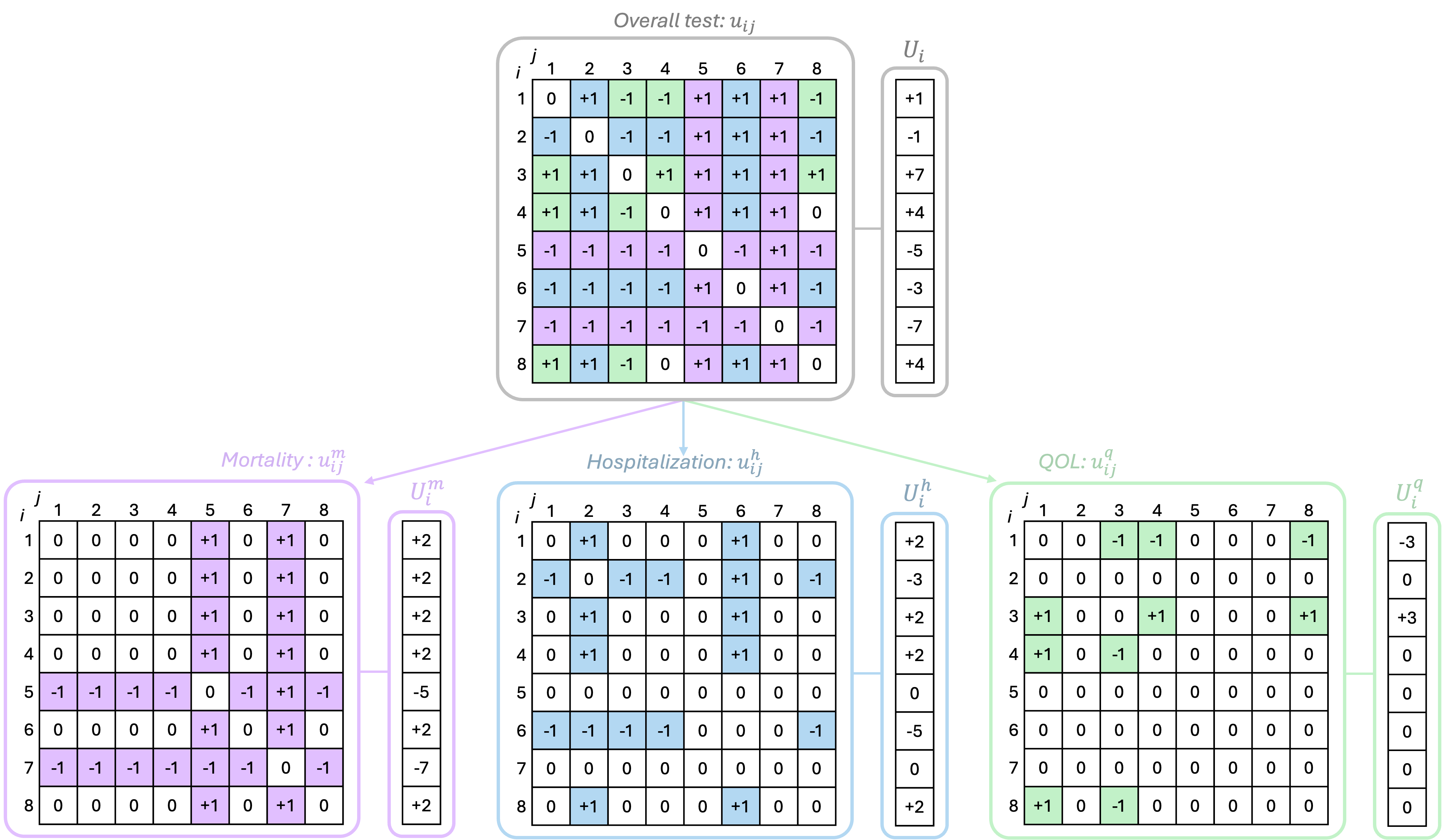
# Introduction

* Joint rank analyses (win ratio, Finkelstein-Schoenfeld test, etc.) have gained popularity in many disease areas, including heart failure. They are popular because they are non-parametric and allow for hierarchical ordering of endpoints based on clinical importance.
* Main challenges and criticisms of the methodology are related to interpretability. After a significant test, we immediately want to know endpoint-specific contributions and understand the key drivers of the overall result.
* Mike to elaborate on common complaints he hears from physicians that are related to the joint-rank methodologies and results.
* The use of these tests relies on the ability to decompose into endpoint-specific effects and contributions, but we tend to use the following, which have their own short-comings:
* Pocock-style win ratio breakdown (as in Fig 2 of <cit.>).
* Independent endpoint assessments. E.g. summarize the treatment effect on one endpoint for everyone in the dataset. This is informative but does not communicate the treatment effect that the joint-rank method sees from the endpoint.
* With three endpoints, frame as: rather than 3 independent endpoint evaluations with sample size N, the joint-rank methods do 3 tests on partitions of $N \times N$ pairwise comparisons. Transition to a description of our approach with partitioned pairwise comparisons and resulting test quantities.

# Methods

Within the context of a randomized clinical trial with a hierarchical composite endpoint, we develop a decomposition of the FS test into partitioned, endpoint-specific tests and corresponding contribution factors. The FS test relies on generalized pairwise comparisons <cit.>, where every participant is compared to every other participant according to the endpoint hierarchy. For illustration, we consider an endpoint that is a hierarchical composite of (1) time to mortality, (2) recurrent hospitalization events, and (3) a quality of life (QOL) measurement evaluated at a fixed time point.

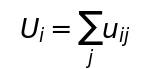
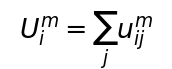
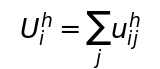
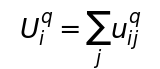
For the comparison of participant *i* to participant *j*, denote the pairwise win, tie, or loss as *uij = +1, \: 0, or -1*, respectively, and define *uij = 0* for *i = j*. In a trial with *N* participants, these pairwise results can be organized into an *N× N* square matrix, demonstrated in the top row of Figure <ref>. The endpoint that generated each pairwise win or loss is shown by the color of the cells.



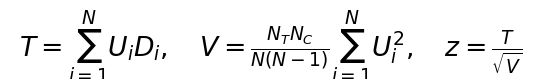
*Figure 1: REDUCE LAP HF II trial complete test statistic, importance factors, and endpoint-specific test statistics*

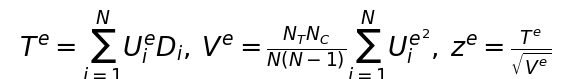
The pairwise comparison results can be partitioned according to the endpoint that generated the result. Conceptually, the pairwise comparison matrix can be divided into endpoint-specific matrices as shown in Figure <ref>. We define the mortality-specific pairwise comparison results as *umij = +1 or -1* if the mortality endpoint determined the pairwise win or loss, and *0* otherwise. Similarly, let *uhij = +1 or -1* if the hospitalization endpoint determined the pairwise win or loss (participants tied on the mortality endpoint comparison, and the tie was broken by the hospitalization endpoint comparison), and *0* otherwise. Finally, let *uqij = +1 or -1* if the pairwise win or loss was determined by the quality of life component (participants tied on the mortality and hospitalization endpoint comparisons, and the tie was broken by the quality of life endpoint comparison), and *0* otherwise.

