Scientific Research Publishing

Advances in Pure Mathematics, 2026, 16(2), 91-117
https://www.scirp.org/journal/apm
ISSN Online: 2160-0384
ISSN Print: 2160-0368

Prioritized and Non-Prioritized Net Benefit Statistics for the Assessment of Benefit-Risk

Yodit Seifu[1*], Revathi Ananthakrishnan[1], Margaret Gamalo[2], John Kolassa[3]

[1]GBDS, Bristol-Myers Squibb, Madison, NJ, USA
[2]Pfizer Inc., Collegeville, PA, USA
[3]Department of Statistics, Rutgers, The State University of NJ, Piscataway, NJ, USA
Email: *yodit.seifu@bms.com

How to cite this paper: Seifu, Y., Ananthakrishnan, R., Gamalo, M. and Kolassa, J. (2026) Prioritized and Non-Prioritized Net Benefit Statistics for the Assessment of Benefit-Risk. *Advances in Pure Mathematics*, **16**, 91-117.
https://doi.org/10.4236/apm.2026.162007

Received: January 18, 2026
Accepted: February 23, 2026
Published: February 26, 2026

Abstract

For a given therapy, patients, health technology assessment agencies and regulatory agencies are interested in evaluating whether the benefit outweighs the associated risks. Existing work [1] uses prioritized composite outcomes to assess benefit and risk; this manuscript extends this work. The assessment of risk is based on a derived score obtained from predefined adverse events of interest. This derived safety score incorporates full aspects of adverse events of interest. After prioritizing the components of the composite outcomes consisting of benefit and risk outcomes, the net benefit or the Wilcoxon-Mann-Whitney statistic is used to assess the composite benefit-risk outcomes. If there is no prioritization, average net benefit (also called the O'Brien method) can be used to assess the treatment differences in the composite outcomes. Via simulation, we evaluate the characteristics of these measures of benefit-risk. An existing sample size derivation is extended to the case where there is no prioritization (average net benefit). Finally, an example motivated by a plaque psoriasis study is presented.

Keywords

Benefit-Risk Assessment, Treatment-and-Disease-Burden, Prioritized Net Benefit, Non-Prioritized Net Benefit

1. Introduction

Patients, health technology assessment agencies, and health authorities all have an interest in evaluating the benefit-risk (BR) of a therapy for the proposed condition of use. Frequently, this is evaluated by assessing efficacy (benefit) and safety (risk) separately. Various authors [1]-[6] have proposed evaluating BR at a patient level

utilizing prioritized composite outcomes consisting of efficacy and safety endpoints; detailed reviews of these discussions exist [7].

Risk assessment using binary safety endpoints, which indicate the presence or absence of an event has been evaluated in [4] and [5], as has been the assessment of risk through an ordinal endpoint based on the severity grade of adverse events (AEs) in [8] and [3]. Binary endpoints and time-to-event endpoints were used to measure risk in [6]. In contrast, measurement of risk with a composite score for adverse events of interest (AEI), considering both the severity and duration of the AE, was used in [1].

Composite endpoints have been used as a primary analysis in several therapeutic areas. Progression-free survival, the composite endpoint of progression or death, assessed in oncology studies, is such an example. Another group of approaches assumes some ordering of importance among efficacy and safety endpoints of interest. For example, in oncology, for the progression-free survival composite endpoint, the order of importance would be death and then progression of disease. The preference-ordered outcomes are referred to as prioritized outcomes. In contrast with composite endpoints such as progression-free survival, prioritized outcomes preserve the multivariate nature of the outcomes. After the priority of importance is defined, treatment interventions can be compared using generalized pairwise comparisons (GPCs). In applying this method, one performs a comparison of every patient in the treatment arm with every patient in the control arm. Within each comparison, one evaluates the different components of the composite outcome in descending order of importance until the patient in the treatment arm is evaluated to have a better or worse outcome compared with the patient in the control arm. If the patient in the treatment arm has a better outcome compared to the control patient, it is called a "win"; if, on the other hand, the control patient has a better outcome, it is called a "loss". If a "win" or a "loss" cannot be established, it is called a "tie". Based on such prioritized outcomes and pairwise comparisons, the net benefit statistics [9] measure the difference between treatment arms. This win statistic [10] is defined as the difference in the proportion of wins between the two arms. The denominator of these proportions is the number of pairwise comparisons, thus accounting for ties. At times, there may not be an obvious prioritization of the outcomes. In such cases, a pairwise comparison and associated win difference is evaluated for each element of the composite outcomes and then an average or total win difference is derived [11]. This statistic is referred to as the adapted O'Brien statistic [12]. See a relevant discussion [13]. We call this averaged win difference statistic the non-prioritized net benefit.

In this paper, we compare the prioritized and non-prioritized net benefit statistics for a study design that utilizes a benefit-risk outcome as the primary outcome and where the risk assessment is via an AEI composite score. We further explore these two types of pairwise comparison methods through an example motivated by plaque psoriasis studies. In Section 2, we review the definition of the AEI composite score and the Wilcoxon-Mann-Whitney parameter. In Section 3, via simu-

lation, we evaluate the sample size requirement for the benefit-risk study designs to be assessed utilizing prioritized and non-prioritized net benefit statistics. In Section 4, a mathematical method for deriving sample size is presented, in the specific context of a dichotomous efficacy score and a continuous measure of safety, and details are presented in **Appendix**. In Section 5, an example based on a plaque psoriasis trial is discussed. Finally, in Section 6, conclusions are drawn.

2. The AEI Composite Score, the Wilcoxon-Mann-Whitney Test, and Non-Prioritized Net-Benefit

In assessing adverse events of interest (AEI), one might derive a safety score that incorporates the multidimensional aspects of AEI, called the AEI composite score [1]. For each AEI, including multiple recurrences of the event, the duration of the event is multiplied by the severity grade, and this value is summed up across the AEI of each patient to derive the safety score. The unit for duration can be days, weeks, or months, depending on the nature of the AEs and the study. See **Figure 1**. Although the choice of unit for AEI duration can influence the AEI score, its impact is limited when the assessment is based on prioritized or non-prioritized outcomes. Because the methodology relies on pairwise comparisons, it is the ranking of values, not their absolute magnitude, that ultimately drives the assessment. For patients with no events, the AEI composite score is zero. If there is no Common Terminology Criteria for Adverse Events (CTCAE) grading of severity, a grade of one, two and three can be used for mild, moderate, and severe AEs, respectively. The derivation of an AEI composite score assumes that the event is treatable, *i.e.*, it is not a permanent condition such as losing sight or death. If an AEI is a permanent condition, this can be added as a separate binary or ordinal endpoint to be compared via pairwise comparisons assessed via win-statistics such as net benefit.

Following other authors [1], our underlying assumption is that the adverse events (AEs) included in the AEI composite score are adverse reactions, and thus, causality has been established. The rationale for proposing the AE composite score, which sums events while accounting for severity and duration, is to provide a more comprehensive assessment of risk for better benefit-risk evaluation. This approach is similar in spirit to how adverse reactions are used in exposure-response models, where collections of adverse reactions are modeled (see, for example, [14]). It is important to note that the AEI composite score may not be applicable to all therapeutic areas. Furthermore, the AEI composite score implicitly assumes a linear trade-off between duration and severity (e.g., a grade 1 event for 30 days is equivalent to a grade 3 event for 10 days). Hence, it should be applied in therapeutic areas and development phases where accounting for the severity and duration of adverse reactions is relevant to the risk assessment. Furthermore, if certain AEI are considered to have differing impact and need to be analyzed separately, then for each category of AEI, a separate AEI composite score can be derived and included in the prioritized outcome.

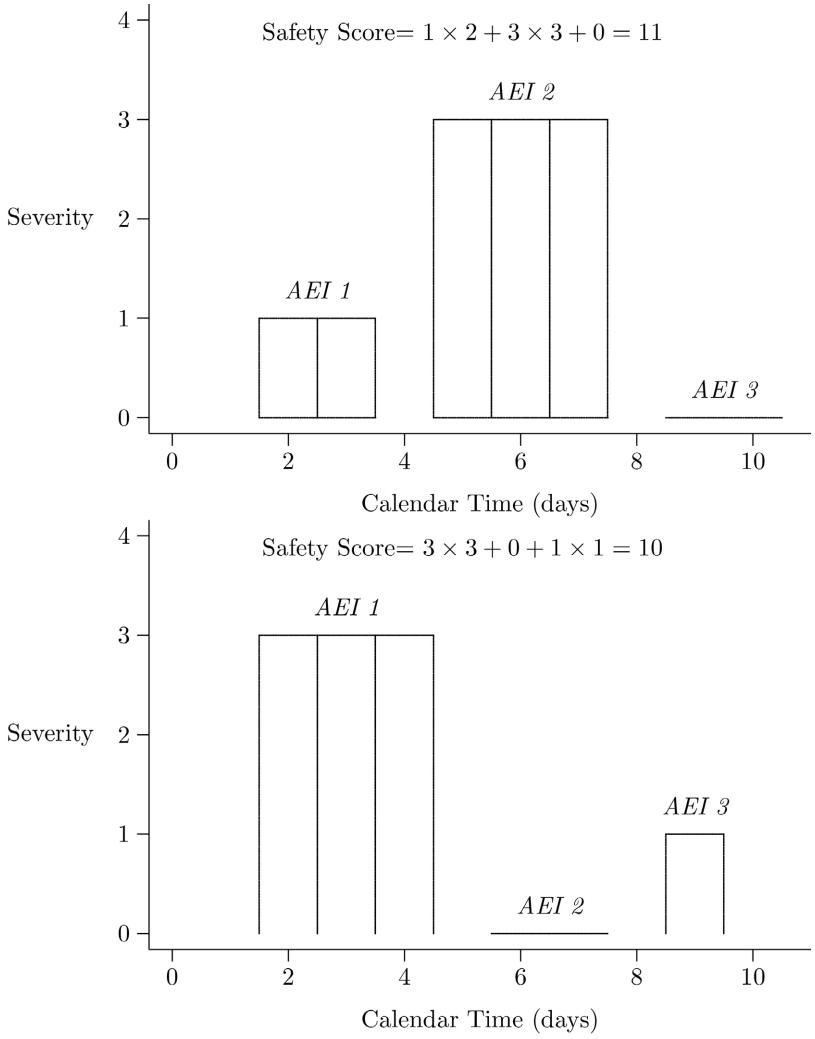

Figure 1. Adverse events of interest composite score.

Schemes for integrating benefit-risk assessment (efficacy and safety) can assess treatment benefit (efficacy) separately and risk (safety) separately and then integrate this information for the overall assessment of benefit-risk. In contrast, benefit-risk assessment can be made at a patient level and further summarized; for example, consider a prioritized outcome consisting of ordinal efficacy and safety outcomes. Furthermore, assume the ordinal efficacy outcome has the highest priority. In deriving the prioritized net benefit, when patients in a pair are being compared, they are first compared according to the efficacy outcome, and if they have the same efficacy value, they get compared over the safety outcome. In the following two subsections, these two approaches are discussed in the context of the composite outcomes consisting of a binary efficacy endpoint and AEI composite score. The associated statistical measures are the prioritized and non-prioritized net benefit scores. In Sections 3 and 4, we compare the prioritized and non-prioritized net benefit via simulation, and discuss new mathematical techniques for approximating power, and solving a power equation for sample size.

