worsening (occurrence of organ dysfunction events) of COVID-19 and death after hospital discharge. Therefore, the HCE for DARE-19 is an endpoint with a stricter definition of recovery by capturing the whole disease burden of patients hospitalized for COVID-19 (both improvement and worsening events). Improvement in clinical status compared to baseline (recovery) is represented as

1. Discharge from hospital before or at Day 30 without in-hospital worsening and alive at Day 30; or
2. Still in hospital at Day 30, but without in-hospital worsening during the 30 days of hospitalization and without oxygen support.

Deterioration in clinical status compared to baseline is represented as occurrence of cardio-renal-metabolic organ dysfunction events during the index hospitalization, prolonged hospitalization, or death.

Additionally, the suggested ranking (Table 1) provides a patient-level ranking through the categorization of all patients into one and only one category. In addition, the timing of events (hospitalization, worsening events and hospital discharge) was used to rank patients in each category and thus increase the power. As noted

| Patient State       | Descriptor                                      | Score |
|---------------------|------------------------------------------------|-------|
| Uninfected          | No clinical or virological evidence of infection | 0     |
| Ambulatory          | No limitation of activities                     | 1     |
|                     | Limitation of activities                        | 2     |
| Hospitalized        | Hospitalized, no oxygen therapy                 | 3     |
| Mild disease        | Oxygen by mask or nasal prongs                  | 4     |
| Hospitalized        | Non-invasive ventilation or high-flow oxygen     | 5     |
| Severe Disease      | Intubation and mechanical ventilation           | 6     |
|                     | Ventilation + additional organ support – pressors, RRT, ECMO | 7     |
| Dead                | Death                                           | 8     |
in [13], the power of statistical tests can be improved if the timing of the events is incorporated in the analysis as well. Therefore, this HCE is essentially designed to capture any increase in the number of patients in the active group compared to placebo who recover/leave the hospital without complications due to organ dysfunction events, as well as reduction in the time to recovery. Prevention of organ dysfunction events is a known effect of dapagliflozin.

**Estimand**

The estimand was introduced by ICH [1] as a conceptual framework for the precise definition of the treatment effect as it applies to the clinical question posed by a given clinical trial objective.

The primary estimand in DARE-19 based on the HCE was the extent to which dapagliflozin (10 mg once daily plus standard of care) improves recovery in adults with cardio-metabolic-renal risk factors and hospitalized with severe respiratory failure due to COVID-19, irrespective of exposure, treatment discontinuation, or concomitant treatment.

Hence, the attributes of the estimand of the DARE-19 trial were the following: the population was the patients with cardio-metabolic-renal risk factors who were hospitalized with severe (but not critical) COVID-19 and treated with dapagliflozin 10 mg (using the intention-to-treat principle) administered once daily (for 30 days) in addition to standard of care; the corresponding endpoint was the hierarchical composite endpoint of recovery in Table 1. The population level summary was the win odds as discussed in the next section. The handling of the intercurrent events of initiation of concomitant treatment and study drug discontinuation is reflected in the estimand with the treatment policy strategy (hence these intercurrent events were disregarded).

### Design of DARE-19 and the Primary Analysis Method

Ordinal outcomes only require statistical methods that rely on the order of outcomes. This is different from assigning scores to various outcomes and analyzing them as a numerical endpoint (e.g. using mean and standard deviation, see for example [14]). An HCE can be analyzed with rank based methods such as the Mann–Whitney-Wilcoxon test, or with rank ANCOVA if stratification and covariates are present. In DARE-19, the primary analysis method was based on the win/Mann–Whitney odds approach, and it corresponds to a win ratio with ties included as half wins [4–6] (based on the initial idea of Finkelstein and Schoenfeld [15]).