For example, consider comparison of participant *i=2* to participant *j=6* in Figure <ref>. Both participants survived to the end of the trial, but participant 2 experienced one hospitalization while participant 6 had two hospitalizations. The endpoint-specific pairwise outcomes are *um2,6 = 0*, *uh2,6 = +1*, and *uq2,6 = 0*.

Partitioning the pairwise results in this way leads to a natural decomposition of the overall FS test quantities. In the FS test, the *uij* are used to construct a net score for each participant, . These correspond to the row sums of the matrix in the top row of Figure <ref>. Similarly, endpoint-specific net scores can be calculated for each participant as , , and . These correspond to the row sums of each endpoint-specific matrix in the bottom row of Figure <ref>. For each participant, the endpoint-specific net scores add to the overall net score (i.e., *Ui = Umi + Uhi + Uqi*).

In the FS test, the individual net scores, *Ui*, and treatment arm assignments, *Di*, are combined to calculate the treatment arm net wins *T*, and variance, *V*. Let *Di = 0 or 1* if participant *i* was in the control or treated group, respectively. Then,



where *NT* and *NC* denote the number of subjects in the treated and control groups, respectively, and *N=NT+NC*. The normal deviate, *z*, is the test statistic used to calculate a *p*-value and test the hypothesis of equality between the treatment and control groups. Similarly, the endpoint-specific quantities are defined as , where *e* represents an endpoint in the hierarchy and, for our example, can be replaced with *m, h,* or *q*.

Importantly, these are not the quantities that arise if each endpoint is tested independently. Instead, they rely on the conditional partitioning of the pairwise comparisons and conform to the hierarchy of the endpoint definition. The conditional partitioning ensures that *T*, the total number of net wins in the treatment arm, is equal to the sum of the net wins from each endpoint-specific partition: *T = Tm + Th + Tq*. A natural decomposition of the overall test statistic, *z*, follows from this relationship:

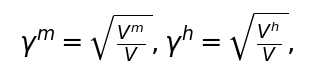
\begin{align\*}

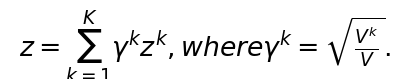
T = T^m + T^h + T^q &\iff z\sqrt{V} = z^m\sqrt{V^m} + z^h\sqrt{V^h} + z^q\sqrt{V^q}\\

&\iff z = z^m\sqrt{\frac{V^m}{V}} + z^h\sqrt{\frac{V^h}{V}} + z^q\sqrt{\frac{V^q}{V}}\\

&\iff z = \gamma^m z^m + \gamma^h z^h + \gamma^q z^q.

\end{align\*}

The overall test statistic, *z*, can be decomposed into a linear combination of the endpoint-specific test statistics *zm*, *zh*, and *zq*, and ratios of standard deviations  and , which we will refer to as “contribution factors.” These factors provide a straightforward expression of the implied utility function for the contribution of each endpoint in the hierarchy to the overall test.

While the endpoint-specific definitions have been provided for a 3-level example endpoint, they can be easily extended to accommodate a hierarchical composite endpoint with *K* elements. More generally, the main result is 

**Properties and interpretation**. The contribution factors are ratios of standard deviations and simplify to $\gamma^k = \sqrt{\frac{V^k}{V}} = \sqrt{\tfrac{\sum\_{i=1}^{N}U\_i^{k^2}}{\sum\_{i=1}^{N}U\_i^2}}$, for a generic endpoint component, $k$. The greater imbalance in wins and losses among subjects (i.e., $|U^k\_i| >> 0$), the higher the contribution factor and the more “separation” power the endpoint has.

Notice that the contribution factors do not depend on treatment assignments. In the decomposition of the overall test statistic, *z*, the endpoint-specific test statistics, *zk*, reflect how each endpoint contributes to the assessment of treatment effect, the contribution factors, *γk*, reflect the relative weight of each endpoint in the final test result. Conceptually, they can be thought of as scaling factors that adjust the contribution of each endpoint-specific test statistic, *zk*, to the overall test statistic, *z*. Higher contribution factors indicate that the endpoint has a larger influence on the overall test. For example, an endpoint component that fails to show a treatment effect can have an endpoint-specific test statistic near zero while still having a large contribution factor if many decisions are made based on that component.

Individual contribution factors are bounded between *0* and *1* because the variances of the endpoint-specific tests must be smaller than the variance of the overall test. Values near 0 indicate that the endpoint component provides relatively little information to the overall test while values near 1 indicate that the endpoint provides a relatively large amount of information. The sum of the contribution factors is bounded below by 1.

**Normalized contribution factors.** The contribution factors provide the recipe for constructing the complete FS test statistic, $z$, from partitioned endpoint-specific statistics, $z\_k$. Similar to a baking recipe using ratios (e.g., 1 part sugar, 1 part butter, 3 parts flour), the contribution factors provide the ratio of the partitioned information that exactly creates the complete test. In the same way the recipe would be comprised of 20% sugar, 20% butter, and 60% flour, the contribution factors can be normalized for a fractional representation.

For example, in a setting with contribution factors of *0.3, 0.3, \text{ and } 0.9*, (i.e., *z = 0.3 · zm + 0.3 · zh + 0.9 · zq*), we can attribute 20%, 20%, and 60% of the complete test statistic to information from the first, second, and third endpoints in the composite, respectively. Normalized contribution factors sum to 1 while keeping their relative weights. Normalizing can simplify interpretation and allow comparison across endpoint definitions or studies.

# Application to the REDUCE LAP HF II Trial Data

REDUCE LAP HF II was a prospective, double-blinded, sham-controlled randomized clinical trial to assess the effectiveness of the Corvia Medical, Inc. Atrial Shunt System in patients with heart failure with preserved ejection fraction <cit.>. The study design and justification has been described previously in <cit.>. The primary analysis method was the FS test, and the primary endpoint of the trial was a hierarchical composite endpoint with the following components:

* cardiovascular (CV) mortality or non-fatal ischemic stroke measured by time of event out to 1 year of follow-up,
* heart failure hospitalizations requiring IV dieresis or urgent outpatient visits requiring intensification of oral diuretics measured by number of recurrent events out to 2 years of follow-up, and, if applicable, time of first event, and
* QOL assessed by the Kansas City Cardiomyopathy Questionnaire overall summary score (KCCQ-OSS) <cit.> measured as change from baseline to 1 year of follow-up.

For pairwise comparisons, follow-up time was limited to the shortest observed time in the pair, and QOL comparisons were considered tied if the KCCQ-OSS change from baseline scores for the pair were within 5 points of each other.

