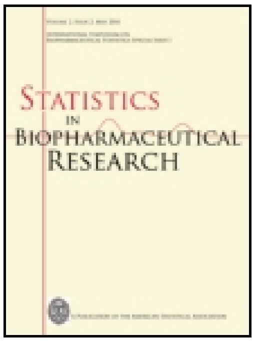

Statistics in Biopharmaceutical Research

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/usbr20

An Expected Score Approach to Ordinal Outcomes in a Bayesian, Response Adaptive, Randomized Trial

Jonathan Beall, Jordan Elm, James Chamberlain, Eric Rosenthal, Jaideep Kapur & Valerie Durkalski-Mauldin

To cite this article: Jonathan Beall, Jordan Elm, James Chamberlain, Eric Rosenthal, Jaideep Kapur & Valerie Durkalski-Mauldin (2023): An Expected Score Approach to Ordinal Outcomes in a Bayesian, Response Adaptive, Randomized Trial, Statistics in Biopharmaceutical Research, DOI: 10.1080/19466315.2023.2169344

To link to this article: https://doi.org/10.1080/19466315.2023.2169344

Published online: 24 Feb 2023.

STATISTICS IN BIOPHARMACEUTICAL RESEARCH
2023, VOL. 00, NO. 0, 1–6
https://doi.org/10.1080/19466315.2023.2169344

Taylor & Francis
Taylor & Francis Group

An Expected Score Approach to Ordinal Outcomes in a Bayesian, Response Adaptive, Randomized Trial

Jonathan Beall[a], Jordan Elm[a], James Chamberlain[b], Eric Rosenthal[c], Jaideep Kapur[d], and Valerie Durkalski-Mauldin[a]

[a]Department of Public Health Sciences, Medical University of South Carolina, Charleston, SC; [b]Division of Emergency Medicine, Children's National Hospital, Washington, DC; [c]Department of Neurology, Massachusetts General Hospital, Boston, MA; [d]Department of Neurology, UVA Health, Charlottesville, VA

ABSTRACT

Ordinal outcomes are common in medicine and can be analyzed in many ways, but the distribution of ordinal data can present unique challenges. The proposed KESETT study is a three-armed, randomized trial comparing two doses of ketamine plus levetiracetam to levetiracetam alone for treating patients with benzodiazepine-refractory status epilepticus. A Bayesian, adaptive clinical trial is proposed employing an ordinal primary outcome at 60 min ranging from 1 (improving consciousness and seizure cessation) to 5 (life-threatening event/death). Based on a previous study, the ordinal outcome is expected to have a bimodal distribution, with the effect of treatment expected to be nonproportional across the outcome scale. As such, approaches relying on assuming proportionality of the odds are not appropriate. We propose for this scenario an analytic approach to compare ordinal outcomes using the expected score derived from the posterior distribution for each treatment group. This approach requires minimal assumptions, maintains the benefit of using the full ordinal scale, is interpretable, and can be used in a Bayesian analysis framework. We compare this new approach under multiple simulated scenarios to three traditional frequentist approaches. The new approach controls Type I error and power, resulting in a sizable reduction in sample size relative to a nonparametric test.

ARTICLE HISTORY
Received January 2022
Accepted December 2022

KEYWORDS
Bayesian; Clinical trial; Muli-arm; Ordinal outcome; Response adaptive randomization; Word

1. Introduction

Previous clinical trials have taken a wide variety of approaches for a primary analysis to assess for an association between an ordinal outcome and a treatment of interest. Ordinal data is frequently dichotomized for simplicity or if the expected treatment benefit is concentrated in one section only. Alternatively, ordinal data may be analyzed using a proportional odds model, a nonparametric test (i.e., Wilcoxon Rank Sum, Cochran-Mantel-Haenszel, etc.) (Saver 2011; Evans et al. 2015; Evans and Follmann 2016), or with parametric tests (i.e., t-test) which assume that the mean of the ordinal outcome follows a normal distribution (Chaisinanunkul et al. 2015; Jovin et al. 2015; Deeds et al. 2020). Both the proportional odds model and t-test have limitations because of their assumptions, which could hinder model performance if these assumptions are violated. For the t-test, while the central limit theorem implies that the mean of the ordinal outcome can be treated as normally distributed, doing so will inherently allocate nonzero probability mass to average values which cannot be obtained from the ordinal data, such as values which are below the minimum observable response. In this article, we propose an analytic approach for an ordinal outcome based on the expected score which avoids assuming the mean of the ordinal outcome follows a normal distribution, maintains the benefit of using the full ordinal range and can be used in a Bayesian analysis framework.