### Win Odds

For the defined ranks for each endpoint, each patient in the intervention group is compared to every patient in the control group to produce a “win”, “loss” or “tie”, based on the derived ranks where a higher rank means a worse outcome. Then, the total number of wins plus half of the total number of ties are divided by the total number of such comparisons. The resulting ratio is called the win proportion of the intervention group against the control group, and it estimates the theoretical win probability of the intervention group having better outcomes than the placebo group. In addition, the concept of win probability for the binary and continuous outcomes has been described as “proportion in favor of treatment” in [16, 17]. Division of the win proportion of the intervention group by the win proportion of the control group produces the win odds (WO) for the intervention group. Division of the total number of wins by the total number of losses without taking into account the ties

### Table 1 The HCE in DARE-19

| I. | Patients alive at the end of follow up (Day 30), without any organ dysfunction event and are discharged from hospital before Day 30 will represent the best cohort [Ranking within this cohort will be based on the time to discharge, with patients being discharged later getting a higher rank] |
| II. | Patients without any organ dysfunction event but hospitalized at the end of follow-up (Day 30) [Ranking within this cohort, from low to high, includes patients not requiring supplemental oxygen, patients requiring supplemental oxygen, and patients on high-flow oxygen devices] |
| III. | Patients who did not die but have only one new or worsened organ dysfunction event [Ranking within this cohort will be based on the timing of the event, with patients having the event sooner getting a higher rank. Type of organ dysfunction will not be considered] |
| IV. | Patients who did not die but have more than one new or worsened organ dysfunction events [Ranking within this cohort will be based on the number of events, with higher number getting a higher rank] |
| V. | Patients dying during the study [Ranking within this cohort will be based on the timing of the event, with patients dying sooner getting a higher rank] |
produces what is called the win ratio as introduced in [6], with this being why the quantity with ties (the win odds) is sometimes called the win ratio with ties [7] (in the presence of ties, the win ratio can be misleading about the degree of similarity of two distributions which is not true for the win odds, see [8]).

**Interpretation of Win Odds**

The win odds is greater than 1 if the intervention group is more likely to experience the better outcomes of interest; conversely a win odds less than 1 represents a less favorable effect in the intervention group as compared to the control group.

Moreover, if the ranks are defined at the patient-level (that is the defined comparison is transitive†), then the interpretation of the win odds is applicable on the patient-level as well, and so the WO will be the odds that a randomly selected patient in the intervention group will have a better outcome than such a patient in the placebo group. In this regard, a better outcome is interpretable as a shift in the clinical scale. However, a definition of ranks that is not patient-level does not enable such patient-level interpretation of the win odds.

**Management of Missing Data**

Two types of missingness need management for the ranking algorithm described in Table 1:

(a) Events with missing date of occurrence (e.g., a patient who is known to have died with an unknown date of death).

(b) Incomplete follow-up of events (e.g., patients who are lost to follow-up or withdraw consent at any time prior to completion of the follow-up period for the study).

For all analyses in DARE-19 (primary and supplementary), events with a missing date of occurrence (type a) were managed by imputing the ranks of these patients with the median rank of the category for the event. For example, if it is known that the patient died, then this patient is in category V (Table 1). Hence, the median rank of the patients who died with known date of death is the imputed rank for the patients who are known to have died with unknown date of death. Missing follow-up of events (type b) were managed in the primary analysis by censoring as discussed in the next section.

**Primary Analysis Method.**

For the patient-level ranks described previously, Cox proportional hazards regression (stratified by country) was applied to estimate the WO for the primary analysis, with ties managed with the Efron method. The statistical test for this treatment comparison was the stratified log-rank test. The advantage of calculating the win odds estimate from Cox regression with the log-rank test for the primary hypothesis is the capability to address type (b) missingness with censoring. As a supplementary analysis, a direct win odds based on the comparison of ranks of the intervention group versus the placebo group was also performed, including only patients with complete follow-up. The confidence interval for the direct WO was constructed using the methods of U-statistics [6–9].

It should be noted that the direct win odds calculation can possibly include patients with premature study discontinuation in the analysis by classifying their comparison to another patient as “tied” when it cannot be classified as better or worse. This way of handling missing data can create non-transitive comparisons; but for DARE-19, it did not have an impact on the results or interpretation of the treatment effect because of the low number of study discontinuations. Another method of managing patients with premature study discontinuations is the use of multiple imputation methods by randomly assigning an imputed outcome after the discontinuation and before Day 30.