2.1. Prioritized Outcomes and the WMW Test

In the case of prioritized pairwise comparisons, the multivariate variables that define the prioritized outcomes can, under certain conditions, be transformed into a univariate rank-based ordering. For example, for a composite outcome consisting of a binary efficacy endpoint and the AEI composite score, a ranking can be achieved (see Section 3 for details). Ranking is an alternative to prioritization. Once ranking is achieved, the Wilcoxon-Mann-Whitney (WMW) test can be performed. Other authors discuss such rankings of prioritized outcomes [15]. The associated WMW parameter is defined as the probability that a randomly picked person in the treatment arm is ranked better than a randomly picked person in the control arm, plus half the probability that they are tied [16]. Hence, the null hypothesis of interest is whether this WMW parameter equals one-half. The *p*-values based on this test should be very similar to those obtained using net benefit or any win statistics (see Section 3 and [17]).

2.2. Non-Prioritized Outcomes and Net-Benefit

For study designs where the primary outcomes are benefit-risk composite outcomes and where risk is measured using an AEI composite score, it may not always be possible to find a definitive prioritization of the outcomes. Furthermore, depending on the severity of the illness and the treatment options that are available, one patient's prioritization of the benefit-risk outcomes may differ from another patient's prioritization. In such cases, it might be important to assess the prioritized net benefit under different types of prioritizations and without any prioritization. If the preference is not to prioritize the outcomes, a pairwise comparison and associated win difference are evaluated for each element of the composite outcomes and then an average or total win difference is derived [12]. One disadvantage of this method is that it evaluates separately benefit (efficacy) and risk (safety) and integrates the two measures via an average or total net benefit. In contrast, in the prioritized case, each treatment patient is compared with every control patient over the set of benefit-risk outcomes to see which patient in a pairwise comparison has a better prioritized benefit-risk outcome. The null hypothesis associated with the non-prioritized net benefit test is that there is no difference in the average or total win probabilities between the treatment and the control arms, while the alternate hypothesis is that there is a difference in this average or total win probabilities.

3. Assessment of Prioritized and Non-Prioritized Pairwise Comparison via Simulation

In this section, we compare the sample size requirements for a study where the primary outcomes are based on composite outcomes consisting of efficacy and treatment-and-disease-burden endpoints (AE composite score).

Following other authors [1], we assume the safety profile of the drug is established, *i.e.*, the AEI are well known and documented. Since these are adverse re-

actions, they are assumed to be caused by the drug. On the other hand, the drug is also expected to provide efficacy response. Furthermore, generally both efficacy and safety are associated with exposure levels; thus, some association between efficacy and AEI should be considered. Hence, efficacy outcomes and adverse events (*i.e.*, incidence, duration, and grade) will be simulated with positive dependence. This positive dependence can be achieved via a latent variable [1].

This manuscript considers the same simulation examples as presented earlier [1]. In brief, the sum of two exponential distributions is used to generate the binary efficacy outcomes, the duration of AEI and the grade of AEI. The latent variable that is used to create the correlation between these three variables has an exponential distribution that is part of the sum of the two exponentials that is used in generating all three variables. More precisely, let X be the efficacy component and Y be the latent variable used to construct the positive dependence between the efficacy and safety endpoints. We assume that X and Y are exponentially distributed with $\mathrm{E}[X]=\lambda_1$ and $\mathrm{E}[Y]=\lambda_2$; refer to the exponential distribution as with expectation λ as $\exp(\lambda)$. To generate n samples with the expected efficacy response rate of β, n samples are generated from $\exp(\lambda_1)$ and $\exp(\lambda_2)$; define the value G_β from $\beta=\mathrm{P}[X+Y\geq G_\beta]$; then subject i in the sample is a responder when $X_i+Y_i\geq G_\beta$ for $i\in\{1,\cdots,n\}$. To simulate the AEI duration, let Z be distributed as $\exp(\lambda_3)$, and n samples are generated from this distribution. For subject i, the resulting duration of the AEI is Y_i+Z_i where Y_i is the previously defined exponentially distributed component common to both efficacy and safety for subject i. Similarly, for the simulation of the severity grades, a sum of two exponential V_i+Y_i is used to generate the severity grades, where the Y_i is the same exponential distribution used to generate efficacy and duration samples. The severity grades are then assigned by first partitioning the unit interval into 5 intervals, corresponding to grades 0 (no-event), 1, 2, 3 and 4, and assigning the grades based upon which interval the sample Y_i+V_i falls. See **Table 1** for these parameterizations.

For this simplest case, where the efficacy outcome is binary, following an earlier example [1], we will assume that there are four AEI that make up the AE composite score and that will be included as part of the primary composite prioritized outcomes. The study consists of two arms, control, and treatment arms, with a 1-to-1 randomization (see **Table 2** and **Table 3**).

When outcomes can be prioritized, win-based statistics, such as the prioritized net benefit, allow assessing treatment effects. However, when the prioritized outcome structure allows for a ranking that mirrors the generalized pairwise statistic (GPS) framework, a Wilcoxon-Mann-Whitney (WMW) test can yield results comparable to GPS. This equivalence holds when the prioritization is such that the primary outcome is a binary endpoint and the secondary is a continuous measure. For the prioritized outcomes under discussion, a "win" for the treatment arm occurs if the treatment patient is a responder while the control is not, or if both are either responders or non-responders but the AEI composite score favors the treat-

ment arm patient. To operationalize this, one can rank all AEI composite outcomes within the responder and non-responder groups, ensuring that all the responders are ranked higher than the non-responders. Performing pairwise comparisons on these ranked values using the Mann-Whitney U-statistic is effectively equivalent to comparing the prioritized outcome via pairwise comparisons where the binary endpoint holds the highest priority. Consequently, the prioritized net benefit and the WMW test should yield similar if not identical *p*-values. It is important to note that if the prioritization is altered, such as making the AEI composite score the highest prioritized outcome, this ranking approach becomes significantly more complex. Such a ranking of composite outcomes is in line with the desirability of outcome rankings (DOOR) [15] and has been discussed by others [18]. For the WMW test, the null hypothesis is that the distribution of the derived variable of benefit-risk is the same between the two groups. Furthermore, the effect parameter associated with this test measures the probability that a randomly picked patient in the treatment arm has a better outcome compared to a randomly picked patient in the control arm, plus half the probability that they are tied [16]. For the non-prioritized outcome, we will use the non-prioritized net benefit. Furthermore, the null hypothesis associated with this test is that there is no difference in the average or total win probabilities between the treatment and the control arms, while the alternate hypothesis is that there is a difference in this average or total win probabilities between the two arms.

Table 1 and Table 2 present the distributional assumptions for the simulations [1]. For all four AEI, the same exponential distributions were used to generate the duration and severity of the AEI.

For the binary efficacy endpoint and AEI composite endpoint described in Table 1 and Table 2, the sample size requirements using prioritized net benefit have been presented earlier through simulation [1]. To demonstrate the equivalence between the prioritized net benefit and the Wilcoxon-Mann-Whitney (WMW) test, we will again derive the sample size requirements via simulation, this time using the WMW test. We will then compare these sample size requirements with the sample size requirements when the outcomes are not prioritized. The associated statistics is the non-prioritized net benefit.

Table 1. Efficacy and exponential distribution assumptions.

Efficacy response rate under the alternate hypothesis (Treatment, Control)	(0.45, 0.20)
Exponential distribution parameters	$\lambda_1 = 1$ (used to generate samples of efficacy outcomes) $\lambda_2 = 7$ days (common exponential distribution)* $\lambda_3 = 1$ day (used to generate samples of duration)* $\lambda_4 = 1$ (used to generate samples of severity)

*The duration of AEI is generated using the sum of two exponentials with parameters λ_2 and λ_3; hence, the mean duration for each AEI is 8 days.

Table 2. Assumed AE event rates for adverse events of interest.

AEI (overall event rate)*	Same AE rates for treatment and control arms (no event, Gr 1, Gr 2, Gr 3, Gr 4)	Different AE rates for treatment and control arms Treatment: (no event, Gr 1, Gr 2, Gr 3, Gr 4) Control: (no event, Gr 1, Gr 2, Gr 3, Gr 4)
AEI 1 (20%, 10%)	(0.80, 0.05, 0.05, 0.05, 0.05)	Treatment: (0.80, 0.05, 0.05, 0.05, 0.05) Control: (0.90, 0.025, 0.025, 0.025, 0.025)
AEI 2 (15%, 7.5%)	(0.85, 0.05, 0.05, 0.03, 0.02)	Treatment: (0.85, 0.05, 0.05, 0.03, 0.02) Control: (0.925, 0.025, 0.025, 0.015, 0.01)
AEI 3 (10%, 5.5%)	(0.90, 0.05, 0.03, 0.02, 0)	Treatment: (0.90, 0.05, 0.03, 0.02, 0) Control: (0.945, 0.025, 0.015, 0.01, 0.005)
AEI 4 (5%, 2.5%)	(0.95, 0.02, 0.01, 0.01, 0.01)	Treatment: (0.95, 0.02, 0.01, 0.01, 0.01) Control: (0.975, 0.01, 0.005, 0.005, 0.005)

AEI: Adverse Event of Interest. Gr: Grade. *Overall event rate for the treatment and control arms (e.g., for AEI 1 treatment arm 0.05 + 0.05 + 0.05 + 0.05 = 0.20, and control arm 0.025 + 0.025 + 0.025 + 0.025 = 0.1).

The sample size was derived under three assumptions. In the first case, for both the treatment and the control arms, the efficacy events are assumed to be positively correlated with the AEI events, duration, and severity. This would be the case if both the treatment and the control arms have the same mechanism of action. For the second case, for the treatment arm, a positive association is assumed between efficacy and the AEI events, duration, and severity, while for the control arm, there is no such association. This would be the case if the control arm is for example a placebo. Finally, for the third case, no association between efficacy and AEI events, duration, and severity is assumed for both the treatment and control arms. Since AEI are defined to be adverse reactions, this third case is not expected to happen in clinical studies. It is added here only for comparison purposes.

For the prioritized comparison, *i.e.*, for the WMW test and statistics, the R package asht [16] was used. For the non-prioritized net benefit, the R package BuyseTest [19] was used. For the simulation-based sample size derivation, the sample size associated with 90% power and with a two-sided type 1 error rate of 5% was iteratively searched. After this initial search, the final sample size derivation is based on 100,000 simulated studies.

In these simulation examples, for the case of prioritized pairwise comparisons, the prioritization of the composite outcomes is that the binary efficacy endpoint has higher priority than the AEI-derived score endpoint. **Table 3** presents the results of the sample size derivations. In the cases where both the treatment and the control arm have the same AEI rates, the non-prioritized net benefit statistics has a lower or similar sample size requirement compared to the prioritized case. In contrast, when the control arm has lower AEI rates compared to the treatment arm, the sample size requirement for the non-prioritized net benefit statistics is

larger. This is to be expected since the non-prioritized net benefit gives an equal weight to both endpoints and the treatment effect for efficacy and safety are in opposite directions. In the case of the prioritized outcome (WMW test), the safety comparison is only made after the comparison with efficacy is made, reducing the impact of the safety endpoint in the derivation of the WMW statistics. On the other hand, if the prioritization is reversed (AEI composite score highest priority and binary endpoint lower priority), the sample size requirements for the prioritized case will be higher than or equal to the non-prioritized case.

