Title: Analysis of an ordinal endpoint for use in evaluating treatments for severe influenza requiring hospitalization

Authors: Ross L. Peterson[1], David M. Vock[1], John H. Powers III[2], Sean Emery[3], Eduardo Fernandez Cruz[4], Sally Hunsberger[5], Mamta K. Jain[6], Sarah Pett[7], James D. Neaton[1]

for the INSIGHT FLU-IVIG Study Group

[1]University of Minnesota School of Public Health, Division of Biostatistics, Minneapolis, MN, USA.

[2] George Washington University School of Medicine, Washington D.C., USA.

[3]Kirby Institute, University of New South Wales, Sydney, Australia.

[4]Hospital General Universitario Gregorio Marañón, Instituto de Investigación Sanitaria Gregorio Marañón, Departamento de Microbiología I /Inmunología, Facultad de Medicina, Universidad Complutense de Madrid.

[5]National Institute of Allergy and Infectious Disease, Biostatistics Research Branch, Rockville, Maryland, USA.

[6]UT Southwestern Medical Center, Department of Internal Medicine, Dallas, Texas, USA.

[7]CRG, Infection and Population Health, UCL and MRC CTU at UCL, University College London, London, UK; Kirby Institute, UNSW, Australia.

Author for correspondence: Ross Peterson, University of Minnesota School of Public Health, Division of Biostatistics, 420 Delaware St. SE MMC 303, Minneapolis, MN, USA.

E-mail: pet00180@umn.edu

Phone Number: 202-641-3048

Funding provided by subcontract 13XS134 under Leidos Biomed's Prime Contract HHSN261200800001E and HHSN2612015000031, NCI/NIAID.

Abstract

Background/Aims A single best endpoint for evaluating treatments of severe influenza requiring hospitalization has not been identified. A novel 6-category ordinal endpoint of patient status is being used in a randomized controlled trial (FLU-IVIG) of intravenous immunoglobulin (IVIG). We systematically examine four factors regarding the use of this ordinal endpoint that may affect power from fitting a proportional odds model: 1) deviations from the proportional odds assumption which result in the same overall treatment effect as specified in the FLU-IVIG protocol and which result in a diminished overall treatment effect; 2) deviations from the distribution of the placebo group assumed in the FLU-IVIG design; 3) the effect of patient misclassification among the 6 categories; and 4) the number of categories of the ordinal endpoint. We also consider interactions between the treatment effect (i.e., Factor 1) and each other factor.

Methods We conducted a Monte Carlo simulation study to assess the effect of each factor. To study factor 1, we developed an algorithm for deriving distributions of the ordinal endpoint in the two treatment groups that deviated from proportional odds while maintaining the same overall treatment effect. For factor 2, we considered placebo group distributions which were more or less skewed than the one specified in the FLU-IVIG protocol by adding or subtracting a constant from the cumulative log odds. To assess factor 3, we added misclassification between adjacent pairs of categories that depend on subjective patient/clinician assessments. For factor 4, we collapsed some categories into single categories.

Results Deviations from proportional odds reduced power at most from 80% to 77% given the same overall treatment effect as specified in the FLU-IVIG protocol. Misclassification and collapsing categories can reduce power by over 40 and 10 percentage points, respectively, when they affect categories with many patients and a discernible treatment effect. But, collapsing categories that contain no treatment effect can raise power by over 20 percentage points. Differences in the distribution of the placebo group can raise power by over 20 percentage points or reduce power by over 40 percentage points depending on how patients are shifted to portions of the ordinal endpoint with a large treatment effect.

Conclusions Provided that the overall treatment effect is maintained, deviations from proportional odds marginally reduce power. However, deviations from proportional odds can modify the effect of misclassification, the number of categories, and the distribution of the placebo group on power. In general, adjacent pairs of categories with many patients should be kept separate to help ensure that power is maintained at the pre-specified level.