2. Motivating Example

This work is motivated by a multicenter, randomized clinical trial which is currently being planned to compare the effectiveness of two doses of ketamine (KET) plus levetiracetam (LEV) versus levetiracetam alone in emergency department treatment of patients with benzodiazepine-refractory status epilepticus (Ketamine for Established Status Epilepticus Treatment Trial, or KESETT). The primary objective is to determine whether the addition of low- or high-dose ketamine (KET) to levetiracetam (LEV) is more effective than LEV alone for Status Epilepticus (SE) among patients one year and older. In this study, LEV is the active comparator. The primary outcome for the study is a composite outcome which is a graded scale that ranges from 1 to 5. The outcome is assessed during the 60 min following study drug initiation and is scored using the scheme below.

1. No clinically evident or electrographic status epilepticus after 15 min, no rescue drugs, and improving mental status by 60 min
2. No clinically evident or electrographic status epilepticus after 15 min, not intubated, but not improving mental status at 60 min
3. No clinically evident or electrographic status epilepticus in the 15 min after dosing, but intubation or requiring rescue medications (including medications used for intubation)

CONTACT Jonathan Beall bealljo@musc.edu Department of Public Health Sciences, Medical University of South Carolina, Charleston, SC.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/SBR.

4. Any clinically evident seizure, continuous electrographic seizure exceeding 5 min in duration, or cumulative electrographic seizure burden exceeding 8 min (i.e., status epilepticus), occurring between 15 min and 60 min after treatment.
5. Life-threatening hypotension or cardiac arrhythmia or death

A prior study of LEV suggested that the distribution of the ordinal data would be bimodal and highly concentrated in two nonadjacent categories, where it is expected that there will be a substantial reduction in the more severe categories with a reduced impact of treatment in the lower categories (Kapur et al. 2019; Chamberlain et al. 2020). Given the hypothesized treatment effect, the proportional odds assumption would be violated. Further, the KESETT clinical investigators sought a Bayesian design. As such, we propose the use of the expected score as the primary measure of comparing treatment arms in a Bayesian, adaptive clinical trial. We compare our proposed design to other analytic approaches under a variety of simulated scenarios to assess the impact of early stopping criteria at interim analyses and Response Adaptive Randomization (RAR).

3. Methods

3.1. Expected Score

Let $\mathbf{X}_t = (x_{1t}, \ldots, x_{n_t t})$ represent a set of ordinal outcomes for subjects $i = 1, \ldots, n_t$ in treatment group t where $x_{it} \in (1, \ldots, L)$. We assume that $\mathbf{X}_t \sim$ Multinomial $(\mathbf{P}_t)$ with $\mathbf{P}_t = (P_{1t}, \ldots, P_{Lt})$ representing the vector of probabilities where $P_{\mathrm{l}t} = Pr(x_{it} = \mathrm{l})$ and $\sum_{\mathrm{l}} P_{\mathrm{l}t} = 1$. Our approach is Bayesian; as such, we assume a Dirichlet prior distribution, Dirichlet (α_0) with $\alpha_0 = (0.1, \ldots, 0.1)$, for $\mathbf{P}_t$. This choice of prior distribution is noninformative and was chosen because it is conjugate for the multinomial likelihood. Given the utilization of the conjugate Dirichlet prior for the multinomial likelihood, our resultant posterior distribution is a Dirichlet distribution where $P_t \sim$ Dirichlet$\left(\sum_i I(x_{it} = 1) + 0.1, \ldots, \sum_i I(x_{it} = L) + 0.1\right)$ where $I(x_{it} = \mathrm{l})$ is an indicator function which equals 1 if subject i in treatment group t has a score of l. For treatment group t, we define the expected score, S_t as $S_t = \sum_{\mathrm{l}} \mathrm{l} \times P_{\mathrm{l}t}$. This quantity of interest will be our primary tool for comparing treatment arms with a lower score indicating a better outcome and is approximately equal to the mean of the observed scores. Our utilization of the expected score is similar to that of the UW-mRS in the MOST and DAWN trials, where the weights applied to our outcome reflect a constant increase in severity when transitioning to the next level in the ordinal outcome (Jovin et al. 2015; Deeds et al. 2020). For these trials, the mean of the UW-mRS score was assumed be normally distributed which is appropriate if the CLT applies. Our approach differs as we construct the expected score using the estimated probabilities of observing each level of the outcome, avoiding a distributional assumption on the expected score. That is, our approach does not rely on an assumption of normality, as the expected score is structurally bound to exist in the observable scoring range of the outcome of interest. Treatment success is defined as a significant reduction in the expected score in either of the treatment arms relative to control. As such, without loss of generality, let S_1 represent the expected score for the control arm of the study. Then, we define the posterior probability of treatment success for treatment arm t as $\Pr(S_t < S_1)$.