**Other Analysis Methods of HCEs**

Although this discussion emphasizes the analysis of an HCE with the win odds, there needs to be recognition that the analysis method for an HCE is separate from how it is defined. An HCE is an endpoint that can be analyzed by the win odds or by another method, for example, ordinal logistic regression. In this regard, the win odds is applicable to an endpoint without hierarchical structure (for example, skewed numeric data) as a version of the Wilcoxon rank-sum test. Therefore, the definition of an HCE in the protocol of a clinical trial is a separate consideration from the analysis method/estimator, and so the statistical analysis plan for such an endpoint can specify other analysis methods/estimators for it.

**Sample Size and Power Calculations**

Gasparayan et al. [9] provides sample size and power calculation formulas for designing new studies (see also [18] for a generalization). DARE-19 was sized to detect a WO of 1.23 at a two-sided significance level of 2.5% (where the overall two-sided significance level of 5% was split

---

† If patient A has a higher rank than patient B, and patient B has a higher rank than patient C, then transitivity implies that patient A has higher rank than patient C.
between dual primary endpoints) [5]. A sample size of 1200 provided 80% power, and the minimum detectable treatment effect was \( WO = 1.15 \).

## Results

### Primary Analysis

The DARE-19 trial did not achieve statistical significance for either of its dual primary endpoints [5]. The HCE had \( WO = 1.09 \) (95% CI 0.97–1.22; \( p = 0.14 \)) (Fig. 2, Table 2). Among the 1250 patients in the DARE-19 trial, 9 patients discontinued the study prematurely and hence could not be explicitly ranked. For the direct win odds with exclusion of the 9 patients with premature discontinuation (see Table 2), the \( WO \) was 1.07 (95% CI 0.94–1.21; \( p = 0.33 \)). Handling of missing data did not affect the results due to the low proportion of patients who prematurely discontinued the study. Overall 7.5% of the comparisons for pairs of patients were ties (47.8% were wins for the dapagliflozin group, 44.6% were wins for the placebo group). Since less than 1% of the patients did not complete the DARE-19 trial, the difference between the direct win odds (for only completers) and the win odds from Cox regression was very small.

### Interpretation of Win Odds

A win odds of 1.09 means that the odds of a randomly selected patient in the intervention group having a better outcome than a randomly selected patient in the placebo group is 1.09. Therefore, in this case, the number needed to treat is 15 as defined in [7]. If the study had been statistically significant, this means that 15 patients are needed to be treated with the intervention treatment rather than with placebo to have one patient with a clinically better outcome (shift to the left in the clinical scale, \( x \)-axis Fig. 2).