Keywords: clinical trials, endpoints, proportional odds model, misspecified model, statistical power

Introduction

Influenza causes 226,000 excess hospitalizations and more than 500,000 deaths worldwide.[1] In spite of the large disease burden, no study has definitively demonstrated substantial clinical efficacy of an antiviral drug in hospitalized influenza patients.[2] For this subpopulation, the proportion of patients dying is small, increasing the challenge to demonstrate treatment effects with all-cause mortality as the sole endpoint. Therefore, the United States Food and Drug Administration (FDA) recommends that the primary endpoint of randomized controlled trials evaluating new treatments include any of the following measures: clinical signs and symptoms, duration of hospitalization, time to normalization of vital signs and oxygenation, requirements for supplemental oxygen or assisted ventilation, and mortality. FDA guidance further states that no single best endpoint has been identified for studying treatments in patients hospitalized by influenza.[2]

Primary endpoints in randomized trials of treatments for hospitalized influenza patients have included continuous measures of virologic activity, time to event outcomes (e.g., time to clinical stability) and binary outcomes (e.g., proportion of patients returning to premorbid status).[3–7] Following the successful completion of a pilot study of intravenous hyperimmune immunoglobulin (IVIG),[8] the International Network for Strategic Initiatives in Global HIV Trials (INSIGHT) initiated a trial of IVIG (FLU-IVIG) to evaluate its efficacy in patients hospitalized with influenza (NCTO2287467).[9] To conduct a study with a feasible sample size and to improve the likelihood of demonstrating benefit relative to a binary outcome, a novel ordinal outcome of patient status serves as the primary endpoint of FLU-IVIG. The ordinal endpoint constructs categories of various

outcome assessments ranked in order of patient status (e.g., from death to resumption of normal activities). To our knowledge, an ordinal endpoint of clinical outcomes has not been used in influenza trials.

To calculate the sample size for a trial with an ordinal endpoint, researchers must make a number of design decisions and assumptions. The purpose of this paper is to describe the ordinal endpoint used in the FLU-IVIG study and to consider the impact of four factors on power: 1) deviations from the proportional odds assumption which result in the same overall treatment effect as specified in the FLU-IVIG protocol and which result in a diminished overall treatment effect; 2) deviations from the distribution of the placebo group that researchers expect to observe in the FLU-IVIG protocol; 3) the effect of patient misclassification among the 6 categories; and 4) the number of categories of the ordinal endpoint. In addition to examining these factors separately, we also consider the effect of interactions between the treatment effect (i.e., factor 1) and each of the other factors.

Methods

The FLU-IVIG study was designed and is being conducted by the INSIGHT Group at sites in the northern and southern hemisphere. FLU-IVIG is a multicenter, double-blind, randomized trial comparing treatment with IVIG versus placebo in hospitalized patients with locally confirmed influenza A or B who have a National Early Warning Score (NEWS) of two or higher.[10] For patients in both groups, the randomized treatment is administered in addition to standard of care treatment which includes anti-viral treatment.

The primary objective is to compare outcomes of patients in the IVIG and placebo groups at 7 days after randomization using an ordinal endpoint constructed with the following 6 mutually exclusive categories:

1) death;

2) intensive care unit hospitalization (In ICU);

3) non-ICU hospitalization, requiring supplemental oxygen;

4) non-ICU hospitalization, not requiring supplemental oxygen;

5) discharged from the hospital, but unable to resume normal activities;

6) discharged from the hospital with resumption of normal activities.

The categories were defined to delineate clear improvement and worsening in patient status and to yield a sufficient spread of the data for showing benefit due to IVIG. Day 7 was chosen as the time point for comparison of ordinal endpoints because pilot data had established that differences between treatment groups in influenza antibody titer levels were greatest compared to placebo in the first few days after treatment with IVIG.[8]

To estimate the sample size, we used data from a cohort study of patients hospitalized with influenza at many of the same sites participating in the FLU-IVIG trial to predict the distribution of the ordinal endpoint in the placebo group for FLU-IVIG.[11–12] The distribution of the ordinal endpoint at day 7 for patients in the cohort study who met the