3.2. Comparators

We will compare our approach to three other analytic choices: t-test, Wilcoxon rank sum, and proportional odds regression. For the t-test and wilcoxon rank sum tests, each active treatment arm will be independently compared against the control arm using a one-sided test. For the proportional odds regression, letting T_{iL} represent an indicator variable which equals 1 if subject i is in the low dose KET arm and 0 otherwise and letting T_{iH} represent an indicator variable which equals 1 if subject i is in the high dose KET arm and 0 otherwise, we construct the model shown below:

$$\operatorname{logit}\left(P\left(x_{it} \leq \mathrm{l}\right)\right) = \beta_0 + \beta_1 \times T_{iL} + \beta_2 \times T_{iH}.$$

Similar to that of the t-test and Wilxocon rank sum tests, we will test β_1 and β_2 independently using one-sided tests.

3.3. Response-Adaptive Randomization

In simulated designs which use RAR, it will be used to allocate subjects to the better performing KET arm in order to maximize the likelihood of patient benefit. For these simulations, the randomization scheme will be equal allocation (1:1:1) for the first 100 patients. Once the primary outcome has been obtained for the first 100 subjects, the RAR will update the allocation probabilities with subsequent updates occurring every 100 patients. Let A_{mt} represent the allocation to treatment group t at interim analysis m where m = 1,…,M. This allocation rate is calculated using $A_{mt} = \frac{R_{mt}}{\sum_t R_{mt}}$, where $R_{mt} = \sqrt{Pr(S_t < S_1)}$. Allocation to the active control arm will be fixed at 33 subjects per enrollment block in order to maintain adequate power. Allocation to treatment arm t will be set to 0 if $A_{mt} < 0.1$, but allocation can begin again at the next interim if $A_{mt} \geq 0.1$.

3.4. Interim Analyses for Early Stopping

In addition to assessing the impact of RAR, we will evaluate designs which allow for early trial termination due to overwhelming efficacy or futility. For designs which allow for early termination, three interim analyses will occur. The first interim will occur once the primary outcome has been obtained for the first 300 subjects and will occur additionally once 400 and 500 primary outcomes have been received. For all frequentist approaches (t-test, wilcoxon rank sum, and proportional odds model), early efficacy boundaries were determined using a Hwang-Shih-DeCani spending function with parameter set to −4 to create O'Brien-Fleming-like boundaries (Hwang, Shih, and De Cani 1990). The boundaries were created assuming a one-sided Type I error rate of 1.25% per active treatment arm, to provide an overall one-sided Type I error rate of 2.5%, and 10% Type II error. Following the approach by Shi and Yin, who demonstrated the translatability between frequentist p-values and Bayesian posterior probabilities, we derived equivalent early stopping criteria for our Bayesian design (Shi and Yin 2021).