### Discussion

HCEs are flexible endpoints, and their construction is feasible in different disease areas and for drugs with different modes of actions. They provide clinically meaningful measures of a patient’s condition throughout the follow up rather than giving priority to the first event or only measurements that pertain to a specific timepoint. Importantly, they can be designed to capture the entire disease burden of the patients by accounting for both improvement and
| Variable                                                                 | Dapa 10 mg (N=625) | Placebo (N=625) | Difference between treatment groups |
|--------------------------------------------------------------------------|---------------------|-----------------|-----------------------------------|
|                                                                          | n (%)               | n (%)           | Win odds                          |
| The hierarchical composite endpoint of all-cause mortality, new/worsened organ dysfunction, in hospital clinical status at day 30 and discharge from hospital | 619 (99.0)          | 622 (99.5)      |                                   |
| Primary                                                                 |                     |                 | 1.09                              |
| Direct                                                                  |                     |                 | 1.07                              |
| Discharge from hospital                                                 | 544 (87.0)          | 532 (85.1)      |                                   |
| Days to discharge                                                       |                     |                 | 1.09 (0.97, 1.22)                 |
| (0, 3]                                                                  | 158 (25.3)          | 146 (23.4)      | 0.136                             |
| (3, 6]                                                                  | 198 (31.7)          | 212 (33.9)      |                                   |
| (6, 9]                                                                  | 109 (17.4)          | 104 (16.6)      |                                   |
| (9, 12]                                                                 | 40 (6.4)            | 36 (5.8)        |                                   |
| (12, 15]                                                                | 23 (3.7)            | 19 (3.0)        |                                   |
| (15, 30]                                                                | 16 (2.6)            | 15 (2.4)        |                                   |
| In hospital clinical status at day 30                                    | 5 (0.8)             | 3 (0.5)         |                                   |
| Hospitalized, not requiring supplemental oxygen                         | 3 (0.5)             | 0               |                                   |
| Hospitalized, requiring supplemental oxygen                             | 1 (0.2)             | 3 (0.5)         |                                   |
| Hospitalized, on high flow oxygen devices                               | 1 (0.2)             | 0               |                                   |
| New/worsened organ dysfunction                                            | 29 (4.6)            | 33 (5.3)        |                                   |
| 1 event†                                                                | 16 (2.6)            | 19 (3.0)        |                                   |
| Multi-organ dysfunction (≥ 2 events)                                     | 13 (2.1)            | 14 (2.2)        |                                   |
| 2 events                                                                | 7 (1.1)             | 2 (0.3)         |                                   |
| 3 events                                                                | 3 (0.5)             | 6 (1.0)         |                                   |
| 4 events                                                                | 2 (0.3)             | 4 (0.6)         |                                   |
| 5 events                                                                | 0                   | 2 (0.3)         |                                   |
| 6 events                                                                | 1 (0.2)             | 0               |                                   |
| 7 events                                                                | 0                   | 0               |                                   |
| 8 events                                                                | 0                   | 0               |                                   |
| All-cause mortality                                                      | 41 (6.6)            | 54 (8.6)        |                                   |
| Subject censored                                                        | 6 (1.0)             | 3 (0.5)         |                                   |
deterioration in clinical status with higher priority for the most severe outcomes.

The concept of HCEs in clinical trials became more commonly incorporated in the heart failure (HF) setting to combine clinical outcome events (for example heart failure hospitalizations and cardiovascular death) with important patient-centered assessments, such as change in symptoms [1, 2]. Some of the recent large HF trials used such an HCE in the statistical testing hierarchy. For the DAPA-HF trial [19], an HCE that accounted for change in HF symptoms score at month 8 and death was a secondary endpoint. The EMPULSE trial [20] had an HCE that accounted for change in HF symptoms score at Day 90, death and heart failure events (hospitalizations or urgent visits) as the primary end-point. For the SOCRATES trial [21], non-disabling stroke, myocardial infarction, major bleeding, disabling stroke and death events were used to construct an HCE (which the authors called DOOR—desirability of outcome ranking). The DOOR endpoint was similar to HCEs for heart failure trials including only death and heart failure hospitalizations, proposed in [6].

Since all patients fully contribute to an HCE and the timing of the events is used to reduce the ties, an HCE is a sensitive endpoint to detect the treatment effect. This consideration is a potential advantage since the endpoint’s enhanced capability to demonstrate therapeutic efficacy could lead to studies with smaller sample sizes and potentially shorter (fixed) duration relative to those based only on “adverse clinical events.” In addition, an HCE accounting for both improvement and deterioration could be more important for studies evaluating interventions in acutely ill patients. However, HCEs pertain to scientific questions that are different from those for “adverse clinical events” and for which longer term outcome studies are necessary to investigate effects on chronic disease in the ambulatory setting. Ultimately, the choice of endpoints depends on multiple factors, including the estimand that accounts for the therapeutic goal, the patient population, and the mechanism of action of the intervention.

A potential limitation of designing trials based on an HCE and the timing of the events is used to reduce the ties, an HCE is a sensitive endpoint to detect the treatment effect. This consideration is a potential advantage since the endpoint’s enhanced capability to demonstrate therapeutic efficacy could lead to studies with smaller sample sizes and potentially shorter (fixed) duration relative to those based only on “adverse clinical events.” In addition, an HCE accounting for both improvement and deterioration could be more important for studies evaluating interventions in acutely ill patients. However, HCEs pertain to scientific questions that are different from those for “adverse clinical events” and for which longer term outcome studies are necessary to investigate effects on chronic disease in the ambulatory setting. Ultimately, the choice of endpoints depends on multiple factors, including the estimand that accounts for the therapeutic goal, the patient population, and the mechanism of action of the intervention.