FLU-IVIG trial inclusion/exclusion criteria is given in Table 1. Because the cohort study is still in progress, we derived a more recent placebo group distribution (updated as of September 1st, 2015) relative to the FLU-IVIG protocol. Therefore, the category percentages used in our investigation differ slightly from the FLU-IVIG protocol (for the category percentages of the placebo group specified in the FLU-IVIG protocol, see the second footnote of Table 1). We refer to these category percentages as the FLU-IVIG design estimates. The FLU-IVIG protocol specifies that a proportional odds model will be used to evaluate the effect of IVIG. Under the proportional odds assumption of the model, the treatment effect is constant across categories between randomized groups. That is, the model assumes that the ratio of the odds of any better versus worse division of the ordinal endpoint (e.g., alive versus dead, discharged versus hospitalized or dead) between IVIG and placebo is constant. In FLU-IVIG, a treatment effect corresponding to a log odds ratio of 0.57 was deemed of interest and attainable. A log odds ratio greater than 0 indicates benefit due to IVIG. Under the proportional odds model, the percentage of subjects in each ordered category for the IVIG group is shown in Table 1 assuming a log odds ratio of 0.57.

Even if the proportional odds assumption is not reasonable, the estimated log odds ratio from erroneously assuming a proportional odds model is still a valid measure of treatment efficacy. In particular, the log odds ratio can be interpreted as the average shift over the 6 ordered categories caused by IVIG and the score test of the log odds ratio is equivalent to the well-known nonparametric Wilcoxon rank-sum test.[13]

In order to detect a log odds ratio of 0.57 assuming proportional odds with 80% power at the 0.05 (2-sided) level of significance, a sample size of 320 patients is required.[14] For reference, Supplemental Table 1 gives the power for detecting different log odds ratios with a sample size of 320 patients. Supplemental Table 2 gives the power for detecting a significant treatment effect under each possible way of dividing the ordinal endpoint into a binary endpoint. The power from each binary endpoint is substantially less than the power from the ordinal endpoint.

Simulation Study Design

We first derived different distributions of the ordinal endpoint in the placebo and IVIG groups under different scenarios of each factor which we describe below. For each scenario, we ran 10,000 simulations of the clinical trial assuming that 320 patients were sampled from the corresponding placebo and IVIG group distributions. For each simulated trial, we analyzed the data assuming a proportional odds cumulative logistic model and computed a Wald test statistic for the treatment effect. The empirical power is the proportion of the 10,000 simulations for which the Wald test statistic was significant. With this approach, estimates of power do not require any large sample approximations.

The reference level of the factors in our simulation experiment corresponds to the assumptions used in the sample size calculation for the FLU-IVIG design; that is, the proportional odds assumption holds, the distribution of the placebo group is determined from the cohort study, no misclassification of patients among the categories of the ordinal endpoint occurs, and the full 6-level ordinal endpoint is used.

Factor 1: Treatment Effect

We first sought to derive distributions of the IVIG group that deviated from proportional odds while maintaining the same overall treatment effect as specified in the FLU-IVIG protocol. By overall treatment effect, we mean the average (across repeated experimentation) estimated log odds ratio for the effect of IVIG relative to placebo from fitting the proportional odds model to the data. We refer to this as the average log odds ratio but note that this is not the arithmetic mean of the log odds ratio for every possible binary division of the ordinal endpoint and is a nonlinear function of the probabilities of each category in the ordinal endpoint in the placebo and IVIG groups.

For large samples, the average log odds ratio of a misspecified proportional odds model is the value for which the expected score function equals zero. Therefore, we can constrain the distribution of the IVIG group such that the average log odds ratio is maintained across deviations from proportional odds (see the Appendix for a derivation). We created a novel algorithm which, given the desired average log odds ratio, the distribution of the ordinal endpoint in the placebo group and the proportions of observations in all but two categories of the treatment group, returns the proportions in the final two categories of the treatment group to maintain the desired average log odds ratio. Code to implement our algorithm in the programming language R is available as a GitHub repository (https://github.com/RPeterson4/Supplementary-Code-for-Evaluating-the-Ordinal-Endpoint-for-FLU-IVIG).