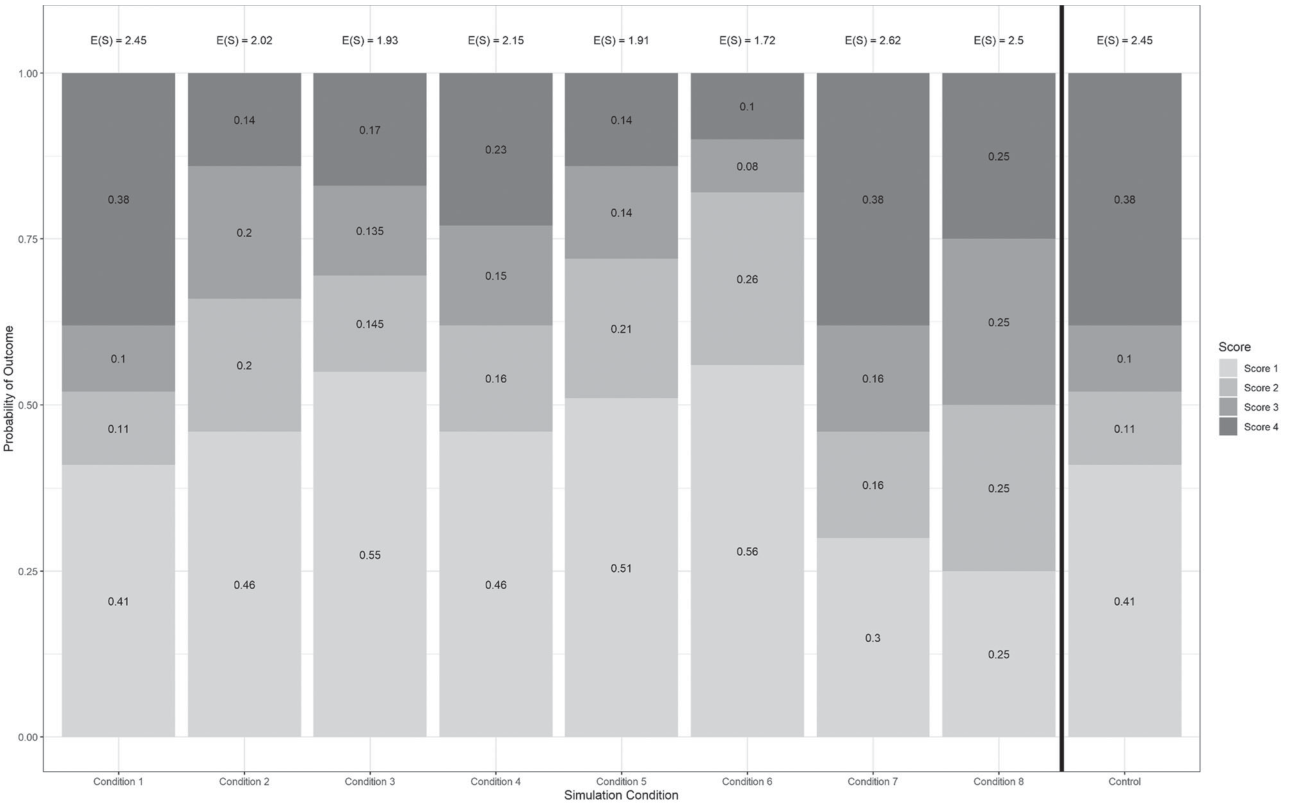

Figure 1. Probability of outcome and expected score for all simulation conditions.

3.5. Success Criteria at Analysis

Let z_{em} be the critical value and p_{em} the corresponding p-value required at the mth analysis according to the Hwang-Shih-DeCani spending function to declare overwhelming efficacy for a single treatment arm. For a trial design which uses a frequentist analytic method (t-test, Wilcoxon rank sum, or proportional odds), the trial will be terminated early for overwhelming efficacy if the resultant p-value for either treatment arm is less than p_{em} for either active treatment arm. For the Bayesian expected score approach, we will terminate the trial for overwhelming efficacy if $\Pr(S_2 < S_1) > (1-p_{em})$ or $\Pr(S_3 < S_1) > (1-p_{em})$.

3.6. Stopping for Overwhelming Futility

Let z_{fm} be the critical value and p_{fm} the corresponding p-value required at the mth analysis according to the Hwang-Shih-DeCani spending function to declare overwhelming futility for a single treatment arm. For a trial design which uses a frequentist analytic method (t-test, Wilcoxon rank sum, or proportional odds), the trial will be terminated early for overwhelming futility if the resultant p-value is greater than p_{fm} for both active treatment arms. For the Bayesian expected score approach, we will terminate the trial for overwhelming futility if $\Pr(S_2 < S_1) < (1 - p_{fm})$ and $\Pr(S_3 < S_1) < (1 - p_{fm})$.

4. Naive Frequentist Sample Size

Based on a previous trial by this team (ESETT) of 478 patients, we found the proportion of LEV patients in each ordinal category (Kapur et al. 2019; Chamberlain et al. 2020). Initial sample size estimates were 220 per group using the following assumptions:

1. 90% power,
2. 2.5% Type I error (one-sided),
3. two-sample Wilcoxon Rank Sum test where the proportions in the LEV group are ($1 = 41\%$, $2 = 11\%$, $3 = 10\%$, $4 = 38\%$, $5 = 0\%$ as observed in ESETT) and the proportions in either of the KET dose groups is ($1 = 46\%$, $2 = 20\%$, $3 = 20\%$, $4 = 14\%$, $5 = 0\%$) is 220 subjects per group. Based on clinical discussion this was a meaningful improvement in this population and represents the clinically important difference.

5. Simulation Study

To evaluate the proposed design and compare it to the initial analytic approach, we evaluated our model under a wide range of conditions in simulation. Specifically, we tested our model under no treatment effect and seven conditions where there was a treatment effect. These simulation conditions are shown in Figure 1. Initially, simulations were conducted with a maximum total $N = 660$ (our naïve sample size), but were redone with a lower maximum because the naïve maximum exceeded 99% power in most conditions, including the clinically important difference stated above. The simulations presented here assumed a maximum sample size of 600. In Figure 1, Condition 1 represents the null case where neither of the treatment arms differ from the control arm. Conditions 2 through 6 all provide an improvement relative to control, with Condition 2 representing the clinically important difference. Conditions 7 and 8 reflect situations where the treatment is detrimental to patient outcomes. All conditions assumed score 5 was 0%.