Prioritized and non-prioritized net benefits have been compared through simulation. Simulation examples that are based on mucopolysaccharidosis type IIIA (MPS IIIA) disease had finding that prioritized net benefit generally had more power than the non-prioritized net benefit due to the largest treatment effects being captured by the highest priority outcome [13]. Simulations of prioritized and non-prioritized net benefits through involving two time-to-event endpoints with varying correlations and treatment effects, as well as a time-to-event and continuous endpoint with varying correlations and treatment effects, found that the non-prioritized net benefit tended to have greater power when the treatment effect was dominated by the second prioritized outcome [12]. When both outcomes had similar treatment effects, the prioritized and non-prioritized net benefits had similar power. These results are similar to the results presented in **Table 3**.

Table 3. Simulation-based sample size derivation.

Case	Distributional assumption	Non-prioritized net benefit		WMW Test	
		N/arm	Power	N/arm	Power
AEI rates: Same rates for Treatment and Control arms[1]					
1	Trt: Efficacy is positively associated with AEI Cont: Efficacy is positively associated with AEI	32	90.37%	50	90.28%
2	Trt: Efficacy is positively associated with AEI Cont: Efficacy is independent from AEI	36	90.62%	32	90.36%
3	Trt: Efficacy is independent from AEI Cont: Efficacy is independent from AEI	161	90.23%	105	90.28%
AEI rates: Different rates for Treatment and Control arms[2]					
4	Trt: Efficacy is positively associated with AEI Cont: Efficacy is positively associated with AEI	128	90.52%	70	90.07%
5	Trt: Efficacy is positively associated with AEI Cont: Efficacy is independent from AEI	87	90.83%	50	90.36%

AEI: Adverse Events of Interest; Cont: Control; Trt: Treatment. Each sample size derivation and power calculation is based on 100,000 sample studies. [1]AEI rates for treatment and control arms are as in Column 2 in **Table 2**. [2]AEI rates for treatment and control arms are as in Column 3 in **Table 2**.

As expected, for this example, the resulting sample size requirement presented for the WMW (derived from the prioritized composite outcomes) is the same as what was obtained when utilizing the prioritized net benefit test statistics [1].

For the WMW test, we have derived the sample size requirement using both the WMW test and the WMW parameter 95% confidence interval (CI). The WMW parameter CI derivation [16] is under the proportional odds assumption. In the example under consideration, since the non-zero AEI events are low, there were several patients that had tied ranks. Even in this case, we have found that the WMW test and associated CI result in similar significance rates and hence, in similar simulation-based sample size requirements.

These examples demonstrate that if a study is to be designed using the prioritized or non-prioritized net benefit statistics, the sample size requirement will be different based on the prioritization we select. At the end of a study, there is also an interest in evaluating the benefit-risk of the drug in preparation for submission to a health authority or for further evaluating the therapy for further development. In such cases, one needs to note the different sample size requirements when evaluating benefit-risk using different types of prioritizations of the outcomes.

4. Mathematical Sample Size Derivation for BR Composite Outcomes to Be Assessed via Prioritized and Non-Prioritized Net Benefit Statistics

4.1. WMW Test/Prioritized Net Benefit

In this section, for the examples in Section 3, we present sample size estimates that are based on a mathematical derivation.

Table 4. Sample size for ranked observations using the formula from [20] vs. simulation-derived sample size.

	$P[Y_{trt} > Y_{ctl}] + 0.5 \times P[Y_{trt} = Y_{ctl}]$	Sample size/arm simulation-based	Sample size/arm under the symmetric distribution assumption
AEI rates: Same rates for Treatment and Control arms			
Case 1	0.652	50	76
Case 2	0.710	32	40
Case 3	0.625	105	113
AEI rates: Different rates for Treatment and Control arms			
Case 4	0.629	70	106
Case 5	0.663	50	66

Y_{trt} and Y_{ctl} are the ranked prioritized outcomes for the treatment and control arms respectively.

For the prioritized pairwise comparison, due to one of the components being a binary endpoint, the problem can be reduced to the univariate pairwise comparison, which can then be tested using the WMW test. Hence, the sample size derivation can also be done utilizing the method developed for the analysis via WMW. In this case, the null hypothesis would be that the probability that a randomly picked patient in the treatment arm does better than a randomly picked control

arm, plus half the probability that they are tied, is half. The sample size derivation assumes that the underlying distribution for both arms is symmetric and requires the value of the probabilities of win, loss, and tie, under the alternate hypothesis [20]. See the discussion [21]. **Table 4** shows the power calculation for the examples in Section 3. In each case, this formula provides a sample size that is higher than what is observed using simulation. The composite AEI variable is zero for most values and the distribution of the non-zero values tends to be skewed. Hence, the underlying distribution of the values that are ranked is not symmetrically distributed. The difference between the simulation-based method and the mathematical derivation becomes larger as the probability of win minus the probability of loss is close to the null value of zero.

4.2. Non-Prioritized Net Benefit

We extended the sample size derivation of [22] to the case where the statistic of interest is the non-prioritized net benefit.

This non-prioritized test statistic is constructed by considering pairs of control and treatment subjects; let i and j index the control and treatment subjects respectively. Let V_{ij} be 1 if the efficacy response for treatment subject j is larger than that of control subject i, −1 if the efficacy response for treatment subject j is smaller than that of control subject i, and 0 if they are tied. Let W_{ij} be 1 if the adverse event response for treatment subject j is more severe than that of control subject i, −1 if the adverse event response for treatment subject j is less severe than that of control subject i, and 0 if they are tied. Then the non-prioritized test statistic is $T = \sum_{i,j}\left(V_{ij} - W_{ij}\right)$. For a large number of control and treatment subjects, the distribution of T is approximately normal.

Under the null hypothesis of no difference between treatment and control subjects, the distributions of V_{ij} and W_{ij} are symmetric about zero, and so the null expectation of T is zero. The test has level α. Let $\mathrm{V}_0[T]$, $\mathrm{E}_A[T]$, and $\mathrm{V}_A[T]$ represent the null variance and alternative expectation and variance of T respectively, and consider one-sided alternatives with $\mathrm{E}_A[T] > 0$. Let Φ represent the standard normal distribution function, and let z_α be the $1-\alpha$ quantile of the standard normal quantile; that is, $\Phi(z_\alpha) = 1-\alpha$, and $z_{0.025} = 1.96$. The standard formula for approximate power for a test statistic that is approximately normal under both the null and alternative distributions is

$$\text{Power} = \Phi\left(\mathrm{E}_A[T] \Big/ \sqrt{\mathrm{V}_A[T]} - \sqrt{\mathrm{V}_0[T]}\, z_\alpha \Big/ \sqrt{\mathrm{V}_A[T]}\right). \tag{1}$$

If one approximates the alternative variance as equal to the null variance, then the approximate power simplifies to

$$\text{Power} = \Phi\left(\mathrm{E}_A[T] \Big/ \sqrt{\mathrm{V}_0[T]} - z_\alpha\right). \tag{2}$$

The standard approximate sample size for this test to obtain power $1-\beta$ is determined by representing $\mathrm{V}_0[T]$ as approximately linear in inverse sample size per arm, and solving either (1) or (2) for this common sample size. Formulas

for moments of T under either hypothesis are given in **Appendix** to this paper.

Since AEI are relatively rare, the AEI composite score is expected to have a large mass at zero. Hence, the above sample size derivation is for the composite endpoint where one is a binary endpoint and the second endpoint is a truncated continuous distribution, *i.e.*, it has a positive mass at zero and has a positive mass for values larger than zero. Cases without a positive mass at zero are accommodated by setting the probability at zero to zero. Other cases, including time-to-event outcomes, may be addressed similarly in future work. Table 5 shows sample sizes that would provide approximately 90% power.

Table 5. Sample sizes for the five simulation design cases.

Case	N per arm	Probability of AEI > 0, conditional on				Mean of non-zero derived safety score (no eff event[1], eff event[2])	
		No efficacy event		Efficacy event		Treatment arm	Control arm
		Treatment arm	Control arm	Treatment arm	Control arm		
1	32	0.9998	0.9787	0.5513	0.07487	(0.03330, 30.00)	(5.7565, 30.00)
2	36	0.9997	0.5816	0.5501	0.58150	(0.03735, 30.00)	(21.0590, 19.92)
3	161	0.5819	0.5810	0.5815	0.58180	(21.01531, 21.04)	(21.0049, 21.12)
4	128	0.9997	0.9998	0.5494	0.49495	(0.11932, 30.00)	(0.2275, 30.00)
5	87	0.9998	0.7667	0.5514	0.76852	(0.08078, 30.00)	(19.6403, 19.30)

[1]No eff event: Mean of the non-zero part of the AEI score when there is no efficacy event. [2]Eff event: Mean of the non-zero part of the AEI score when there is an efficacy event.
In all cases, efficacy for treatment and control groups is 0.45 and 0.20 respectively.

Case	Original sample size	Approximate power via Formula (1) (%)	Approximate power via Formula (2) (%)	Approximate sample size for 90% power
1	32	99.99	98.39	18
2	36	82.95	87.61	39
3	161	92.99	92.00	150
4	128	97.72	97.39	88
5	87	86.19	90.77	85

Our sample size derivation requires assumptions on the efficacy probability and the probability of the AEI score being positive in the presence and absence of an efficacy event. To derive the sample size, we treat efficacy as a Bernoulli trial, and hence one must specify the two probabilities of success for the control and treatment arms. We treat the adverse event process as conditional on the efficacy process, and treat this process as continuous, with an additional point mass at zero. Hence, in order to specify the alternative, one must specify the four conditional probabilities of a zero adverse event value, and the four densities of the non-zero part of the adverse event. Our simulations take the positive part of the adverse event score as exponential, with four expectations that the user must specify. An-

alytic calculations are quite long, and are presented in **Appendix** to this paper. Table 5 presents these probabilities estimated via simulations for the simulation example cases presented in Tables 1-4.

The upper panel lists probability inputs for each of these simulations. The lower panel gives the asymptotic approximation to the power for this model, and the approximate sample size for 90% power.

Compare these powers to the simulation approximation to the true power in Table 3. In three of these cases, the second, third, and fifth cases, the asymptotic approximation (2) to the target power is quite accurate. In the other two cases, the approximation is less accurate, since the extreme adverse event duration rates result in the exponential model failing to fit. In these cases with extreme differences in adverse event duration means, our approximation as implemented performs poorly. Work is in progress to fit a more sophisticated model than the exponential model.

5. Plaque Psoriasis Therapy Example

The example in this section is motivated by the two plaque psoriasis pivotal studies comparing the immunotherapy Risankizumab with Ustekinumab and matching placebo [23]. In both trials the randomization was 3:1:1 ratio in favor of Risankizumab. GPC has been used for assessing treatment benefit via multiple prioritized outcomes, including PRO (patient-reported outcome) endpoints [24] [25]. Similarly, in our assessment of the treatment benefit of Risankizumab relative to Ustenkinumab, we will include multiple efficacy outcomes and PRO outcomes. In these trials, in many of the efficacy and patient-reported outcome (PRO) measures, the Risankizumab-treated patients had on average a numerically better outcome. In this example, we will use the approximate average observed efficacy and PRO results for these two pivotal studies. Table 6 summarizes the response rates and mean efficacy and PRO endpoint values from the two studies (UltIMMa-1 and UltIMMa-2 [24]), as well as the resulting approximate average values applied in this illustrative example.

Table 6. Summary statistics from the two pivotal studies (UltlMMa-1 and UltlMMa-2) comparing Risankizumab and Ustekinumb.