The trial randomized 626 participants, and the primary analysis result was neutral with a win ratio of 1.0 and an FS *z-*statistic near zero <cit.>.

<cit.> identified a biologically defined responder population composed of half of the participants (N=313) who did not have latent pulmonary vascular disease. The post-hoc analysis of this population yielded a win ratio of 1.51 and a statistically significant FS test with test statistic *z = 2.63*. A trial to confirm the benefit of the Corvia Atrial Shunt System in this population is underway.

To understand how each component of the hierarchical endpoint influenced the positive FS test in the responder population, the partitioned, endpoint-specific test statistics and contribution factors were calculated (Table <ref>). The sum of the partitioned test statistics, weighted by the contribution factors, results in the observed FS test statistic, *2.63 = 0.145×(-1.37) + 0.653 × (1.89) + 0.744 × (2.15)*.[This FS statistic may vary slightly from previously published numbers due to the timing of the data export as we are using data exported on September 9, 2024.]

The CV mortality or non-fatal ischemic stroke test statistic, *zm = -1.37*, arises from only two events in the dataset. Both events were in the treated group. The corresponding contribution factor, *γm = 0.145* reflects the relatively small influence of this endpoint to the overall test result. The positive hospitalization-specific test statistic, *zh = 1.89*, and positive QOL-specific test statistic, *zq = 2.15* suggest that treated patients tended to fare better than control patients on both endpoints. The corresponding contribution factors, *γh = 0.653* and *γq = 0.744*, reflect moderate influence on the overall test from each endpoint.

The normalized contribution factors indicate that *9.4%*, *42.4%*, and *48.2%* of the overall FS test is informed by the mortality, hospitalization, and QOL endpoints, respectively. To provide a comparison, the fractional breakdown of decisions in the win ratio are included in Figure <ref> as *1.5%*, *33.4%*, and *65.2%*. Compared to the normalized contribution factors, the win-ratio breakdown indicates a larger contribution of the QOL endpoint and a smaller contribution of hospitalization and CV mortality.

We find that the win ratio decision breakdown tends to systematically under-weigh the influence of endpoints with a small number of events (particularly for endpoints higher in the hierarchy) and over-weigh the influence of endpoints lower in the hierarchy compared to the **Normalized contribution factors.** This behavior is discussed in more detail in Section <ref>.

10pt8pt

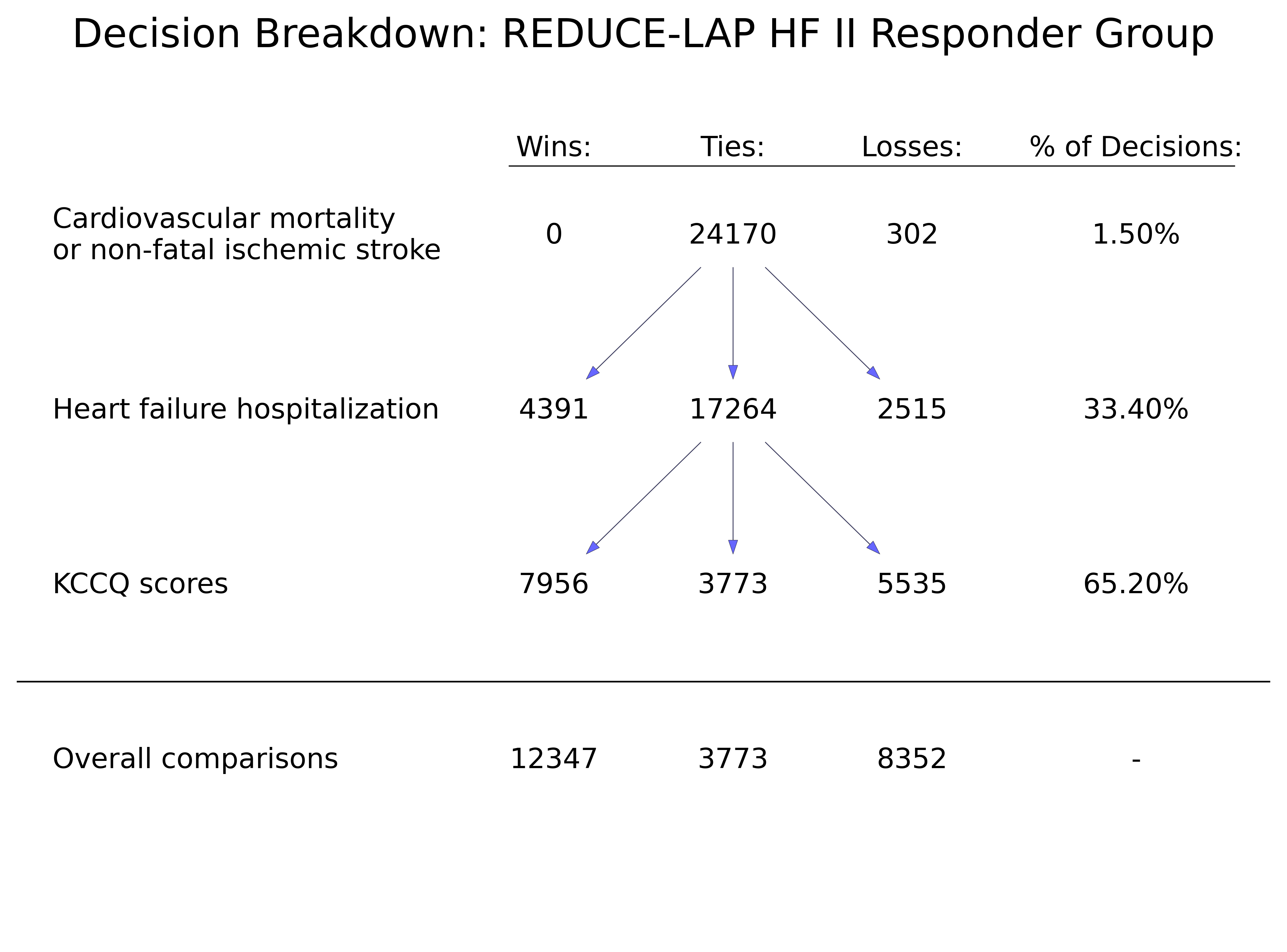
Endpoint component Partition test-statistic Contribution factor Normalized contr. factor