A potential limitation of designing trials based on an HCE is its better feasibility for fixed-follow-up designs, mainly because the pairwise comparisons of patients with different follow-up times can be challenging [22]. However, the definition of an HCE may still have issues for a fixed-follow-up design when the extent to which patients are not followed for all events of interest for the HCE is not minimal. Thus, a careful assessment needs to be conducted for the potential extent and reasons for missingness of the data that pertain to the definition of an HCE before designing an HCE. When the expected amount of missingness is minimal or can be reasonably managed for the definition of an HCE for a fixed-follow-up design, the resulting patient-level rankings can enable the ordering of patients according to the individual values of their
ranks; and this property makes the treatment effect interpretable at a common, patient-level clinical scale. A challenge associated with the use of HCEs is the lack of standardization of these endpoints in different therapeutic areas and hence the possibility for arbitrary aspects for its definition in some situations; and this potential deficiency may lead to treatment effect estimates that are not interpretable and have little meaning for patients. Additionally, in some situations, establishing a threshold for a clinically meaningful treatment effect can be quite challenging for an HCE measured by the win odds, although graphical displays like the bar chart in Fig. 2 or those for corresponding cumulative distributions can be helpful for this purpose.

In trials of patients hospitalized with severe COVID-19, an HCE can be defined to include improvement in clinical status as well as deterioration, including death. This definition of the HCE provides a stricter definition of recovery, and it accounts for in-hospital worsening events and deaths after discharge. Additionally, the timing of all events can be useful to distinguish between patients who recover or who experience worsening, thus making the HCE more sensitive to capture the treatment effect. An HCE can be analyzed with the win/Mann–Whitney odds (win ratio with ties), and it does not require distributional assumptions for estimation (including the proportionality assumption). The win odds for an HCE can be useful for designing new trials, and it can provide a clinically meaningful treatment effect estimate with respect to the estimand framework, and potentially lead to more efficient trials with smaller sample size and shorter duration. Such trials can be especially important when clinical answers are needed expeditiously, such as during a pandemic.

Conclusions

In summary, HCEs are flexible endpoints that can be adapted for use in different therapeutic areas, with win odds as the analysis method. DARE-19 is an example of a COVID-19 trial with an HCE as one of the primary endpoints. This HCE provided clinically interpretable treatment effect estimates in the COVID-19 setting.

Author Contributions

All authors contributed to the interpretation of the results. The first draft of the manuscript was prepared by SBG and JB who had unrestricted access to the data. The Article was reviewed and approved by all authors.

Funding

DARE-19 was an investigator initiated collaborative study, with the study design and procedures operationalized through collaboration between Saint Luke’s Mid America Heart Institute (sponsor) and Astra-Zeneca (funding source).

Data Availability

Data for this article are not available.

Declarations

Conflict of interest

SBG, JO, OFB and RE are employees and stockholders of AstraZeneca. JB was an employee of AstraZeneca at the time the manuscript was being prepared and now is employed at Bristol-Myers-Squibb. EKK has nothing to declare. GGK is the Principal Investigator of a biostatistics grant from AstraZeneca. He is also the Principal Investigator for biostatistics grants from other biopharmaceutical sponsors that have no relationship to the submitted work. OB has received research Grants from AstraZeneca, Pfizer, Servier, Novartis, Amgen, Bayer, and Boehringer-Ingelheim. MNK has received a research grant for the conduct of this study from AstraZeneca. He has also received Grant and research support from AstraZeneca. He has received a Grant and honoraria from Boehringer-Ingelheim, and honoraria from AstraZeneca, Alnylam, Amgen, Bayer, Eli Lilly, Esperion, Merck (Diabetes and Cardiovascular), Janssen, Novo Nordisk, Pharmacosmos and Vifor Pharma.