For the nonzero treatment effect conditions, we evaluated our model in scenarios where one arm had no treatment effect with the other having a treatment effect and where both arms had a treatment effect. We will henceforth reference the conditions where one arm had a treatment effect as "One Different Arm" and the conditions where both arms had a treatment effect as "Two Different Arms." We further investigated the impact of using RAR and allowing for early stopping at an interim analysis due to overwhelming efficacy or futility. Our simulations were conducted in R using the R library nimble, where the model was provided with 10,000 burn-in iterations, 15,000 total iterations, with the samples not thinned. As such, inference is based off 5000 iterations. We generated 10,000 datasets per condition.

6. Results

Tables 1–4 show the probability of declaring either active treatment arm as providing statistically significant benefit relative to control. Table 1 presents the results for a design which uses RAR and interim analyses to allow for early stopping. Table 2 shows the results for a design which uses RAR but does not have interim analyses to allow for early stopping. Table 3 presents the results for a design which does not uses RAR but does have interim analyses to allow for early stopping. Table 4 shows the results for a design which does not use RAR and does not have interim analyses to allow for early stopping. From these tables, we see that Type I error is near 1.25% per arm for all conditions where there is no treatment effect. For the Two Different Arms simulation, it is worth noting that the probability of declaring at least one arm significant would be the sum of the three probabilities listed in each of the tables. When comparing the results in Tables 1 and 2 to the results in Tables 3 and 4, we can see that the utilization of RAR provides an increase in power when there is one treatment arm which provides benefit relative to control with the other arm not providing benefit. Further, we see that utilization of the Hwang-Shih-DeCani spending functions for early stopping criteria results in no significant changes in the Type I error and power when compared to designs which do not allow for early termination. Lastly, Table 5 shows the average sample size for the designs which have interim analyses to allow for early stopping and demonstrates that there is a substantial reduction in the expected number of enrolled subjects when allowing for early termination for overwhelming efficacy or futility.

Table 1. Probability of declaring a treatment arm successful across simulation conditions for designs with RAR and interim analyses which allow for early stopping.

Simulation	Condition	Expected Score	Wilcoxon rank sum	*t*-test	Proportional odds
Null	1	0.013|0.012|0.001	0.013|0.011|0	0.013|0.011|0.001	0.013|0.011|0.001
One different	2	0|0.898|0.002	0|0.824|0.002	0|0.895|0.002	0.001|0.81|0.005
	3	0|0.972|0.002	0|0.958|0.001	0|0.971|0.002	0|0.956|0.002
	4	0.002|0.535|0.004	0.002|0.486|0.004	0.002|0.53|0.004	0.003|0.474|0.005
	5	0|0.985|0.002	0|0.959|0.002	0|0.984|0.001	0|0.955|0.004
	6	0|0.999|0.001	0|0.998|0.001	0|0.999|0.001	0|0.997|0.002
	7	0.012|0|0	0.012|0|0	0.011|0|0	0.015|0|0
	8	0.011|0.004|0	0.011|0.004|0	0.011|0.004|0	0.021|0.002|0
Two different	2	0.29|0.302|0.361	0.275|0.286|0.33	0.289|0.301|0.36	0.271|0.287|0.352
	3	0.267|0.27|0.455	0.275|0.274|0.437	0.267|0.268|0.457	0.267|0.268|0.452
	4	0.271|0.256|0.148	0.243|0.232|0.138	0.269|0.255|0.146	0.251|0.238|0.147
	5	0.238|0.243|0.515	0.256|0.261|0.463	0.237|0.244|0.514	0.245|0.253|0.487
	6	0.064|0.06|0.876	0.145|0.135|0.72	0.067|0.065|0.868	0.125|0.117|0.757
	7	0|0|0	0|0|0	0|0|0	0|0|0
	8	0.003|0.004|0	0.003|0.004|0.001	0.003|0.004|0	0.005|0.006|0.001

NOTE: Cells are interpreted as "Probability of declaring Low Dose KET arm successful and not High Dose KET|Probability of declaring High Dose KET arm successful not Low Dose KET|Probability of declaring both Low and High Dose KET arms successful."

Table 2. Probability of declaring a treatment arm successful across simulation conditions for designs with RAR and no interim analyses to allow for early stopping.