1. CV mortality/stroke (*m*) *zm = -1.37* *γm = 0.145* *nγm* = 0.094

2. Heart failure events (*h*) *zh = 1.*89 *γh = 0.653* *nγh* = 0.424

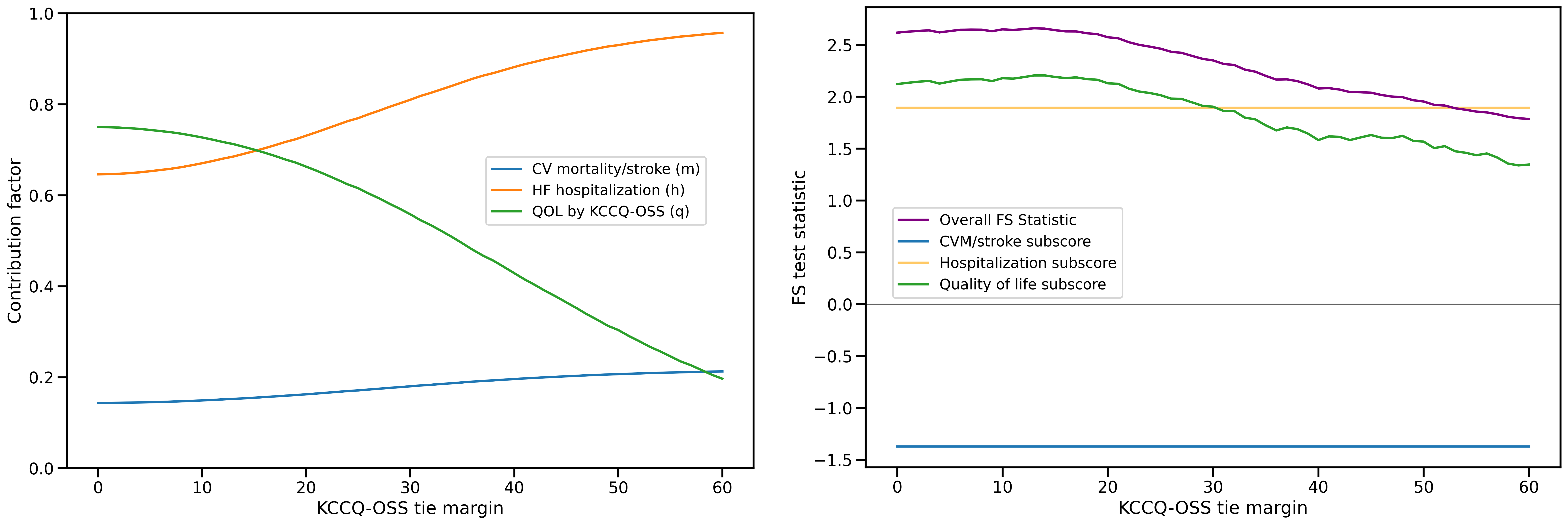
3. QOL via KCCQ-OSS (*q*) *zq = 2.15* *γq = 0.744* *nγq* = 0.482

Partitioned Finkelstein-Schoenfeld Test Statistics, Contribution Factors, and Normalized Contribution Factors for REDUCE LAF HF II Responder Population



*Figure 2: Partitioned Finkelstein-Schoenfeld Test Statistics, Contribution Factors, and Normalized Contribution Factors for REDUCE LAF HF II Responder Population*

Contribution factors can provide insight into the sensitivity of the observed FS score to the definition of win and loss for each component of the hierarchical composite endpoint. For example, the REDUCE LAP HF II trial employed a 5-point tie margin for comparisons on the KCCQ-OSS endpoint. The tie margin operates on a pairwise basis, meaning that if a pair of KCCQ-OSS change from baseline outcomes are within 5 points of one another, the comparison result will be a tie. Other heart failure trials have employed different tie margins for pairwise comparisons on the KCCQ-OSS endpoint, including dichotomized and ordinalized definitions <cit.>, <cit.>. The selection of an appropriate tie margin is largely a question of clinical judgment, and post-trial sensitivity analyses surrounding that choice can be handled naturally with the contribution factor framework.



*Figure 3: Win Ratio Decision Fractions for REDUCE LAF HF II Responder Population*

In Figure <ref>, we explore the behavior of the overall FS test and the contribution factor breakdown for the REDUCE LAP HF II responder data when different KCCQ-OSS tie margins are applied within the pairwise comparison framework. As the tie margin increases, fewer pairwise decisions are made on the KCCQ-OSS endpoint. As a result, the KCCQ-OSS contribution factor decreases, the KCCQ-OSS endpoint-specific FS test statistic decreases, and the overall FS test statistic decreases.

The contribution factors provide a logical framework for addressing questions of clinical trial design and interpretation. For example, the sensitivity analysis above could be used to identify the KCCQ tie margin at which the hospitalization and quality of life endpoints contribute equally to the composite. Alternatively, a researcher may want to know at what tie margin does the overall composite no longer represent a statistically significant improvement for treated patients.