We considered three treatment effect scenarios (T1–T3 below) that deviate from proportional odds while maintaining an average log odds ratio of 0.57 assuming the other three factors are not altered. The deviation from proportional odds is strong enough in each of these scenarios to yield, on average, a significant p-value for the test of the proportional odds assumption (at the 0.05 level) across the samples. We also considered two treatment effect scenarios (T4 and T5) with a log odds ratio of 0.57 across a subset of all possible binary divisions of the categories of the ordinal endpoint but zero elsewhere. In these scenarios, the overall treatment effect is diminished. The six treatment effect scenarios are:

- T0: Proportional odds is satisfied and the average log odds ratio is 0.57 (FLU-IVIG design assumption).
- T1: The treatment effect constantly weakens across the ordinal endpoint. The log odds ratio is 2.6 between the binary outcome of alive and dead patients, and then constantly decreases by 0.6 with each successive binary division of the ordinal endpoint (e.g., the log odds ratio is 2.0 for Hospitalized, not in ICU, on oxygen or better versus death or in ICU).
- T2: The treatment effect is constant and positive across the most severe categories of the ordinal endpoint. Specifically, the log odds ratio of 1.16 for the first four binary divisions of the ordinal endpoint (ordering the scale from most severe outcome to least severe). There is no treatment effect for the last binary division (Discharged, back to normal activities or worse versus Discharged, not back to normal activities).
- T3: The treatment only benefits patients in the discharged categories. That is, the log odds ratio is 1.16 for the last binary division and 0 for all other binary divisions.

- T4: The log odds ratio is 0.57 for the first four binary divisions and 0 for the last binary division.
- T5: The log odds ratio is 0.57 for the last binary division and 0 for the first four binary divisions.

Factor 2: Distribution of the Placebo Group

To systematically alter the distribution of the placebo group, note that the cumulative log odds of being in a more versus less severe category for each possible binary split of the ordinal endpoint (see Supplemental Table 3) uniquely determines the placebo group distribution. To derive different distributions of the placebo group, we added or subtracted a constant from each of the cumulative log odds of being in a more versus less severe category from the placebo group design estimate (see the Appendix for a derivation). Adding (subtracting) a constant increases the proportion of patients with more (less) severe outcomes of the ordinal endpoint. Note that 62.9% of subjects are in the discharged categories of the ordinal endpoint for the placebo group design estimate. Therefore, having more (fewer) patients in more severe categories will yield a less (more) skewed distribution. The five distributions of the placebo group are:

- P0: The placebo group distribution for the FLU-IVIG design.
- P1: Add 0.5 to the cumulative log odds of P0 (less skewed distribution).
- P2: Add 1 to the cumulative log odds of P0 (less skewed distribution).
- P3: Subtract 0.5 from the cumulative log odds of P0 (more skewed distribution).
- P4: Subtract 1 from the cumulative log odds of P0 (more skewed distribution).

Factor 3: Misclassification

For our purposes, we studied misclassification among adjacent pairs of categories that may be difficult to distinguish between for significant numbers of patients. This misclassification may result from a combination of the subjective nature of the categories, inconsistent clinician judgment, and patients' memory of their recovery. To study the effect of misclassification, we considered scenarios investigated by Whitehead who supposed 20% misclassification between two pairs of categories.[14] Whitehead represented misclassification by exchanging certain percentages of patients between categories that could be misclassified (see the Appendix for an example).

Here, we assumed that 20% and 40% of patients in the non-ICU hospitalized categories and the discharged categories could be misclassified. We chose the non-ICU hospitalized categories because use of oxygen during the day can be variable, and the discharged categories for depending on the patient's memory of when they resumed normal activities. We assumed the misclassification rate to be constant across both randomized groups because the study is double-blind (i.e., nondifferential misclassification). We also considered scenarios in which 20% misclassification affected either the non-ICU hospitalized categories or the discharged categories but not the other. The type I error rate does not change under the nondifferential misclassification we assumed. The five levels of misclassification are:

- M0: No misclassification (FLU-IVIG design assumption).