Simulation	Condition	Expected Score	Wilcoxon rank sum	*t*-test	Proportional odds
Null	1	0.012|0.012|0.001	0.012|0.011|0.002	0.012|0.011|0.001	0.012|0.011|0.001
	2	0|0.899|0.014	0|0.83|0.013	0|0.897|0.014	0|0.814|0.02
	3	0|0.968|0.011	0|0.955|0.011	0|0.967|0.011	0|0.952|0.013
	4	0|0.546|0.012	0|0.497|0.012	0|0.543|0.012	0.001|0.485|0.015
	5	0|0.978|0.012	0|0.956|0.011	0|0.978|0.011	0|0.949|0.017
One different	6	0|0.988|0.012	0|0.988|0.012	0|0.988|0.012	0|0.982|0.018
	7	0.013|0|0	0.012|0|0	0.012|0|0	0.015|0|0
	8	0.011|0.003|0.001	0.01|0.003|0.001	0.011|0.003|0.001	0.017|0.002|0.001
Two different	2	0.066|0.072|0.819	0.087|0.094|0.719	0.067|0.073|0.816	0.078|0.087|0.754
	3	0.027|0.027|0.94	0.036|0.037|0.916	0.028|0.027|0.938	0.033|0.033|0.924
	4	0.16|0.16|0.366	0.151|0.147|0.327	0.161|0.157|0.363	0.151|0.148|0.347
	5	0.013|0.016|0.968	0.031|0.035|0.917	0.013|0.016|0.968	0.026|0.031|0.93
	6	0|0|1	0.002|0.001|0.996	0|0|1	0.002|0.001|0.997
	7	0|0|0	0|0|0	0|0|0	0|0|0
	8	0.002|0.004|0	0.002|0.004|0.001	0.002|0.004|0	0.004|0.006|0.001

NOTE: Cells are interpreted as "Probability of declaring Low Dose KET arm successful and not High Dose KET|Probability of declaring High Dose KET arm successful not Low Dose KET|Probability of declaring both Low and High Dose KET arms successful."

Table 3. Probability of declaring a treatment arm successful across simulation conditions for designs with no RAR and interim analyses to allow for early stopping.

Simulation	Condition	Expected Score	Wilcoxon rank sum	*t*-test	Proportional odds
Null	1	0.012|0.01|0.001	0.011|0.01|0.001	0.012|0.01|0	0.011|0.01|0.001
One different	2	0|0.875|0.002	0|0.792|0.002	0|0.871|0.001	0.001|0.77|0.004
	3	0|0.963|0.002	0|0.948|0.002	0|0.963|0.002	0|0.943|0.003
	4	0.002|0.512|0.004	0.001|0.462|0.003	0.001|0.509|0.003	0.002|0.447|0.004
	5	0|0.974|0.002	0|0.939|0.002	0|0.974|0.002	0|0.93|0.004
	6	0|0.998|0.002	0|0.997|0.001	0|0.999|0.001	0|0.995|0.003
	7	0.014|0|0	0.012|0|0	0.013|0|0	0.017|0|0
	8	0.012|0.003|0.001	0.011|0.004|0.001	0.012|0.004|0.001	0.022|0.003|0.001
Two different	2	0.306|0.295|0.356	0.287|0.281|0.331	0.305|0.296|0.354	0.287|0.278|0.352
	3	0.264|0.265|0.464	0.27|0.274|0.443	0.263|0.264|0.466	0.261|0.264|0.463
	4	0.269|0.269|0.148	0.246|0.24|0.136	0.266|0.268|0.146	0.25|0.25|0.144
	5	0.237|0.231|0.527	0.259|0.253|0.467	0.239|0.236|0.521	0.253|0.247|0.484
	6	0.064|0.059|0.877	0.133|0.136|0.731	0.066|0.064|0.869	0.116|0.116|0.768
	7	0|0.001|0	0|0.001|0	0|0.001|0	0|0.001|0
	8	0.003|0.004|0	0.004|0.004|0	0.003|0.004|0	0.006|0.006|0.001

NOTE: Cells are interpreted as "Probability of declaring Low Dose KET arm successful and not High Dose KET|Probability of declaring High Dose KET arm successful not Low Dose KET|Probability of declaring both Low and High Dose KET arms successful."

Table 4. Probability of declaring a treatment arm successful across simulation conditions for designs with no RAR and no interim analyses to allow for early stopping.