# Behavior and Simulation

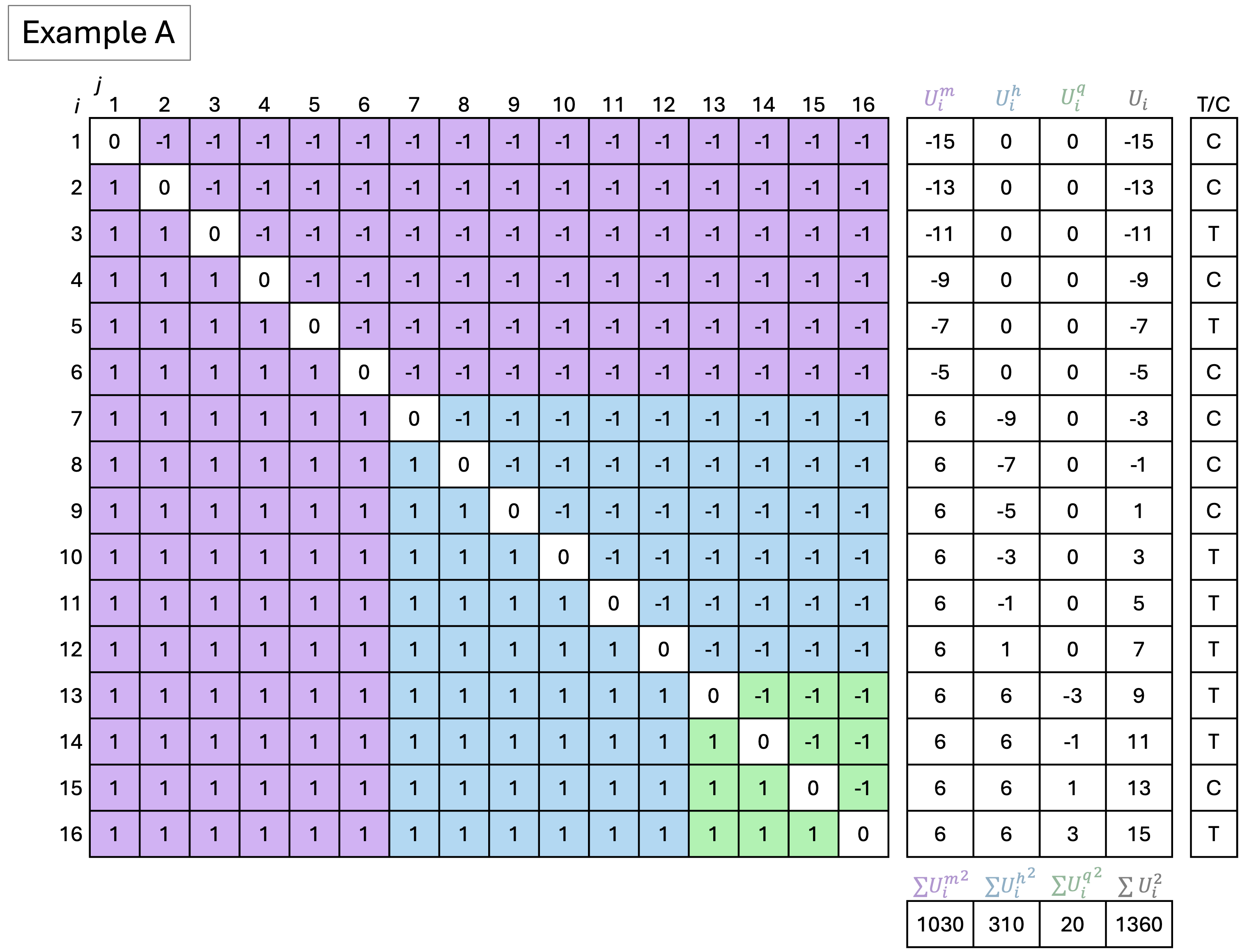
## Examples

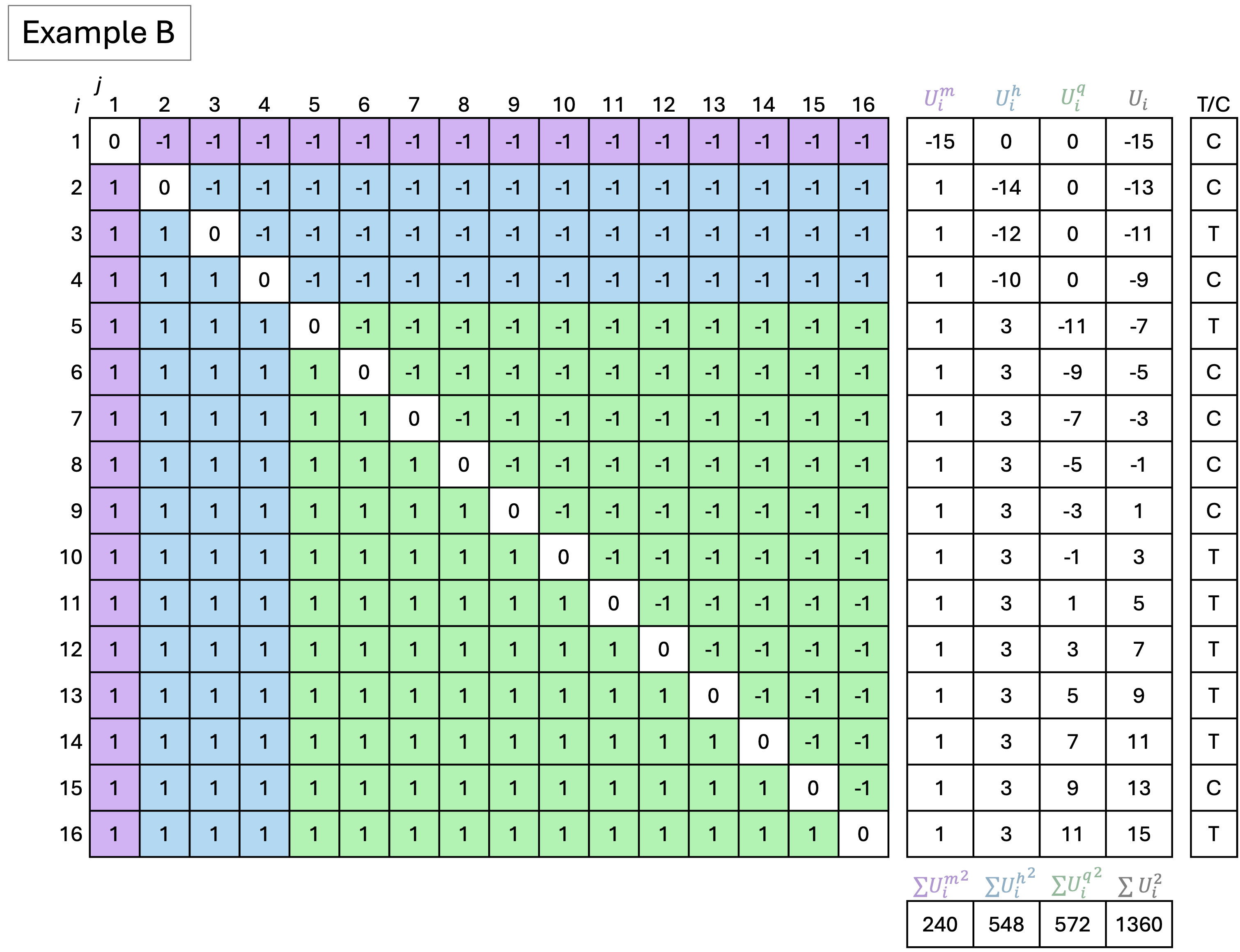
In this section, we construct examples with two small datasets to demonstrate behavior and interpretation of the contribution factors. We continue to use the three-level hierarchical composite endpoint defined in Section <ref>. Figure <ref> depicts the pairwise comparison results from the example trial datasets, each with *N=16* participants. The datasets have been constructed so that the pairwise comparison outcomes (the *uij \in \{-1,0,1\}*) are the same, but the contribution of each endpoint to the outcome changes.

In the first dataset (example A), more decisions are made on the mortality endpoint, and fewer decisions are made on the QOL endpoint. For the second dataset (example B), fewer decisions are made on the morality endpoint and more on the QOL endpoint.

The contribution factors are calculated by first deriving the endpoint-specific net scores (*Uim*, *Uih*, and *Uiq*) and the overall net score (*Ui*) for each participant. Next, the sum of the squared net scores (, , , and ) across all participants are calculated. For example A, these values are , , , and  and for example B, they are , , , and . Finally, the contribution factors for each endpoint are obtained by taking the positive square root of the ratio of the endpoint-specific to the overall sum of squared net scores. For example A, we obtain , , and . For example B, this calculation yields contributions factors of *γm ≈ 0.420*, *γh ≈ 0.635*, and *γq ≈ 0.649*.

Both examples have been constructed with identical overall FS test statistics (*z = 1.68*), yet they have very different underlying decompositions. Although the contribution factors do not rely on treatment assignment, it is beneficial to interpret them in the context of the treatment effect observed in the endpoint-specific partitioned tests. In example A, there are positive treatment effects for mortality and hospitalization (*zm = 1.09*, *zh = 1.65*) but a negative effect for quality of life (*zq = -0.43*). The positive effects in the first two partitions of the hierarchy carry more influence in the overall test and are only partially offset by the negative effect for the third endpoint, which carries less relative weight. In example B, there is a moderate positive treatment effect for all three outcomes (*zm = 1.00*, *zh = 0.74*, and *zq = 1.21*). The contribution of mortality to the overall test is lower while the contributions of both hospitalization and QOL are higher.