	UltlMMa-1*		UltlMMa-2*		Average values used in the simulation in Cases 1 & 2	
Parameter at week 16	Risankizumab	Ustekinumb	Risankizumab	Ustekinumb	Treatment	Control
N	304	100	294	99	NA	NA
PSAI 90	75.3%	42.0%	74.8%	47.5%	75%	45%
sPGA (0 or 1)	87.8%	63.0%	83.7%	61.6%	85%	62%
DLQI 0 or 1	65.8%	43.0%	66.7	46.5	66%	45%
Change in PSS at week 16 (SE)	−5.6 (0.2)	−4.4 (0.3)	−6.4 (0.2)	−5.6 (0.3)	−6	−5

*Values reproduced from [23], SE = Standard Error.

For safety, two of the highlighted categories of AEs are severe AEs and infections [23]. In our analysis, we will use these categories for our safety burden assessment. However, the frequencies of these events have been altered from what is presented in the label and in [23]. The severe AEs and the infections are assumed to be not overlapping, *i.e.*, severe AEs are severe AEs that are not infections. In the simulations, a patient can only have one infection or severe AE. In our examples, we have considered three cases: 1) efficacy and PRO effects similar to those observed in the comparison between Risankizumab and Ustekinumab [23] and with AEI favoring the control arm; 2) the same efficacy and PRO effects as in Case 1), but with AEI favoring the treatment arm; and 3) reduced efficacy and PRO effects compared to Case 1), while maintaining the same AEI effect as in Case 1).

Table 7 presents the assumptions for the efficacy, safety (severe AEs and infection), and PRO endpoints.

Table 7. Assumed efficacy, PRO, safety values (motivated by data in [23]).

Measure	Case 1	Case 2	Case 3
	(Treatment, Control)	(Treatment, Control)	(Treatment, Control)
Number per group (N)	(100, 100)	(100, 100)	(100, 100)
PASI 90 at week 16 (probability)	(0.75, 0.45)	(0.75, 0.45)	(0.75, 0.6)
sPGA 0 or 1 at week 16 (probability)	(0.85, 0.62)	(0.85, 0.62)	(0.85, 0.735)
DLQI 0 or 1 at week 16 (probability)	(0.66, 0.45)	(0.66, 0.45)	(0.66, 0.555)
Change in PSS at week 6 (mean, standard deviation*)	[(−6, 3.46), (−5, 2.99)]	[(−6, 3.46), (−5, 2.99)]	[(−6, 3.46), (−5.5, 2.99)]
Infection (no event, Gr 1, Gr 2, Gr 3, Gr 4)	[(0.80, 0.05, 0.05, 0.05, 0.05), (0.90, 0.025, 0.025, 0.025, 0.025)]	[(0.90, 0.025, 0.025, 0.025, 0.025), (0.80, 0.05, 0.05, 0.05, 0.05)]	[(0.80, 0.05, 0.05, 0.05, 0.05), (0.90, 0.025, 0.025, 0.025, 0.025)]
Severe AEs (no event, Gr 1, Gr 2, Gr 3, Gr 4)	[(0.95, 0.0, 0.0, 0.03, 0.02), (0.99, 0.0, 0.0, 0.005, 0.005)]	[(0.99, 0.0, 0.0, 0.005, 0.005), (0.95, 0.0, 0.0, 0.03, 0.02)]	[(0.95, 0.0, 0.0, 0.03, 0.02), (0.99, 0.0, 0.0, 0.005, 0.005)]

*Standard deviation is derived using the mean-squared errors presented in [23]. DLQI = Dermatology Life Quality Index, PASI = Psoriasis Area and Severity Index, PSS = Psoriasis Symptom Scale, sPGA = static Physician's Global Assessment.

The assumed time-point of the analysis is Week 16. For efficacy, we considered the binary efficacy endpoints of achieving Psoriasis Area and Severity Index (PASI) reduction of 90% or more and achieving the static Physician's Global Assessment (sPGA) score of 0 or 1. For PRO endpoints, we considered the binary endpoint of achieving the Dermatology Life Quality Index (DLQI) score of 0 or 1. For the PRO Psoriasis Symptom Scale (PSS), the continuous value of change from baseline of the PSS score was used. The efficacy and PRO endpoints were simulated using a multivariate normal distribution where the PASI, sPGA and DLIQI corresponding variables are generated from the standard normal distribution, while the mean change from baseline and associated standard deviation observed

are used to generate the PSS endpoints. See Table 6 and Table 7 for the assumed response rates for the binary endpoints and for the continuous PRO endpoint (PSS). The correlation between PASI and sPGA is assumed to be 0.7, while the correlation for all other endpoints is assumed to be 0.5. Once the multivariate data are simulated, the quantiles are used to generate the binary endpoints from this multivariate simulated data. For example, for the new treatment arm, to generate binary PASI 90 data that has a response rate of 85%, all the corresponding Psoriasis Area and Severity Index (PASI) values from the marginal standard normal distribution are set to response if the simulated value is less than or equal to the standard normal distribution quantile corresponding to a cumulative probability of 85%, *i.e.*, if the values are less than or equal to 1.0364. For the safety endpoint, an AEI composite score is derived from infections, or severe AEs that are not infections using exponential distributions as in the examples discussed in Section 3.

Table 8. Simulated study under Case 1 assumptions (100 patients per arm).

Winners and treatment effect according to prioritized scoring system (%)					
Endpoint	Trt win	Control win	Neutral	Proportion of effect	Cumulative effect
PASI 90	36.85	14.85	48.30	22.00	22.00
sPGA 0 or 1	10.56	4.50	33.24	6.06	28.06
DLQI 0 or 1	11.60	3.58	18.06	8.02	36.08
Change from baseline PSS	10.26	7.80	0.00	2.46	38.54
AEI derived score	0.00	0.00	0.00	0.00	38.54
Winners and treatment effect according to non-prioritized scoring system (%)					
Endpoint	Trt win	Control win	Neutral	Proportion of effect	Cumulative effect
AEI derived score	8.90	22.70	68.40	−13.80	−2.76
PASI 90	36.85	14.85	48.30	22.00	1.64
sPGA 0 or 1	35.20	11.20	53.60	24.00	6.44
DLQI 0 or 1	41.48	12.48	46.04	29.00	12.24
Change from baseline PSS	61.64	38.36	0.00	23.28	16.90

The prioritization of the endpoints might be different from the perspective of patients, payers, and health authorities. Even among patients, the prioritization of the endpoints may be different depending on the severity of the patient's illness. For example, a patient with severe plaque psoriasis may be willing to tolerate more serious adverse events in favor of a therapy with greater efficacy. In a prioritized net-benefit analysis, efficacy outcomes would be given the highest priority for such patients. Conversely, a patient with mild disease might prefer a therapy with a better safety profile, even if it offers less efficacy, prioritizing safety outcomes instead. In view of this, there is interest in assessing differences in benefit-risk using efficacy, PRO, and safety endpoints by utilizing three different prioritiza-

tions of these composite endpoints. For the three cases that were considered, initially we applied the following order of priority from highest to lowest: 1) PASI 90, 2) sPGA 0 or 1, 3) DLQI 0 or 1, 4) PSS change from baseline endpoint, and 5) AEI composite score constructed from severe AEs and infection events. The prioritized net benefit statistics was then used to assess the treatment difference in win proportions. Second, we reversed the priority of the safety endpoint. The AEI composite score, instead of being the last prioritized endpoint, became the highest prioritized outcome. After the AEI composite score, the same prioritization is kept for the efficacy and PRO endpoints. Lastly, equal weights were used for all the endpoints. Table 8 presents the assessment of one simulated study under Case 1 assumptions, using both prioritized and non-prioritized net benefit. In both analyses, the contribution of PASI 90 endpoint is considerably higher than that of other variables. The impact is greater when it is the highest-prioritized endpoint.

Table 9. Simulation summary for Case 1.

Outcomes[1]	Average cumulative treatment effect (2.5 and the 97.5 percentile of the simulation)	Average percent of neutral pairwise comparisons
Prioritization: 1) PASI 90, 2) sPGA 0 or 1, 3) DLQI 0 or 1, 4) Change from baseline PSS, 5) AEI derived score		
PASI 90	29.97 (17.0, 43.0)	47.50
sPGA 0 or 1	34.54 (20.8, 47.8)	34.58
Overall	38.09 (23.22, 52.4)	
Prioritization: 1) AEI-derived score, 2) PASI 90, 3) sPGA 0 or 1, 4) DL QI 0 or 1, 5) Change from baseline PSS		
AEI-derived score	−11.41 (−21.7, −1.2)	70.12
PASI 90	9.58 (−5.3, 24.8)	33.33
Overall	15.3 (−0.9, 31.4)	
Non-prioritized net benefit*		
AEI-derived score	−2.27 (−4.3, −0.03)	70.15
PASI 90	3.74 (0.4, 7.0)	47.47
Overall	16.03% (7.9, 24.1)	

[1]Average cumulative effect presents the average of 10,000 simulation treatment effects for the highest priority and the cumulative treatment effect of the two highest priorities. The average percentage of neutrality presents the average of 10,000 percentage of pairwise comparison that results in neutral outcome for the highest priority and after the two highest priority comparisons. *For non-prioritized net benefit, the cumulative treatment effect is averaged over the number of outcomes, *i.e.*, scaled by 5. For neutral pairwise comparisons, the average percentage of neutral outcomes after comparison with the outcome is presented.

Trials with a sample size of 100 patients per arm were generated 10,000 times (see Table 7 for the distribution of the endpoints). For Case 1, Table 9 summarizes the results of the three types of pairwise comparisons, by presenting the mean win

proportion difference and the 2.5th and 97.5th percentiles of the win proportion difference (for non-prioritized pairwise comparisons, it is the average win proportion).

From the results of this analysis, depending on the prioritization of the composite endpoints, an endpoint may not be relevant, *i.e.*, not used at all. The change from baseline PSS endpoint is a continuous endpoint; hence, ties are unlikely to occur when a pair is compared using this endpoint. For such continuous variables, a clinically meaningful difference is more relevant than a mere numerical difference. Therefore, pairs should be classified as whether the values differ by more than the predefined clinical threshold. The use of such a threshold was proposed [9] and has been studied for time-to-event endpoints [8].

The AEI-derived score is zero for most patients, but when it is non-zero, it also is a continuous variable. Hence, if the AEI-derived score is given the highest priority, it may lead to the continuous PSS score not being used in any comparison. In our simulation, this was not the case. For Case 1, when the PASI 90 endpoint has the highest priority, on average more than 50% of the pairwise comparisons result in a win or a loss; hence, the net-benefit is dominated by this endpoint. See **Table 9**. In contrast, when the AEI composite score has the highest priority, on average, approximately 30% of the pairwise comparisons result in a win or a loss, while the first and second highest priority together result on average in just 33% being neutral after these two comparisons. This indicates the great impact the PASI 90 endpoint has on the prioritized net-benefit statistics.