Simulation	Condition	Expected Score	Wilcoxon rank sum	*t*-test	Proportional odds
Null	1	0.013|0.01|0.002	0.012|0.009|0.002	0.012|0.009|0.002	0.011|0.009|0.002
One different	2	0|0.876|0.011	0|0.8|0.011	0|0.876|0.01	0|0.771|0.017
	3	0|0.959|0.012	0|0.948|0.011	0|0.959|0.012	0|0.941|0.015
	4	0.001|0.522|0.013	0|0.472|0.012	0.001|0.517|0.013	0.001|0.454|0.015
	5	0|0.969|0.013	0|0.934|0.013	0|0.968|0.013	0|0.925|0.018
	6	0|0.987|0.013	0|0.986|0.012	0|0.987|0.013	0|0.98|0.019
	7	0.014|0|0	0.013|0|0	0.013|0|0	0.017|0|0
	8	0.011|0.004|0.001	0.011|0.004|0.001	0.011|0.004|0.001	0.02|0.002|0.002
Two different	2	0.068|0.061|0.832	0.085|0.086|0.734	0.068|0.061|0.83	0.079|0.079|0.766
	3	0.022|0.024|0.948	0.03|0.034|0.924	0.022|0.025|0.946	0.028|0.031|0.93
	4	0.157|0.158|0.381	0.148|0.147|0.338	0.157|0.156|0.377	0.149|0.147|0.357
	5	0.015|0.013|0.968	0.037|0.034|0.912	0.015|0.014|0.967	0.03|0.03|0.926
	6	0|0|1	0.001|0.001|0.998	0|0|1	0.001|0.001|0.998
	7	0|0|0	0|0|0	0|0|0	0|0|0
	8	0.003|0.003|0.001	0.003|0.003|0.001	0.003|0.003|0.001	0.004|0.004|0.002

NOTE: Cells are interpreted as "Probability of declaring Low Dose KET arm successful and not High Dose KET|Probability of declaring High Dose KET arm successful not Low Dose KET|Probability of declaring both Low and High Dose KET arms successful."

Table 5. Average sample size across simulation conditions for designs with interim analyses which allow for early stopping.

Simulation	Condition	Expected Score	Wilcoxon rank sum	*t*-test	Proportional odds
Null	1	406.68 |410.41	406.81 |410.13	406.21 |410.12	406.87 |410.22
One different	2	428.6 |436.21	453.77 |460.42	432.97 |439.24	460.53 |469.08
	3	381.84 |389.89	395.83 |405.19	384.72 |392.46	398.42 |409.2
	4	493.11 |491.24	495.05 |494.51	494.6 |492.83	497.07 |496.69
	5	365.99 |373.54	398.88 |407.73	369.77 |375.62	403.55 |415.54
	6	307.59 |310.06	326.57 |334.16	308.53 |311.03	328.67 |338.29
	7	379.49 |381.69	379.39 |381.48	379.39 |381.55	380.37 |382.83
	8	392.78 |395.4	389.95 |392.75	392.33 |395.01	391.52 |394.9
Two different	2	400.11 |397.47	433.44 |430.88	404.08 |401	418.3 |415.75
	3	354.41 |351.78	369.85 |367.39	356.7 |354.58	364.28 |362.36
	4	482.94 |483.96	491.68 |493.23	485.34 |485.98	486.35 |487.54
	5	342.21 |343.09	374.61 |376.43	344.92 |345.72	365.09 |365.83
	6	303.33 |303.14	317.01 |317.18	303.71 |303.65	312.98 |313.03
	7	331.12 |332.06	331.46 |332.5	330.99 |331.99	331.7 |332.94
	8	375.09 |376.61	370.78 |371.56	374.93 |376.38	372.64 |373.53

Cells are interpreted as "Average sample size for design with interim analysis with early stopping criteria and with RAR |Average sample size for design with interim analysis with early stopping criteria and without RAR".

7. Discussion

In order to assess our proposed Bayesian, adaptive design for an ordinal outcome, we conducted a simulation study for a wide range of distributions of ordinal scores, including some extreme cases that are not clinically expected (Conditions 4 and 8). The power and Type I error rate were well controlled under all scenarios of clinical interest. With this design we were able to use a Bayesian framework for the ordinal outcome and reduce the average sample size through the inclusion of pre-specified adaptations. Further, we found that our proposed approach is similar to the *t*-test, with both approaches having similar Type I error, power, and average sample sizes. Additionally, it is worth nothing that the results for the proportional odds model are similar to that of the Wilcoxon rank sum test, which is consistent with the finding that the proportional odds model is asymptotically equivalent to the rank sum test regardless of if the proportional odds assumption holds (Harrell 2015).