References

1. Packer M. 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. 2016;134(21):1664–78.
2. Packer M. Proposal for a new clinical end point to evaluate the efficacy of drugs and devices in the treatment of chronic heart failure. J Cardiac Fail. 2001;7(2):176–82.
3. ICH E9(R1). Harmonised guideline addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials. Final version. Adopted on 20 November 2019.
4. Kosiborod M, et al. Effects of dapagliflozin on prevention of major clinical events and recovery in patients with respiratory failure because of COVID-19: design and rationale for the DARE-19 study. Diabetes Obes Metab. 2021;23(4):886–96.
5. Kosiborod MN, Esterline R, Furtado RH, et al. Dapagliflozin in patients with cardiometabolic risk factors hospitalised with COVID-19 (DARE-19): a randomised, double-blind, placebo-controlled, phase 3 trial. Lancet Diabetes Endocrinol. 2021. https://doi.org/10.1016/S2213-8587(21)00180-7.
6. Pocock SJ, Ariti CA, Collier TJ, et al. The win ratio: a new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. Eur Heart J. 2012;33(2):176–82.
7. Gasparyan SB, Folkvaljon F, Bengtsson O, et al. Adjusted win ratio with stratification: calculation methods and interpretation. Stat Methods Med Res. 2021;30(2):580–611.
8. Brunner E, Vandemeulebroecke M, Mütze T. Win odds: an adaptation of the win ratio to include ties. Stat Med. 2021;40(14):3367–84.
9. Gasparyan SB, Kowalewski E, Folkvaljon F, et al. Power and sample size calculation for the win odds test: application to an ordinal endpoint in COVID-19 trials. J Biopharm Stat. 2021. https://doi.org/10.1080/10543406.2021.1968893.
10. WHO. R&D. WHO R&D blueprint novel coronavirus (COVID-19) therapeutic trial synopsis; 2020. https://www.who.int/blueprint
11. Zarin DA, Rosenfeld S. Lack of harmonization of coronavirus disease ordinal scales. Clin Trials. 2021;18(2):263–4.

12. AstraZeneca [Online]//Drugs@FDA: FDA-approved drugs. https://www.accessdata.fda.gov/drugsatfda_docs/label/2021/202293s024lbl.pdf. Accessed 23 Dec 2021

13. Dodd LE, Follmann D, Wang J, et al. Endpoints for randomized controlled clinical trials for COVID-19 treatments. Clin Trials. 2020;17(5):472–82.

14. Taylor AL, Ziesche S, Yancy C, Carson P, D’Agostino R Jr, Ferdinand K, Taylor M, Adams K, Sabolinski M, Worcel M, Cohn JN. Combination of isosorbide dinitrate and hydralazine in blacks with heart failure. N Engl J Med. 2004;351(20):2049–57.

15. Finkelstein DM, Schoenfeld DA. Combining mortality and longitudinal measures in clinical trials. Stat Med. 1999;18(11):1341–54.

16. Buyse M. Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. Stat Med. 2010;29(30):3245–57.

17. Bebu I, Lachin JM. Large sample inference for a win ratio analysis of a composite outcome based on prioritized components. Biostatistics. 2016;17(1):178–87.

18. Gasparyan SB, Kowalewski EK, Koch GG Comments on “Sample size formula for a win ratio endpoint” by R.X. Yu and J. Ganju. Statistics in Medicine. 2022; 1–3. https://doi.org/10.1002/SIM.9379.

19. McMurray JJ, Solomon SD, Inzucchi SE, et al. Dapagliflozin in patients with heart failure and reduced ejection fraction. N Engl J Med. 2019;381(21):1995–2008.

20. Voors AA, Angermann CE, Teerlink JR, Collins SP, Kosiborod M, Biegus J, Ferreira JP, Nassif ME, Psopta MA, Tromp J, Borleffs C. The SGLT2 inhibitor empagliflozin in patients hospitalized for acute heart failure: a multinational randomized trial. Nat Med. 2022;28:1–7.

21. Evans SR, Knutsson M, Amarenco P, et al. Methodologies for pragmatic and efficient assessment of benefits and harms: application to the SOCRATES trial. Clin Trials. 2020;17(6):617–26.

22. Rauch G, Kunzmann K, Kieser M, et al. A weighted combined effect measure for the analysis of a composite time-to-first-event endpoint with components of different clinical relevance. Stat Med. 2018;37(5):749–67.

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.