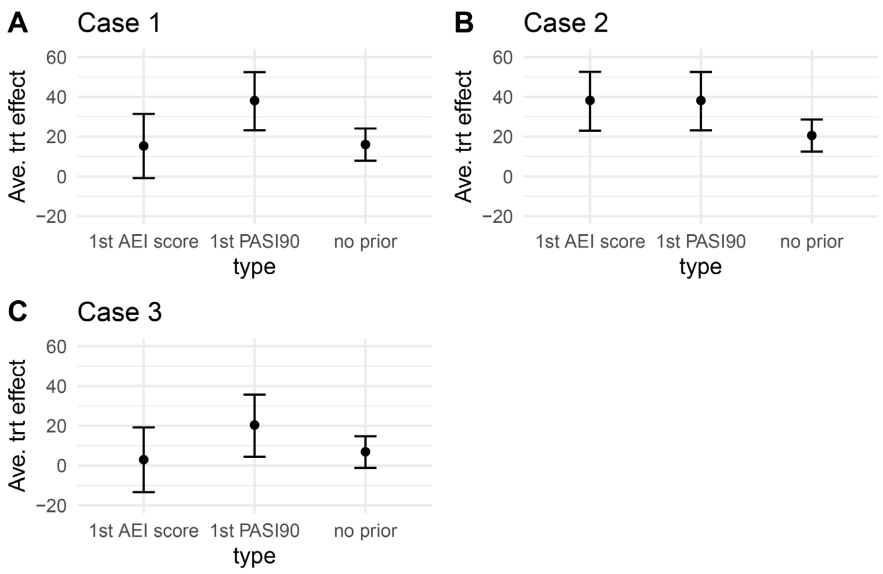

Figure 2. The average treatment effect measured by prioritized and non-prioritized net benefit and associated 2.5th and 97.5th percentiles.

The simulation results for Cases 1 - 3 are presented graphically in **Figure 2** and **Figure 3**. In all three cases, when PASI 90 is the highest prioritized outcome, the average and the 2.5th percentiles are consistently above "0" (see **Figure 2**). This is

because, on average, more than 40% of the comparisons result in a "win" or "loss" (see Figure 3), driving the net benefit by this outcome's comparison. Changing prioritization has a significant impact in Case 1 and Case 3 (see Figure 2), as the AEI score favors the control while all efficacy and PRO outcomes favor the treatment arm, reflecting the complex relationship between the outcomes and the prioritization. The operating characteristics of prioritized net benefit for cases where there are two prioritized outcomes have been assessed via simulation in [13]. For these examples, the overall power of the prioritized net-benefit depends, in a complex manner, on the entire variance-covariance structure of the set of outcomes.

Additionally, in all three cases, the non-prioritized net benefit showed the smallest variability (see Figure 2). In all three cases, the non-prioritized net-benefit was more sensitive than the prioritized net-benefit, when the AEI score had the highest priority. Similar observations were made in comparisons of the prioritized and non-prioritized net benefit, via simulation [12] and [25] (see Section 3). However, as stated in Section 2.2, the prioritized net benefit assesses the benefit-risk variables at the patient level.

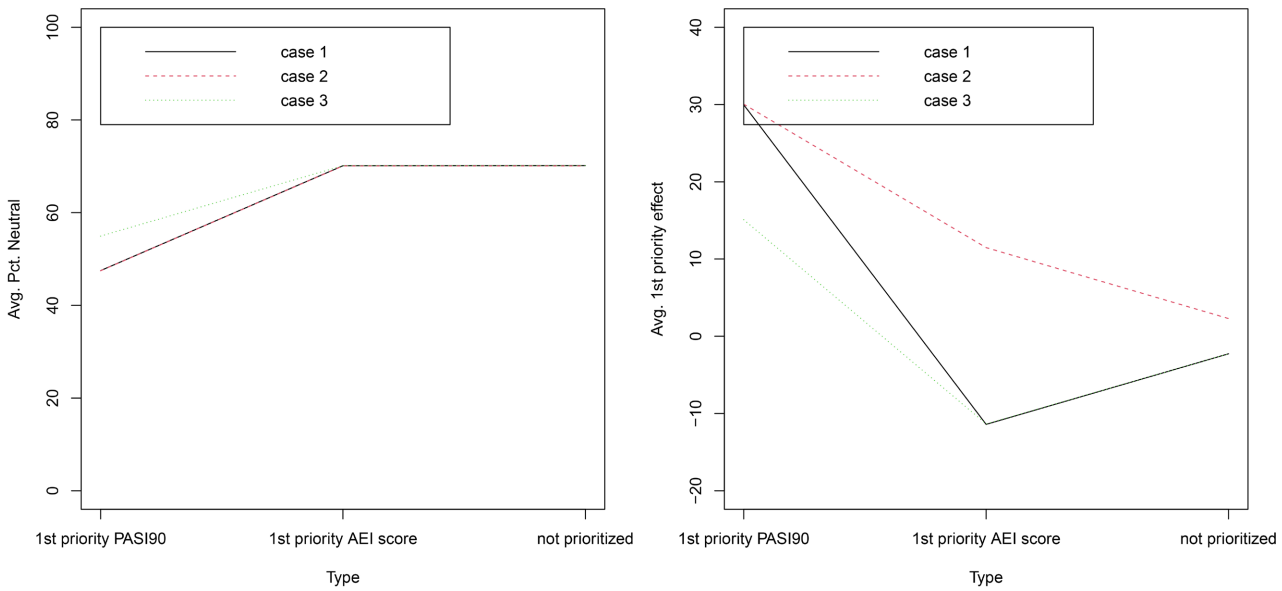

Figure 3. For the prioritized net benefit, average percentage of neutral comparisons and the first prioritized treatment effect. For non-prioritized net benefit, average percentage of neutral effects and the mean effect based on the AEI composite score.

The results of these three comparisons show that depending on the prioritization of the benefit and risk endpoints, the treatment effects differ.

6. Conclusions

BR assessment is an important aspect of drug evaluation for pharmaceutical companies, health authorities, payers and even patients. Furthermore, different patient groups may have different therapy needs. For example, a patient who has a severe form of a disease might be willing to accept a higher chance of a safety risk

associated with a therapy, while this may not be the case for a patient who has a less severe form of the disease. BR assessment using GPC methodologies has been discussed in [1] [4]-[6]. BR might be evaluated at a patient level utilizing prioritized composite endpoints consisting of efficacy and safety endpoints, where safety is assessed using the adverse events of interest (AEI) composite score [1]. In this paper, we considered both the prioritized and non-prioritized forms of the proposed BR endpoint.

Via the simulation example in [1], we have shown that the sample size requirement may differ depending on whether we use a prioritized or non-prioritized BR endpoint. As expected, when the efficacy benefit favors the treatment arm and the safety benefit favors the control arm, the sample size requirements for the non-prioritized outcome increase substantially.

For the BR outcomes under consideration, we also have demonstrated how sample size can be derived using mathematical formula. For the prioritized case, we can use a widely available sample size derivation for the WMW test. For the non-prioritized net benefit, we have provided an approximation to the power, as in (2). This sample size derivation is geared to the more complex AEI composite score endpoint. The sample size derivation takes into consideration the various probabilities of having a safety event and their association with the efficacy events. In cases with moderate AEI probabilities and rates for duration, the more complicated model allowing multiple multi-level AEI is well-approximated with the simpler single-level model, and power is well approximated asymptotically. In more extreme cases, the complicated model is poorly approximated by the single-event single-level exponential model. More work needs to be done to fit this distribution better. Thus, this reliance on the simple exponential model is a drawback of our method, and an opportunity for extension. An area for extension is the use of more general gamma distributions, or even heavier-tailed distributions such as the Pareto.

Finally, via an example based on plaque psoriasis, we demonstrated how the effect sizes and the direction of the effect could differ depending on the prioritization of the outcomes. Other researchers report consistent observations [12] [13] [25]. It is important to note that when using pairwise comparisons for prioritized composite endpoints, depending on the nature of the variables (e.g., normally distributed) and order of the prioritization, a variable may not be relevant. However, such cases are expected to be rare. Generally, for continuous variables, clinically relevant thresholds are defined [9], which may result in ties.

When assessing BR using pairwise comparisons of prioritized composite endpoints such as net benefit, it is important to consider diverse types of prioritizations, especially when the order of the prioritization is not clear. In the case of prioritized net benefit, the prioritization accounts for the relationship between risk (safety) and benefit (efficacy) at the patient level. In contrast, for the average net benefit, marginal risks and benefits are assessed and hence can potentially provide a misleading benefit-risk assessment [4] [26]. Through the examples we considered, depend-

ing on the prioritization of the composite endpoint, the sample size requirement for demonstrating treatment difference may be different. The sample size requirements for the non-prioritized net benefit can be assessed using the formula we have provided (available via the AESim R package).

One limitation of the AEI composite score used in assessing risk in our BR assessment is that it is calculated by multiplying the duration of an event by its severity grade and summing these products. This approach implicitly assumes a linear trade-off between duration and severity. For example, a grade 1 event lasting 30 days is treated as equivalent to a grade 3 event lasting 10 days, which may not reflect clinical judgment. A prolonged grade 1 event may impose minimal burden on the patient, whereas a short grade 4 event may require hospitalization. A potential solution for this issue would be partitioning AEI into two composite scores: one based on lower-grade events (e.g., grades 1 - 2) and another based on higher-grade events. In a prioritized net-benefit framework, the composite score based on higher-grade events can be assigned a higher priority than the composite score based on lower-grade events.

We provided useful tools for planning studies with non-prioritized endpoints with multiple multi-level AEI; these tools approximate moments of the appropriate test statistics and calculate sample size and power from a normal approximation. The sample size derivation that is provided is for any non-prioritized composite endpoint consisting of a binary and continuous (for example, exponentials with a point mass at zero) endpoints.

Acknowledgements

The authors would like to thank Dr. Jeff Maca for his insights on the framework of incorporating severity and duration into the assessment of adverse events. This material is based in part upon work supported by, and while John Kolassa was serving at, the National Science Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Seifu, Y., Mt-Isa, S., Duke, K., Gamalo-Siebers, M., Wang, W., Dong, G., *et al.* (2022) Design of Paediatric Trials with Benefit-Risk Endpoints Using a Composite Score of Adverse Events of Interest (AEI) and Win-Statistics. *Journal of Biopharmaceutical Statistics*, **33**, 696-707. https://doi.org/10.1080/10543406.2022.2153202

[2] Péron, J., Roy, P., Ding, K., Parulekar, W.R., Roche, L. and Buyse, M. (2015) Assessing the Benefit-Risk of New Treatments Using Generalised Pairwise Comparisons: The Case of Erlotinib in Pancreatic Cancer. *British Journal of Cancer*, **112**, 971-976. https://doi.org/10.1038/bjc.2015.55

[3] Péron, J., Giai, J., Maucort-Boulch, D. and Buyse, M. (2019) The Benefit-Risk Balance

of Nab-Paclitaxel in Metastatic Pancreatic Adenocarcinoma. *Pancreas*, **48**, 275-280. https://doi.org/10.1097/mpa.0000000000001234

[4] Buyse, M., Saad, E.D., Peron, J., Chiem, J., De Backer, M., Cantagallo, E., *et al.* (2021) The Net Benefit of a Treatment Should Take the Correlation between Benefits and Harms into Account. *Journal of Clinical Epidemiology*, **137**, 148-158. https://doi.org/10.1016/j.jclinepi.2021.03.018

[5] Backer, M.D., Sengar, M., Mathews, V., Salvaggio, S., Deltuvaite-Thomas, V., Chiêm, J., *et al.* (2023) Design of a Clinical Trial Using Generalized Pairwise Comparisons to Test a Less Intensive Treatment Regimen. *Clinical Trials*, **21**, 180-188. https://doi.org/10.1177/17407745231206465

[6] Piffoux, M., Ozenne, B., De Backer, M., Buyse, M., Chiem, J. and Péron, J. (2024) Restricted Net Treatment Benefit in Oncology. *Journal of Clinical Epidemiology*, **170**, Article ID: 111340. https://doi.org/10.1016/j.jclinepi.2024.111340

[7] Buyse, M., Verbeeck, J., Saad, E.D., Backer, M.D., Deltuvaite-Thomas, V. and Molenberghs, G. (2025) Handbook of Generalized Pairwise Comparisons: Methods for Patient-Centric Analysis. Chapman and Hall/CRC.