*Figure 5: Pairwise Comparisons and Contribution Factor Quantities for Example Datasets*

Pairwise Comparisons and Contribution Factor Quantities for Example Datasets

10pt8pt

2\*Example 3c|Contribution Factors 3c|Normalized Contribution Factors 3c|Partitioned FS *z*-scores

2-10

Mortality Hosp. QOL Mortality Hosp. QOL Mortality Hosp. QOL

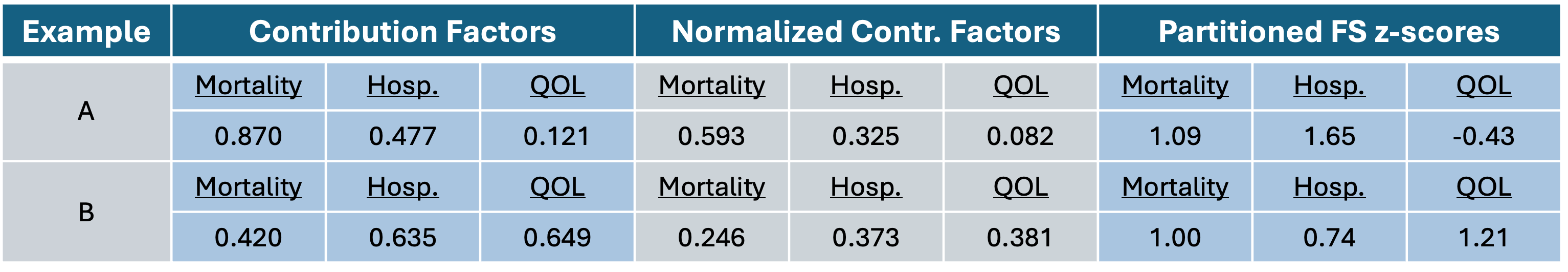
A 0.870 0.477 0.121 0.593 0.325 0.082 1.09 1.65 -0.43

B 0.420 0.635 0.649 0.246 0.373 0.381 1.00 0.74 1.21

Contribution Factors, Normalized Contribution Factors, and Partitioned Finkelstein-Schoenfeld Test Statistics for Example Datasets

Figure <ref> shows the pairwise comparison outcomes and decision breakdowns for the win ratio. As discussed previously, the win ratio relies on a restricted set of pairwise comparisons of treated versus control subjects (treated versus treated and control versus control pairs are excluded). In example A, 40 pairwise comparisons are decided by the mortality endpoint, 21 are decided by the hospitalization endpoint, and 3 are decided by the quality of life endpoint. The corresponding decision fractions are 0.625, 0.328, and 0.047. In example B, the number of pairwise comparisons decided by mortality, hospitalization, and quality of life are 8, 21, and 35, respectively, so the decision fractions are 0.125, 0.328, and 0.547.

The decision fractions convey information about the number of comparisons between treatment and control subjects decided by each endpoint, but not whether each comparison was a win or loss. By contrast, the win and loss information is included in the contribution factor through the standard deviation of each endpoint-specific test which is derived from the endpoint-specific net scores. An endpoint with a greater imbalance of wins and losses will have larger standard deviation relative to the standard deviation of the overall FS test and a larger importance factor. While both contribution factors and decision fractions can provide similar qualitative results, only the contribution factor accounts for both the hierarchical ordering and variability of the endpoints and can be tied directly to the partitioned FS tests.



*Figure 6: Caption*

The hierarchical nature of the composite endpoint plays an important role in differentiating between the contributions of individual components, even when decision fractions are identical. In Figure <ref>, the decision fractions for the hospitalization endpoint are 0.328 for examples A and B. The importance factors are 0.477 and 0.635, respectively, because the standard deviation of the test for hospitalization is smaller for example A (9.09) than example B (12.08). The normalized importance factors for hospitalization of 0.325 and 0.373 for example A and B, respectively, reflect the difference in how hospitalization contributes to the overall test, even though the decision fractions are identical across scenarios. This difference arises because the contribution factors preserve the hierarchy of the composite endpoint.