All simulations assumed the same distribution for the control group (LEV), which based on previous studies was expected to be bimodal and concentrated in both extremes. It is likely that a different underlying control group distribution of scores or changes in assumed treatment effect would result in different operating characteristics, including scenarios in which the normality assumption or proportional odds may be optimal. Nevertheless, these results are encouraging and provide an approach which may prove useful for other settings. As with the rank sum test, our approach does not immediately generalize into a framework for which we can adjust for baseline covariates.

Many find ordinal outcomes more difficult to interpret than a binary response rate. This could be because of confusion around the meaning of measures of central tendency for ordinal outcomes, if they are not commonly in use. The ordinal outcome for KESETT is new and was developed by the KESETT clinical investigators following the desirability of outcome ranking (DOOR) approach (Evans et al. 2015; Evans and Follmann 2016). The expected score is approximately equivalent to the mean of the ordinal outcome and can be analyzed without assuming normality of the mean. For the KESETT design, the clinically important difference in the outcome probabilities represented a difference in expected scores of 0.43 (2.45 for LEV alone vs. 2.02 for add-on treatment group). Changes to the coding of this scale would affect the interpretation of the expected score, as well as the operating characteristics. These scale codings should be clinically determined and benchmarked with previous large trials to define clinically important differences.

Motivated by a trial in status epilepticus, we have developed a Bayesian, response-adaptive design for ordinal data which is applicable without the assumptions of normality of the mean or proportionality of the odds. Our approach performs similarly to the t-test and maintains optimal operating characteristics for our defined clinically important difference.

Funding

This work was supported by the National Institute of Neurological Disorders and Stroke (NINDS) under grants U01NS088034, U01NS088023, and U24NS100655.

References

Chaisinanunkul, N., Adeoye, O., Lewis, R. J., Grotta, J. C., Broderick, J., Jovin, T. G., et al. (2015), "Adopting a Patient-Centered Approach to Primary Outcome Analysis of Acutre Stroke Trials Using a Utility-Weighted Modified Rankin Scale," *Stroke*, 46, 2238–2243. [1]

Chamberlain, J. M., Kapur, J., Shinnar, S., Elm, J., Holsti, M., Babcock, L., et al. (2020), "Efficacy of Levetiracetam, Fosphenytoin, and Valproate for Established Status Epilepticus by Age Group (ESETT): A Double-Blind, Responsive-Adaptive, Randomised Controlled Trial," *Lancet*, 395, 1217–1224. [2,3]

Deeds, S. I., Barreto, A., Elm, J., Derdeyn, C. P., Berry, S., Khatri, P., Moy, C., Janis, S., Broderick, J., Grotta, J., and Adeoye, O. (2020), "The Multiarm Optimization of Stroke Thrombolysis Phase 3 Acute Stroke Randomized Clinical Trial: Rationale and Methods," *International Journal of Stroke*, 16, 873–880. [1,2]

Evans, S. R., and Follmann, D. (2016), "Using Outcomes to Analyze Patients Rather than Patients to Analyze Outcomes: A Step Towards Pragmatism in Benefit-Risk Evaluation," *Statistics in Biopharmaceutical Research*, 8, 386–393. [1,6]

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. [1,6]

Harrell, F. (2015), "Ordinal Logistic Regression," in *Regression Modeling Strategies*. Springer Series in Statistics, pp. 311–325, Cham: Springer. [5]

Hwang, I. K., Shih, W. J., and De Cani, J. S. (1990), "Group Sequential Designs Using a Family of Type I Error Probability Spending Functions," *Statistics in Medicine*, 9, 1439–1445. [2]

Jovin, T. G., Saver, J. L., Ribo, M., Pereira, V., Furlan, A., Bonafe, A., et al. (2015), "Diffusion-Weighted Imaging or Computerize Tomogrpahic Perfusion Assessment with Clinical Mismatch in the Triage of Wake Up and Alte Presenting Strokes Undergoing Neurointervention with Trevo (DAWN) Trial Methods," *International Journal of Stroke*, 12, 641–652. [1,2]

Kapur, J., Elm, J., Chamberlain, J. M., Barsan, W., Cloyd, J., Lowenstein, D., et al. (2019), "Randomized Trial of Three Anticonvulsant Medications for Status Epilepticus," *New England Journal of Medicine*, 22, 2103–2113. [2,3]

Saver, J. L. (2011), "Optimal End Points for Acute Stroke Therapy Trials: Best Ways to Measure Treatment Effects of Drugs and Devices," *Stroke*, 11, 2356–2362. [1]

Shi, H., and Yin, G. (2021), "Reconnecting p-value and Posterior Probability Under One- and Two-Sided Tests," *The American Statistician*, 75, 265–275. [2]