[8] Péron, J., Roy, P., Ozenne, B., Roche, L. and Buyse, M. (2016) The Net Chance of a Longer Survival as a Patient-Oriented Measure of Treatment Benefit in Randomized Clinical Trials. *JAMA Oncology*, **2**, 901-905. https://doi.org/10.1001/jamaoncol.2015.6359

[9] Buyse, M. (2010) Generalized Pairwise Comparisons of Prioritized Outcomes in the Two-Sample Problem. *Statistics in Medicine*, **29**, 3245-3257. https://doi.org/10.1002/sim.3923

[10] Dong, G., Huang, B., Wang, D., Verbeeck, J., Wang, J. and Hoaglin, D.C. (2020) Adjusting Win Statistics for Dependent Censoring. *Pharmaceutical Statistics*, **20**, 440-450. https://doi.org/10.1002/pst.2086

[11] Ramchandani, R., Schoenfeld, D.A. and Finkelstein, D.M. (2016) Global Rank Tests for Multiple, Possibly Censored, Outcomes. *Biometrics*, **72**, 926-935. https://doi.org/10.1111/biom.12475

[12] Verbeeck, J., Spitzer, E., de Vries, T., van Es, G.A., Anderson, W.N., Van Mieghem, N.M., *et al.* (2019) Generalized Pairwise Comparison Methods to Analyze (Non)prioritized Composite Endpoints. *Statistics in Medicine*, **38**, 5641-5656. https://doi.org/10.1002/sim.8388

[13] Deltuvaite-Thomas, V. and Burzykowski, T. (2021) Operational Characteristics of Generalized Pairwise Comparisons for Hierarchically Ordered Endpoints. *Pharmaceutical Statistics*, **21**, 122-132. https://doi.org/10.1002/pst.2156

[14] Niebecker, R., Maas, H., Staab, A., Freiwald, M. and Karlsson, M.O. (2019) Modeling Exposure-Driven Adverse Event Time Courses in Oncology Exemplified by Afatinib. *CPT*: *Pharmacometrics & Systems Pharmacology*, **8**, 230-239. https://doi.org/10.1002/psp4.12384

[15] Evans, S.R., Rubin, D., Follmann, D., Pennello, G., Huskins, W.C., Powers, J.H., *et al.* (2015) Desirability of Outcome Ranking (DOOR) and Response Adjusted for Duration of Antibiotic Risk (RADAR). *Clinical Infectious Diseases*, **61**, 800-806. https://doi.org/10.1093/cid/civ495

[16] Fay, M.P. and Malinovsky, Y. (2018) Confidence Intervals of the Mann-Whitney Parameter That Are Compatible with the Wilcoxon-Mann-Whitney Test. *Statistics in Medicine*, **37**, 3991-4006. https://doi.org/10.1002/sim.7890

[17] Gasparyan, S.B., Kowalewski, E.K., Folkvaljon, F., Bengtsson, O., Buenconsejo, J., Adler, J., *et al.* (2021) Power and Sample Size Calculation for the Win Odds Test: Applica-

tion to an Ordinal Endpoint in COVID-19 Trials. *Journal of Biopharmaceutical Statistics*, **31**, 765-787. https://doi.org/10.1080/10543406.2021.1968893

[18] CIOMS Working Group (2023) Benefit-Risk Balance for Medical Products.

[19] Ozenne, B. and Peron, J. (2024) BuyseTest: Implementation of the Generalized Pairwise Comparisons. R Package Version 3.0.2.

[20] Noether, G.E. (1987) Sample Size Determination for Some Common Nonparametric Tests. *Journal of the American Statistical Association*, **82**, 645-647. https://doi.org/10.1080/01621459.1987.10478478

[21] Kornacki, A., Bochniak, A. and Kubik-Komar, A. (2017) Sample Size Determination in the Mann-Whitney Test. *Biometrical Letters*, **54**, 175-186. https://doi.org/10.1515/bile-2017-0010

[22] Zhou, T.J., LaValley, M.P., Nelson, K.P., Cabral, H.J. and Massaro, J.M. (2022) Calculating Power for the Finkelstein and Schoenfeld Test Statistic for a Composite Endpoint with Two Components. *Statistics in Medicine*, **41**, 3321-3335. https://doi.org/10.1002/sim.9419

[23] Gordon, K.B., Strober, B., Lebwohl, M., Augustin, M., Blauvelt, A., Poulin, Y., *et al.* (2018) Efficacy and Safety of Risankizumab in Moderate-to-Severe Plaque Psoriasis (Ultimma-1 and Ultimma-2): Results from Two Double-Blind, Randomised, Placebo-Controlled and Ustekinumab-Controlled Phase 3 Trials. *The Lancet*, **392**, 650-661. https://doi.org/10.1016/s0140-6736(18)31713-6

[24] Verbeeck, J., Dirani, M., Bauer, J.W., Hilgers, R., Molenberghs, G. and Nabbout, R. (2023) Composite Endpoints, Including Patient Reported Outcomes, in Rare Diseases. *Orphanet Journal of Rare Diseases*, **18**, Article No. 262. https://doi.org/10.1186/s13023-023-02819-x

[25] Deltuvaite-Thomas, V., De Backer, M., Parker, S., Deneux, M., Polgreen, L.E., O'Neill, C., *et al.* (2023) Generalized Pairwise Comparisons of Prioritized Outcomes Are a Powerful and Patient-Centric Analysis of Multi-Domain Scores. *Orphanet Journal of Rare Diseases*, **18**, Article No. 321. https://doi.org/10.1186/s13023-023-02943-8

[26] Yuan, S.S., Seifu, Y., Wang, W. and Colopy, M. (2021) Estimands in Safety and Benefit-Risk Evaluation. In: *Quantitative Drug Safety and Benefit-Risk Evaluation: Practical and Cross-Disciplinary Approaches*, Chapman and Hall/CRC, 317-349. https://doi.org/10.1201/9780429488801-18

Appendix: Providing Moment Calculations for Benefit Risk Win Statistics

Suppose that Y_e represents an efficacy event for an experimental individual; Y_e is binary, with probability of success $\mathrm{P}\left[Y_e = 1\right] = \pi_e$. Suppose that Y_c and π_c are analogous quantities for control individuals. Suppose that there are m control observations and n experimental observations. Below, we describe various experimental designs under which, if the initial efficacy consideration is a tie (that is, if $Y_e = Y_c$), then additional information is collected.

Additional non-negative variables Z_e and Z_c, with continuous distributions except for perhaps a positive probability at zero. Heuristically treat these as representing safety measurements. Let

$$F_{e,i} = \mathrm{P}\left[Z_e \le z \mid Y_e = i\right] \text{ for } i \in \{0,1\},$$

$$F_{c,i} = \mathrm{P}\left[Z_c \le z \mid Y_c = i\right] \text{ for } i \in \{0,1\}.$$

Express $F_{e,i}(z) = \kappa_{e,i} + \left(1-\kappa_{e,i}\right)\int_0^z f_{e,i}(x)\mathrm{d}x$ for $z \ge 0$, and $F_{c,i}(z) = \kappa_{c,i} + \left(1-\kappa_{c,i}\right)\int_0^z f_{c,i}(x)\mathrm{d}x$ for $z \ge 0$. Here, we allow Y_e and Z_e to be dependent, and similarly allow Y_c and Z_c to be dependent, but require (Y_e, Z_e) to be independent of (Y_c, Z_c). Let

$$\begin{aligned} G_{ij}^0 &= \mathrm{P}\left[U_{e1} > U_{c1} \mid Y_{c1} = i, Y_{e1} = j\right] \\ &= \int_0^\infty \int_v^\infty f_{e,j}(u) f_{c,i}(v)\mathrm{d}u\,\mathrm{d}v \\ &= \int_0^\infty \int_0^\infty f_{e,j}(x+v) f_{c,i}(v)\mathrm{d}u\,\mathrm{d}v \end{aligned}$$

$$\begin{aligned} G_{i,jk}^1 &= \mathrm{P}\left[U_{e1} > U_{c1}, U_{e2} > U_{c1} \mid Y_{c1} = i, Y_{e1} = j, Y_{e2} = k\right] \\ &= \int_0^\infty \int_w^\infty \int_w^\infty f_{c,i}(w) f_{e,j}(v) f_{e,k}(u)\mathrm{d}u\,\mathrm{d}v\,\mathrm{d}w \\ &= \int_0^\infty \int_0^\infty \int_0^\infty f_{c,i}(w) f_{e,j}(x+w) f_{e,k}(y+w)\mathrm{d}u\,\mathrm{d}x\,\mathrm{d}y \end{aligned}$$

$$\begin{aligned} G_{i,jk}^2 &= \mathrm{P}\left[U_{e1} < U_{c1}, U_{e2} < Y_{c1} \mid Y_{c1} = i, Y_{e1} = j, Y_{e2} = k\right] \\ &= \int_0^\infty \int_{-\infty}^w \int_{-\infty}^w f_{c,i}(w) f_{e,j}(v) f_{e,k}(u)\mathrm{d}u\,\mathrm{d}v\,\mathrm{d}w \\ &= \int_0^\infty \int_{-\infty}^0 \int_{-\infty}^0 f_{c,i}(w) f_{e,j}(x+w) f_{e,k}(y+w)\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}w \end{aligned}$$

$$\begin{aligned} G_{j,i,k}^3 &= \mathrm{P}\left[U_{e1} < U_{c1}, U_{e2} > U_{c1} \mid Y_{c1} = i, Y_{e1} = j, Y_{e2} = k\right] \\ &= \int_0^\infty \int_{-\infty}^w \int_w^\infty f_{c,i}(w) f_{e,j}(v) f_{e,k}(u)\mathrm{d}u\,\mathrm{d}v\,\mathrm{d}w \\ &= \int_0^\infty \int_{-\infty}^0 \int_0^\infty f_{c,i}(w) f_{e,j}(x+w) f_{e,k}(y+w)\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}w \end{aligned}$$

$$\begin{aligned} G_{ij,k}^4 &= \mathrm{P}\left[U_{e1} > U_{c1}, U_{e1} > U_{c2} \mid Y_{c1} = i, Y_{c2} = j, Y_{e1} = k\right] \\ &= \int_0^\infty \int_{-\infty}^w \int_{-\infty}^w f_{e,k}(w) f_{c,j}(v) f_{c,i}(u)\mathrm{d}u\,\mathrm{d}v\,\mathrm{d}w \\ &= \int_{-\infty}^\infty \int_{-\infty}^0 \int_{-\infty}^0 f_{e,k}(w) f_{c,j}(x+w) f_{c,i}(y+w)\mathrm{d}x\,\mathrm{d}y\,\mathrm{d}w \end{aligned}$$

$$\begin{aligned} G_{ij,k}^5 &= \mathrm{P}\left[U_{e1} < U_{c1}, U_{e1} < U_{c2} \mid Y_{c1} = i, Y_{c2} = j, Y_{e1} = k\right] \\ &= \int_{-\infty}^\infty \int_w^\infty \int_w^\infty f_{e,k}(w) f_{c,j}(v) f_{c,i}(u)\mathrm{d}u\,\mathrm{d}v\,\mathrm{d}w \end{aligned}$$

$$G_{i,j,k}^{6} = \mathrm{P}\left[U_{c1} < U_{e1} < U_{c2} \mid Y_{c1} = i, Y_{c2} = j, Y_{e1} = k\right]$$
$$= \int_{-\infty}^{\infty}\int_{-\infty}^{w}\int_{w}^{\infty} f_{e,k}(w) f_{c,j}(u) f_{c,i}(v)\,\mathrm{d}u\,\mathrm{d}v\,\mathrm{d}w$$

Under the null hypothesis, and assuming equal distributions regardless of values of Y_e and Y_c, then $G_{ij}^{0} = 1/2$, $G_{i,jk}^{1} = 1/3$, $G_{i,jk}^{2} = 1/3$, $G_{i,jk}^{3} = 1/6$, $G_{i,jk}^{4} = 1/3$, $G_{i,jk}^{5} = 1/3$, and $G_{i,jk}^{6} = 1/6$.