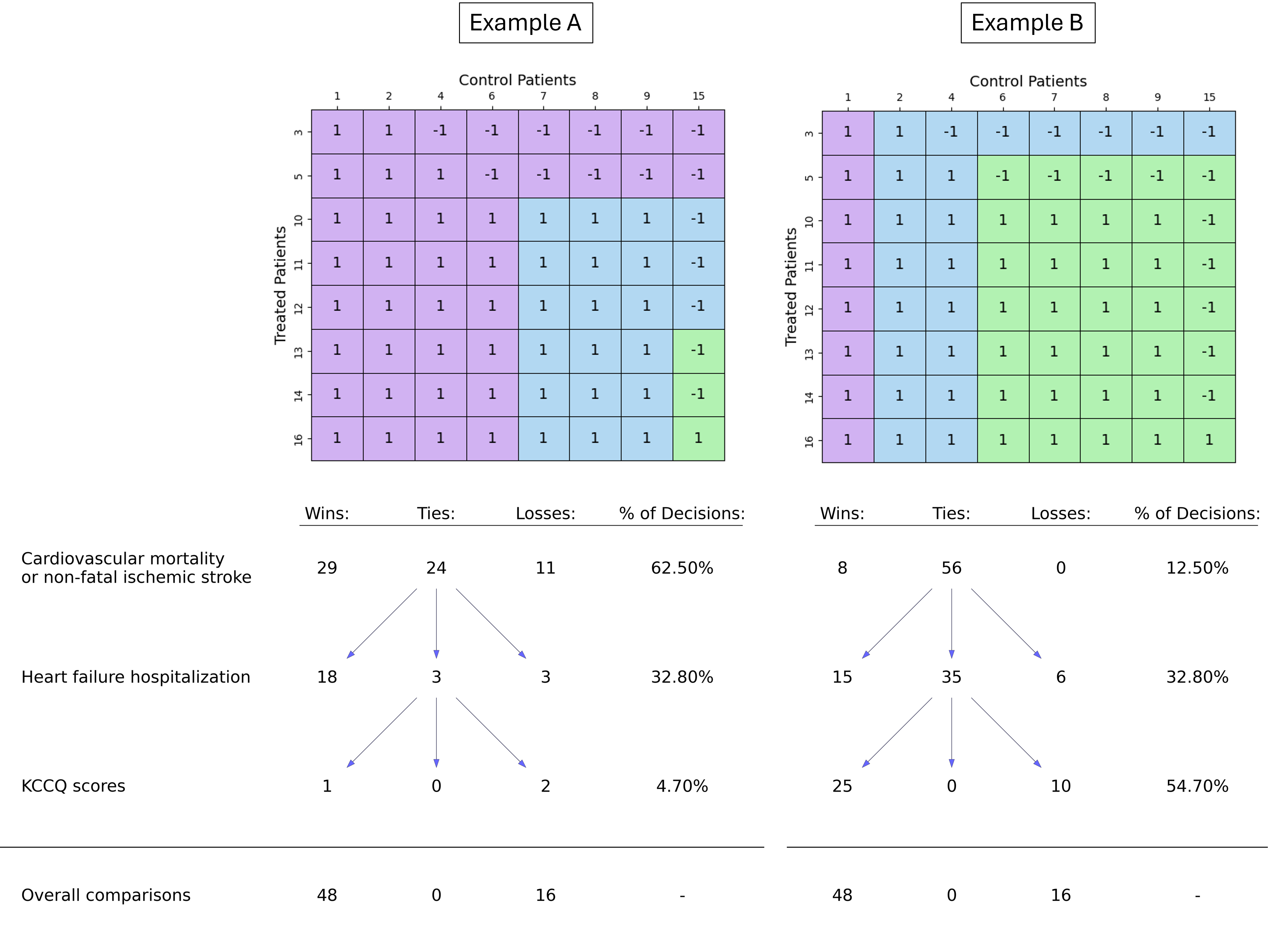
## Simulation study

We explore the behavior of contribution factors through repeated clinical trial simulations. Each trial includes *N=200* subjects with outcomes for three endpoints in the composite endpoint defined above. The simulation study crosses three factors, each aimed at changing the relative amount of information for endpoint with respect to the overall FS test.

* Average cardiovascular-related mortality control rate: 0, 0.04, 0.08, 0.15, 0.25, 0.40, and 0.70 events per person per year, on average.
* Average recurrent hospitalization control rate: 0.05, 0.1, 0.2, 0.5, 0.7, 1.1, and 1.5 events per person per year, on average.
* KCCQ-OSS tie margin: 0, 0.1, 5, and 20. When pairwise comparisons are conducted, the KCCQ-OSS endpoint will be considered a tie if the observations are within the tie margin of one another.

Simulated subjects are assigned to treatment or control with equal randomization. We do not apply a treatment effect to the cardiovascular-related mortality endpoint. For the hospitalization endpoint, we assume a standard deviation of 2.5 times the average control rate and that treated subjects have a *50%* reduction in the incidence of hospitalization events. For the KCCQ-OSS endpoint, we assume that control and treated subjects have an average 12-month change from baseline of 10 points and 6 points, respectively, with a standard deviation 20 points. Accrual was simulated with a 220-day ramp-up period to a steady average enrollment rate of 0.4 subjects per day. The trials use a common close design where subjects are followed for 2 years or until the common close date, defined to be 12 months after the last randomization. We assume a dropout rate of 0.015 subjects per year. Simulated trials are included if at least one death and at least one hospitalization occurred so that we are always comparing results from a 3-level hierarchical composite endpoint. Additional simulation details can be found in the supplementary material.

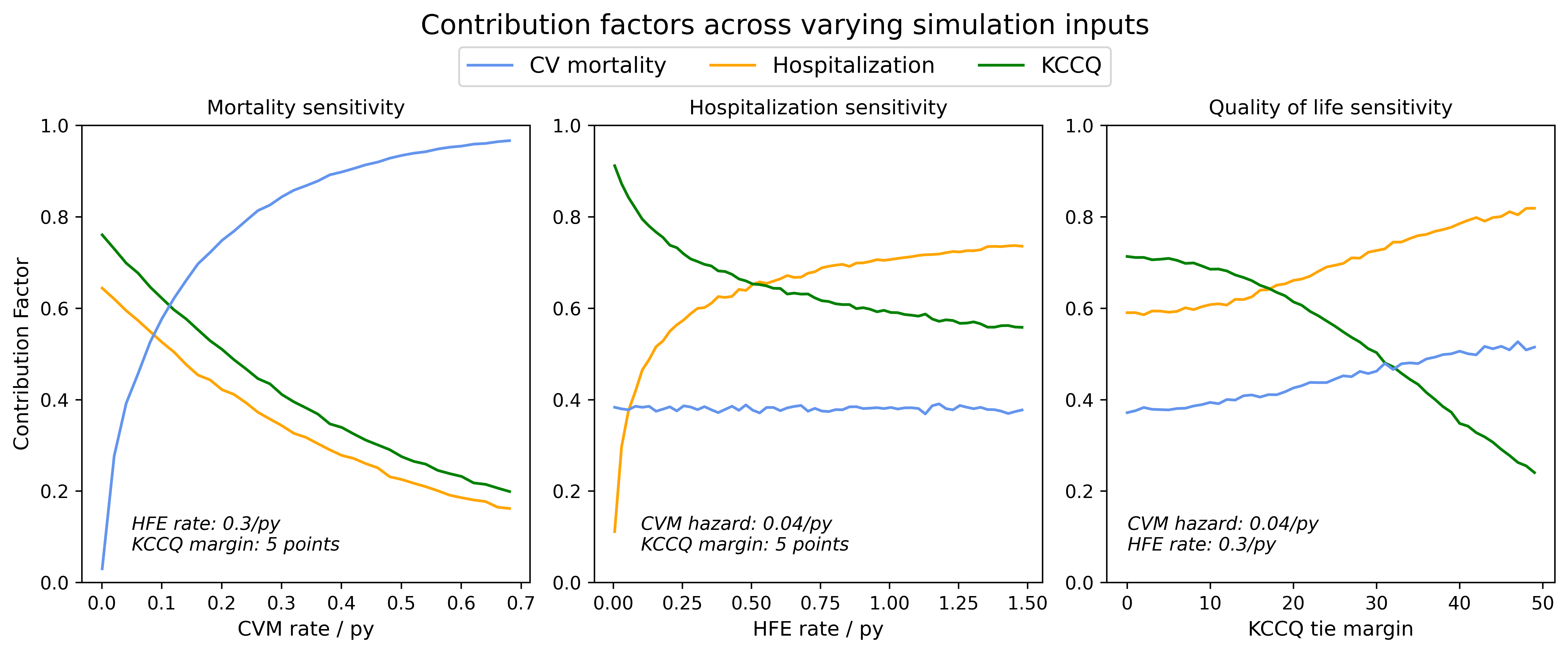
Figure <ref> shows how the contribution factors change as the simulation inputs are varied. The left plot shows the remarkable sensitivity of the contribution factors to the underlying cardiovascular mortality hazard. The CVM contribution factor rapidly approaches one as the mortality rate increases beyond a rare event, even as the hospitalization rate and KCCQ tie margin are held constant. The center plot shows the behavior of the contribution factors as the hospitalization rate increases: the mortality contribution remains relatively constant since comparisons on mortality occur before hospitalization. However, as the underlying hospitalization rate increases the hospitalization contribution factor gradually increases while the quality of life contribution gradually decreases. Finally, the right most plot shows the reduction of the quality of life contribution to the endpoint as the tie-margin is increased. Even under fixed mortality and heart failure rates the contribution factors for these earlier elements in the hierarchy increase as the quality of life margin increases. This can be attributed to the reduction in variance of the overall FS test as more ties are introduced by the larger indifference margin.