Let

$$V_{k,\ell} = Y_{e,\ell} - Y_{c,k} = \begin{cases} 1 & \text{if } \ell \text{ beats } k \\ -1 & \text{if } k \text{ beats } \ell \\ 0 & \text{if } k \text{ and } \ell \text{ tie} \end{cases}, \quad W_{k,\ell} = \begin{cases} 1 & \text{if } Z_{e,\ell} > Z_{c,k} \\ -1 & \text{if } Z_{e,\ell} = Z_{c,k} \\ 0 & \text{if } Z_{e,\ell} < Z_{c,k} \end{cases}.$$

Let

$$\check{H}_{ij} = \mathrm{P}\left[W_{11} = 1 \mid Y_{c,1} = i, Y_{e,1} = j\right] = \mathrm{P}\left[Z_{e,1} > Z_{c,1} \mid Y_{c,1} = i, Y_{e,1} = j\right]$$
$$= \left(1-\kappa_{e,j}\right)\left[\kappa_{c,i} + \left(1-\kappa_{c,i}\right)G_{ij}\right]$$

$$\hat{H}_{ij} = \mathrm{P}\left[W_{11} = -1 \mid Y_{c,1} = i, Y_{e,1} = j\right] = \mathrm{P}\left[Z_{e,1} < Z_{c,1} \mid Y_{c,1} = i, Y_{e,1} = j\right]$$
$$= \left(1-\kappa_{c,i}\right)\left[\kappa_{e,j} + \left(1-\kappa_{e,j}\right)\left(1-G_{ij}\right)\right]$$

Let

$$\begin{aligned}
\tilde{H}_{i,jk}^{1} &= \mathrm{P}\left[W_{1,1} = 1, W_{1,2} = 1 \mid Y_{c,1} = i, Y_{e,1} = j, Y_{e,2} = k\right] \\
&+\mathrm{P}\left[W_{1,1} = -1, W_{1,2} = -1 \mid Y_{c,1} = i, Y_{e,1} = j, Y_{e,2} = k\right] \\
&-\mathrm{P}\left[W_{1,1} = 1, W_{1,2} = -1 \mid Y_{c,1} = i, Y_{e,1} = j, Y_{e,2} = k\right] \\
&-\mathrm{P}\left[W_{1,1} = -1, W_{1,2} = 1 \mid Y_{c,1} = i, Y_{e,1} = j, Y_{e,2} = k\right] \\
&= \mathrm{P}\left[Z_{e1} > Z_{c1}, Z_{e2} > Z_{c1} \mid Y_{c,1} = i, Y_{e,1} = j, Y_{e,2} = k\right] \\
&+\mathrm{P}\left[Z_{e1} < Z_{c1}, Z_{e2} < Z_{c1} \mid Y_{c,1} = i, Y_{e,1} = j, Y_{e,2} = k\right] \\
&-\mathrm{P}\left[Z_{e1} > Z_{c1}, Z_{e2} < Z_{c1} \mid Y_{c,1} = i, Y_{e,1} = j, Y_{e,2} = k\right] \\
&-\mathrm{P}\left[Z_{e1} < Z_{c1}, Z_{e2} > Z_{c1} \mid Y_{c,1} = i, Y_{e,1} = j, Y_{e,2} = k\right] \\
&= \left(1-\kappa_{ej}\right)\left(1-\kappa_{ek}\right)\left[\kappa_{ci} + \left(1-\kappa_{ci}\right)G_{i,jk}^{1}\right] \\
&+\left(1-\kappa_{ci}\right)\left[\kappa_{ej}\kappa_{ek} + \kappa_{ej}\left(1-\kappa_{ek}\right)G_{ik} + \kappa_{ek}\left(1-\kappa_{ej}\right)G_{ij} + \left(1-\kappa_{ek}\right)\left(1-\kappa_{ej}\right)G_{i,jk}^{2}\right] \\
&-\left(1-\kappa_{ci}\right)\left(1-\kappa_{ej}\right)\left[\kappa_{ek}G_{ij} + \left(1-\kappa_{ek}\right)G_{k,i,j}^{3}\right] \\
&-\left(1-\kappa_{ci}\right)\left(1-\kappa_{ek}\right)\left[\kappa_{ej}G_{ik} + \left(1-\kappa_{ej}\right)G_{j,i,k}^{3}\right]
\end{aligned}$$

$$\begin{aligned}
\tilde{H}_{i,jk}^{2} &= \mathrm{P}\left[W_{1,1} = 1, W_{2,1} = 1 \mid Y_{c,1} = i, Y_{e,1} = j, Y_{c,2} = k\right] \\
&+\mathrm{P}\left[W_{1,1} = -1, W_{2,1} = -1 \mid Y_{c,1} = i, Y_{e,1} = j, Y_{c,2} = k\right] \\
&-\mathrm{P}\left[W_{1,1} = 1, W_{2,1} = -1 \mid Y_{c,1} = i, Y_{e,1} = j, Y_{c,2} = k\right] \\
&-\mathrm{P}\left[W_{1,1} = -1, W_{2,1} = 1 \mid Y_{c,1} = i, Y_{e,1} = j, Y_{c,2} = k\right] \\
&= \mathrm{P}\left[Z_{e1} > Z_{c1}, Z_{e1} > Z_{c2} \mid Y_{c,1} = i, Y_{e,1} = j, Y_{c,2} = k\right] \\
&+\mathrm{P}\left[Z_{e1} < Z_{c1}, Z_{e1} < Z_{c2} \mid Y_{c,1} = i, Y_{e,1} = j, Y_{c,2} = k\right] \\
&-\mathrm{P}\left[Z_{e1} > Z_{c1}, Z_{e1} < Z_{c2} \mid Y_{c,1} = i, Y_{e,1} = j, Y_{c,2} = k\right]
\end{aligned}$$

$$
\begin{aligned}
&-\mathrm{P}\left[Z_{e1}<Z_{c1}, Z_{e1}>Z_{c2} \mid Y_{c,1}=i, Y_{e,1}=j, Y_{c,2}=k\right] \\
&=\left(1-\kappa_{ej}\right)\left[\kappa_{ci}\kappa_{ck}+\kappa_{ci}\left(1-\kappa_{ck}\right)G_{kj}+\kappa_{ck}\left(1-\kappa_{ci}\right)G_{ij}+\left(1-\kappa_{ci}\right)\left(1-\kappa_{ck}\right)G^{4}_{ik,j}\right] \\
&+\left(1-\kappa_{ci}\right)\left(1-\kappa_{ck}\right)\left[\kappa_{ej}+\left(1-\kappa_{ej}\right)G^{5}_{ik,j}\right] \\
&-\left(1-\kappa_{ck}\right)\left(1-\kappa_{ej}\right)\left[\kappa_{ci}G_{kj}+\left(1-\kappa_{ci}\right)G^{6}_{i,j,k}\right] \\
&-\left(1-\kappa_{ci}\right)\left(1-\kappa_{ej}\right)\left[\kappa_{ck}G_{ij}+\left(1-\kappa_{ck}\right)G^{6}_{k,j,i}\right]
\end{aligned}
$$

Take $k\in\{1,\cdots,m\}$, and $\ell\in\{1,\cdots,n\}$. First moments are given by

$$
\mathrm{E}\left[V_{1,1}\right]=\pi_e\left(1-\pi_c\right)-\pi_c\left(1-\pi_e\right)
$$

$$
\begin{aligned}
\mathrm{E}\left[W_{k,\ell}\right]&=\left(1-\pi_c\right)\left(1-\pi_e\right)\left(\check{H}_{00}-\hat{H}_{00}\right)+\pi_c\left(1-\pi_e\right)\left(\check{H}_{10}-\hat{H}_{10}\right) \\
&+\left(1-\pi_c\right)\pi_e\left(\check{H}_{01}-\hat{H}_{01}\right)+\pi_c\pi_e\left(\check{H}_{11}-\hat{H}_{11}\right).
\end{aligned}
$$

Here are second moments for same row and column:

$$
\mathrm{E}\left[V^2_{k,\ell}\right]=1-\mathrm{P}\left[Y_{e,\ell}=Y_{c,k}\right]=1-\pi_e\pi_c-\left(1-\pi_c\right)\left(1-\pi_e\right)
$$

$$
\begin{aligned}
\mathrm{E}\left[W^2_{k,\ell}\right]&=1-\mathrm{P}\left[Z_{e,\ell}=Z_{c,}\right] \\
&=1-\kappa_{e,0}\kappa_{c,0}\left(1-\pi_e\right)\left(1-\pi_c\right)-\kappa_{e,1}\kappa_{c,0}\pi_e\left(1-\pi_c\right) \\
&-\kappa_{e,0}\kappa_{c,1}\left(1-\pi_e\right)\pi_c-\kappa_{e,1}\kappa_{c,1}\pi_e\pi_c
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{E}\left[V_{k,\ell}W_{k,\ell}\right]&=\pi_e\left(1-\pi_c\right)\check{H}_{01}+\pi_c\left(1-\pi_e\right)\hat{H}_{10}-\pi_c\left(1-\pi_e\right)\check{H}_{10}-\pi_e\left(1-\pi_c\right)\hat{H}_{01} \\
&=\pi_e\left(1-\pi_c\right)\left(\check{H}_{01}-\hat{H}_{01}\right)+\pi_c\left(1-\pi_e\right)\left(\hat{H}_{10}-\check{H}_{10}\right)
\end{aligned}
$$

Take $\ell\neq m$. Here are entries for the same row but different columns:

$$
\begin{aligned}
\mathrm{E}\left[V_{k,\ell}V_{k,m}\right]&=\mathrm{P}\left[Y_{e,\ell}=0,Y_{c,k}=1,Y_{e,m}=0\right]+\mathrm{P}\left[Y_{e,\ell}=1,Y_{c,k}=0,Y_{e,m}=1\right] \\
&=\pi_c\left(1-\pi_e\right)^2+\left(1-\pi_c\right)\pi_e^2
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{E}\left[W_{k,\ell}W_{k,m}\right]&=\mathrm{P}\left[W_{k,\ell}=1,W_{k,m}=1\right]+\mathrm{P}\left[W_{k,\ell}=-1,W_{k,m}=-1\right] \\
&-\mathrm{P}\left[W_{k,\ell}=1,W_{k,m}=-1\right]-\mathrm{P}\left[W_{k,\ell}=-1,W_{k,m}=1\right] \\
&=\tilde{H}^1_{0,00}\left(1-\pi_e\right)^2\left(1-\pi_c\right)+\tilde{H}^1_{1,00}\pi_c\left(1-\pi_e\right)^2 \\
&+\tilde{H}^1_{0,10}\pi_e\left(1-\pi_e\right)\left(1-\pi_c\right)+\tilde{H}^1_{0,01}\pi_e\left(1-\pi_e\right)\left(1-\pi_c\right) \\
&+\tilde{H}^1_{1,10}\pi_e\left(1-\pi_e\right)\pi_c+\tilde{H}^1_{1,01}\pi_c\pi_e\left(1-\pi_e\right) \\
&+\tilde{H}^1_{0,11}\pi_e^2\left(1-\pi_c\right)+\tilde{H}^1_{1,11}\pi_e^2\pi_c
\end{aligned}
$$

and

$$
\begin{aligned}
\mathrm{E}\left[V_{k,\ell}W_{k,m}\right]&=\mathrm{P}\left[Y_{e,\ell}=0,Y_{c,k}=1,W_{k,m}=-1\right]-\mathrm{P}\left[Y_{e,\ell}=0,Y_{c,k}=1,W_{k,m}=1\right] \\
&+\mathrm{P}\left[Y_{e,\ell}=1,Y_{c,k}=0,W_{k,m}=1\right]-\mathrm{P}\left[Y_{e,\ell}=1,Y_{c,k}=0,W_{k,m}=-1\right] \\
&=\mathrm{P}\left[Y_{e,\ell}=0,Y_{c,k}=1,W_{k,m}=-1,Y_{e,m}=0\right] \\
&+\mathrm{P}\left[Y_{e,\ell}=0,Y_{c,k}=1,W_{k,m}=-1,Y_{e,m}=1\right] \\
&-\mathrm{P}\left[Y_{e,\ell}=0,Y_{c,k}=1,W_{k,m}=1,Y_{e,m}=0\right] \\
&-\mathrm{P}\left[Y_{e,\ell}=0,Y_{c,k}=1,W_{k,m}=1,Y_{e,m}=1\right]
\end{aligned}
$$

$$
\begin{aligned}
&+\mathrm{P}\left[Y_{e,\ell}=1,Y_{c,k}=0,W_{k,m}=1,Y_{e,m}=0\right]\\
&+\mathrm{P}\left[Y_{e,\ell}=1,Y_{c,k}=0,W_{k,m}=1,Y_{e,m}=1\right]\\
&-\mathrm{P}\left[Y_{e,\ell}=1,Y_{c,k}=0,W_{k,m}=-1,Y_{e,m}=0\right]\\
&-\mathrm{P}\left[Y_{e,\ell}=1,Y_{c,k}=0,W_{k,m}=-1,Y_{e,m}=1\right]\\
&=\pi_c\left(1-\pi_e\right)^2\hat{H}_{10}+\pi_c\pi_e\left(1-\pi_e\right)\hat{H}_{11}\\
&-\pi_c\left(1-\pi_e\right)^2\check{H}_{10}-\pi_c\pi_e\left(1-\pi_e\right)\check{H}_{11}\\
&+\left(1-\pi_c\right)\pi_e\left(1-\pi_e\right)\check{H}_{00}+\left(1-\pi_c\right)\pi_e^2\check{H}_{01}\\
&-\left(1-\pi_c\right)\pi_e\left(1-\pi_e\right)\hat{H}_{00}-\left(1-\pi_c\right)\pi_e^2\hat{H}_{01}\\
&=\pi_c\left(1-\pi_e\right)^2\left(\hat{H}_{10}-\check{H}_{10}\right)+\pi_c\pi_e\left(1-\pi_e\right)\left(\hat{H}_{11}-\check{H}_{11}\right)\\
&-\left(1-\pi_c\right)\pi_e\left(1-\pi_e\right)\left(\hat{H}_{00}-\check{H}_{00}\right)-\left(1-\pi_c\right)\pi_e^2\left(\hat{H}_{01}-\check{H}_{01}\right)
\end{aligned}
$$

Take $j\neq k$. Here are entries for the same column but different rows:

$$
\begin{aligned}
\mathrm{E}\left[V_{j,m}V_{k,m}\right]&=\mathrm{P}\left[Y_{c,j}=0,Y_{c,k}=0,Y_{e,m}=1\right]+\mathrm{P}\left[Y_{c,j}=1,Y_{c,k}=1,Y_{e,m}=0\right]\\
&=\pi_e\left(1-\pi_c\right)^2+\left(1-\pi_e\right)\pi_c^2
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{E}\left[W_{j,m}W_{k,m}\right]&=\mathrm{P}\left[W_{j,m}=1,W_{k,m}=1\right]+\mathrm{P}\left[W_{j,m}=-1,W_{k,m}=-1\right]\\
&\quad-\mathrm{P}\left[W_{j,m}=1,W_{k,m}=-1\right]-\mathrm{P}\left[W_{j,m}=-1,W_{k,m}=1\right]\\
&=\tilde{H}^2_{00,0}\left(1-\pi_e\right)\left(1-\pi_c\right)^2+\tilde{H}^2_{10,0}\pi_c\left(1-\pi_c\right)\left(1-\pi_e\right)\\
&\quad+\tilde{H}^2_{01,0}\pi_c\left(1-\pi_e\right)\left(1-\pi_c\right)+\tilde{H}^2_{00,1}\pi_e\left(1-\pi_c\right)^2\\
&\quad+\tilde{H}^2_{11,0}\pi_c^2\left(1-\pi_e\right)+\tilde{H}^2_{10,1}\pi_c\pi_e\left(1-\pi_c\right)\\
&\quad+\tilde{H}^2_{01,1}\pi_c\pi_e\left(1-\pi_c\right)+\tilde{H}^2_{11,1}\pi_c^2\pi_e
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{E}\left[V_{j,m}W_{k,m}\right]&=\mathrm{P}\left[Y_{e,m}=0,Y_{c,j}=1,W_{k,m}=-1\right]-\mathrm{P}\left[Y_{e,m}=0,Y_{c,j}=1,W_{k,m}=1\right]\\
&\quad+\mathrm{P}\left[Y_{e,m}=1,Y_{c,j}=0,W_{k,m}=1\right]-\mathrm{P}\left[Y_{e,m}=1,Y_{c,j}=0,W_{k,m}=-1\right]\\
&=\mathrm{P}\left[Y_{e,m}=0,Y_{c,j}=1,W_{k,m}=-1,Y_{c,k}=0\right]\\
&\quad+\mathrm{P}\left[Y_{e,m}=0,Y_{c,j}=1,W_{k,m}=-1,Y_{c,k}=1\right]\\
&\quad-\mathrm{P}\left[Y_{e,m}=0,Y_{c,j}=1,W_{k,m}=1,Y_{c,k}=0\right]\\
&\quad-\mathrm{P}\left[Y_{e,m}=0,Y_{c,j}=1,W_{k,m}=1,Y_{c,k}=1\right]\\
&\quad+\mathrm{P}\left[Y_{e,m}=1,Y_{c,j}=0,W_{k,m}=1,Y_{c,k}=0\right]\\
&\quad+\mathrm{P}\left[Y_{e,m}=1,Y_{c,j}=0,W_{k,m}=1,Y_{c,k}=1\right]\\
&\quad-\mathrm{P}\left[Y_{e,m}=1,Y_{c,j}=0,W_{k,m}=-1,Y_{c,k}=0\right]\\
&\quad-\mathrm{P}\left[Y_{e,m}=1,Y_{c,j}=0,W_{k,m}=-1,Y_{c,k}=1\right]\\
&=\pi_c\left(1-\pi_c\right)\left(1-\pi_e\right)\hat{H}_{00}+\pi_c^2\left(1-\pi_e\right)\hat{H}_{10}\\
&\quad-\pi_c\left(1-\pi_c\right)\left(1-\pi_e\right)\check{H}_{00}-\pi_c^2\left(1-\pi_e\right)\check{H}_{10}\\
&\quad+\left(1-\pi_c\right)^2\pi_e\check{H}_{01}+\left(1-\pi_c\right)\pi_e\pi_c\check{H}_{11}\\
&\quad-\left(1-\pi_c\right)^2\pi_e\hat{H}_{01}-\left(1-\pi_c\right)\pi_c\pi_e\hat{H}_{11}
\end{aligned}
$$

$$= \pi_c(1-\pi_c)(1-\pi_e)\left(\hat{H}_{00} - \check{H}_{00}\right) + \pi_c^2(1-\pi_e)\left(\hat{H}_{10} - \check{H}_{10}\right)$$
$$-(1-\pi_c)^2\pi_e\left(\hat{H}_{01} - \check{H}_{01}\right) - (1-\pi_c)\pi_e\pi_c\left(\hat{H}_{11} - \check{H}_{11}\right)$$

Let $T = \sum_i \sum_j \left(V_{ij} - W_{ij}\right)$.

$$\mathrm{E}[TT] = \sum_i \sum_j \sum_k \sum_\ell \mathrm{E}\left[\left(V_{ij} - W_{ij}\right)\left(V_{k\ell} - W_{k\ell}\right)\right]$$
$$= \sum_i \sum_j \left(\mathrm{E}\left[V_{ij}V_{ij}\right] - 2\mathrm{E}\left[V_{ij}W_{ij}\right] + \mathrm{E}\left[W_{ij}W_{ij}\right]\right)$$
$$+ \sum_i \sum_j \sum_{\ell \neq j} \left(\mathrm{E}\left[V_{ij}V_{i\ell}\right] - \mathrm{E}\left[V_{ij}W_{i\ell}\right] - \mathrm{E}\left[W_{ij}V_{i\ell}\right] + \mathrm{E}\left[W_{ij}W_{i\ell}\right]\right)$$
$$+ \sum_i \sum_j \sum_{k \neq i} \left(\mathrm{E}\left[V_{ij}V_{kj}\right] - \mathrm{E}\left[V_{ij}W_{kj}\right] - \mathrm{E}\left[W_{ij}V_{kj}\right] + \mathrm{E}\left[W_{ij}W_{kj}\right]\right)$$
$$+ \sum_i \sum_j \sum_{k \neq i} \sum_{\ell \neq j} \left(\mathrm{E}\left[V_{ij}V_{k\ell}\right] - \mathrm{E}\left[V_{ij}W_{k\ell}\right] - \mathrm{E}\left[W_{ij}V_{k\ell}\right] + \mathrm{E}\left[W_{ij}W_{k\ell}\right]\right)$$
$$= mn\left(\mathrm{E}\left[V_{11}V_{11}\right] - 2\mathrm{E}\left[V_{11}W_{11}\right] + \mathrm{E}\left[W_{11}W_{11}\right]\right)$$
$$+ mn(n-1)\left(\mathrm{E}\left[V_{11}V_{12}\right] - 2\mathrm{E}\left[V_{11}W_{12}\right] + \mathrm{E}\left[W_{11}W_{12}\right]\right)$$
$$+ m(m-1)n\left(\mathrm{E}\left[V_{11}V_{21}\right] - 2\mathrm{E}\left[V_{11}W_{21}\right] + \mathrm{E}\left[W_{11}W_{21}\right]\right)$$
$$+ m(m-1)n(n-1)\left(\mathrm{E}\left[V_{11}\right] - \mathrm{E}\left[W_{11}\right]\right)^2$$

Appendix: Software Used for These Calculations

The following R software packages were used in these calculations:

• BuyseTest [19]: R package used for evaluating prioritized and non-prioritized net benefit,

• asht [16]: R package used for evaluating the WMW test and associated confidence interval,

• AESim: R package that provides sample size derivation for a binary efficacy endpoint and AEI composite score to be analyzed via non-prioritized net benefit.

No other data were used or produced in this research.