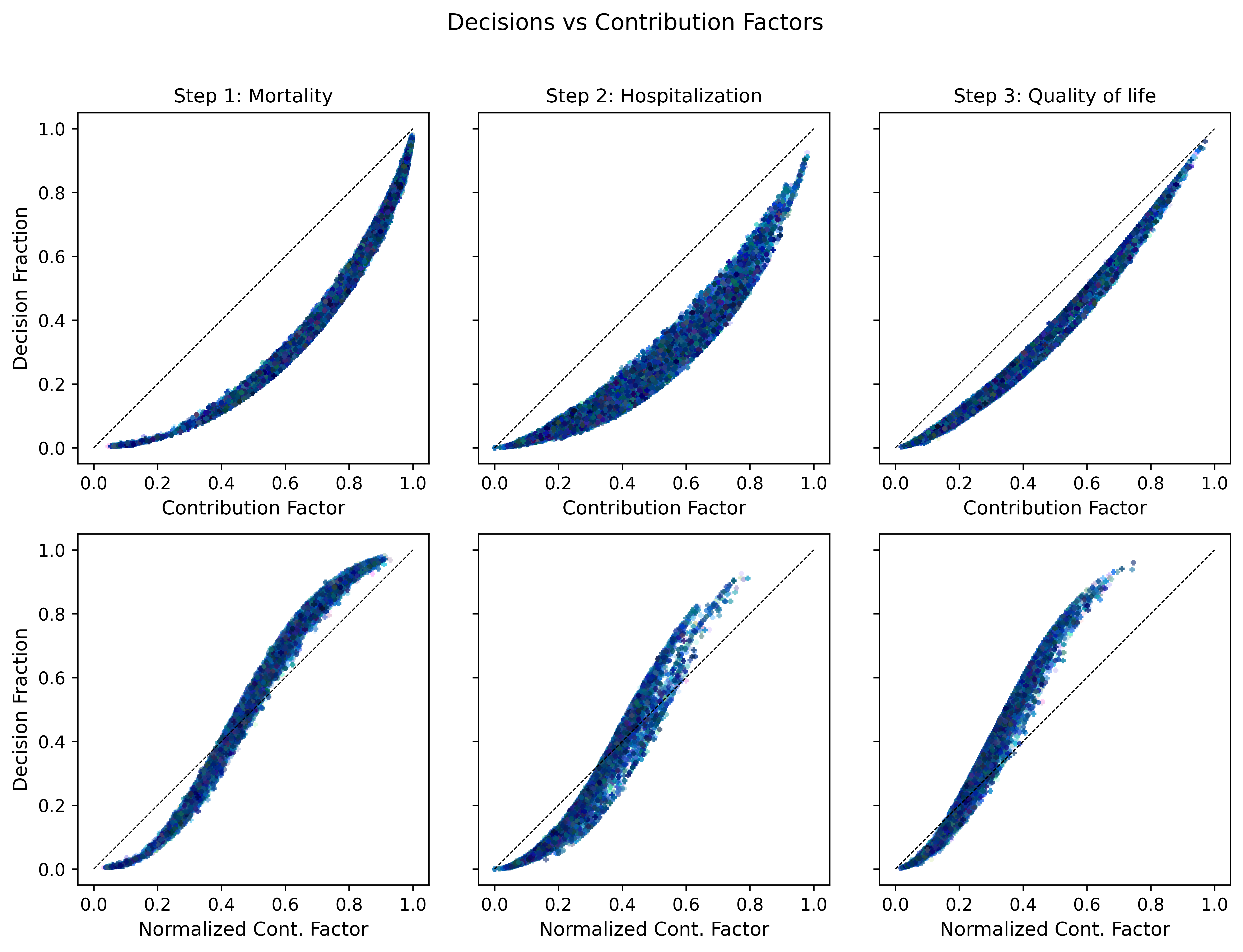
*Figure 7: Contribution Factors, Normalized Contribution Factors, and Partitioned Finkelstein-Schoenfeld Test Statistics for Example Datasets*

Figure <ref> shows the behavior of contribution factors relative to win-ratio decision fractions for each component of the composite endpoint in the simulated trials. Points in the top three plots lie to the lower right of the line of identity (*x=y*). This pattern is expected because the sum of the contribution factors is bounded below by one. The distance from the identity line is most pronounced for the first element of the hierarchical composite and diminishes for subsequent elements.

The bottom three plots in Figure <ref> show the normalized contribution factors relative to win-ratio decision fractions for each endpoint and Figure <ref> shows the same results in a single plot. Compared to the contribution factors, win ratio decision fractions tend to understate the contribution of endpoints that break a small number of ties (lower left corners of plots). This is particularly apparent for the mortality and hospitalization endpoints. In contrast, win ratio decision fractions tend to overstate the contribution of endpoints that break a moderate or large amount of ties (upper right corners of plots). This arises because the win ratio is a point estimate designed to capture treatment effect magnitude as the discrepancy in wins when treated subjects are compared to control subjects. The fractional breakdown of the win ratio decisions gives information about the contribution of each endpoint to the signal of the point estimate. On the other hand, the FS test and the contribution factors account for the fact that events on high-ranking endpoints tend to reduce the variability of what can be measured by lower-ranking endpoints in the comparison hierarchy (e.g., if a patient dies, the opportunity for pairwise decisions on hospitalization or QOL endpoints are eliminated when that patient is in the pair). Thus, even when few events are observed, higher-ranking endpoints lend more information to the FS test than just the paired decisions generated by the endpoint. This behavior is naturally captured by the contribution factors because they arise from the partitioned FS test construction.



*Figure 8: Pairwise Comparison Outcomes and Decisions For Example Datasets*



*Figure 9: Comparison of Contribution Factor Changes for Simulation Study. (A) Change in Contribution Factors for changing CVM rates and constant HFE rate and KCCQ margin (B) Change in Contribution Factors for changing HFR rates and constant CVM rate and KCCQ margin (C) Change in Contribution Factors for changing KCCQ tie margin and constant CVM rate and HFE rate.*

A single simulated trial is highlighted by the circled plotted points in Figure <ref>. In this trial, mortality was the deciding endpoint in 12.8% of paired comparisons, while the decision fraction was 36.3% for hospitalization and 50.9% for quality of life. This contrasts with the normalized contribution factors of 24.9%, 39.6%, and 35.5% for the mortality, hospitalization, and quality of life components, respectively. Using the decision fraction alone to interpret this composite endpoint could lead to a misleading conclusion that the quality of life measure was responsible for the majority of the composite result, when in reality, heart failure hospitalization provided the largest relative contribution to the FS test, followed closely by the quality of life measurement.

# Discussion

* Summarise the testing structure with the staging framing (Nathan).
* Blurp we moved from the methods section down to discussion: “The partitioned, endpoint-specific tests can provide insight into the information contained in each endpoint after the hierarchy has been considered. However, the authors caution against stand-alone interpretation of these tests. The partitioned tests are not independent of one another, are prone to small sample sizes, and should be viewed as parts of a whole picture.”