- M1: 20% misclassification between the non-ICU hospitalized categories and the discharged categories.
- M2: 40% misclassification between the non-ICU hospitalized categories and the discharged categories.
- M3: 20% misclassification between the non-ICU hospitalized categories.
- M4: 20% misclassification between the discharged categories

Factor 4: Number of Categories

Misclassification between adjacent pairs of categories can be eliminated by collapsing each into a single category. Thus, we examined collapsing the non-ICU hospitalized categories and the discharged categories. Furthermore, the discharged categories contain the largest percentage of patients in each scenario on average, implying that collapsing them may have an outsize effect on power. Conversely, we collapsed the four most severe categories because they contain the smallest percentage of patients. We also collapsed the ordinal endpoint into a binary hospitalized or dead versus discharged endpoint, which is a clinically relevant cut-point. The six levels of collapsing categories are:

- C0: The full 6-category ordinal endpoint (FLU-IVIG design assumption).
- C1: Collapse the non-ICU hospitalized categories and the discharged categories.
- C2: Collapse the non-ICU hospitalized categories.
- C3: Collapse the discharged categories.
- C4: Collapse the hospitalization or death categories.

- C5: Collapse the hospitalization or death categories and the discharged categories to make for a binary endpoint.

Interactions

As detection of the treatment effect is of primary interest for the FLU-IVIG trial, we explored the effect on power if the treatment effect (factor 1) and the placebo group distribution, misclassification, or number of categories also deviated from the levels assumed in the design of FLU-IVIG. This yielded three groups of two-way interactions: 1) treatment effect scenarios and distributions of the placebo group; 2) treatment effect scenarios and levels of misclassification; and 3) treatment effect scenarios and number of categories. Due to the interacting factors, the overall treatment effect may differ from the 0.57 log odds ratio specified in the FLU-IVIG protocol.

Results

Main Effects

Provided that the average log odds ratio was maintained, treatment effect scenarios that violated proportional odds only marginally reduced power (see Table 1). For example, under treatment effect scenario T2 in which the treatment benefit is only evident over the most severe categories of the ordinal endpoint (hospitalization or death categories), power declined from 80% to 77.3%. However, both scenarios in which the log odds ratio was 0.57 for some binary divisions but 0 for the rest greatly reduced power, mainly due

to the decline in the average log odds ratio from 0.57 (0.31 and 0.25 under T4 and T5, respectively).

Changes in the distribution of the placebo group from the FLU-IVIG design led to moderate differences in power. Less skewed placebo group distributions yielded slightly higher power, from 80% to 80.9% and 81.9% in scenarios P1 and P2, respectively (see Table 1). Conversely, distributions of the placebo group which were more skewed led to modest declines in power, with the most skewed distribution returning the largest loss of power from 80% to 73.8%.

Table 2 shows that misclassification among the categories always reduced power by lowering the average log odds ratio. Scenarios in which there was greater misclassification, the misclassification involved more categories, or the misclassification was between categories containing many patients decreased power the most. For example, limiting the misclassification to the discharged categories, which comprise 62.9% of patients in the distribution of the placebo group, reduced power from 80% to 70.1%. Expanding the 20% misclassification to include the non-ICU hospitalized categories, which together contain 30.6% of patients, only additionally reduced power from 70.1% to 69.7%.

Reducing the number of categories always lowered power (see Table 3). Generally, power declined more when multiple categories or categories with many patients were combined. For example, collapsing the discharged categories reduced power from 80% to 65.6%, while having a binary hospitalized or dead versus discharged endpoint reduced

power from 80% to 63.9%. Collapsing the non-ICU hospitalized categories or the four most severe categories did not substantially reduce power, mainly due to the small percentage of patients in those categories.

Interactions

The power for all possible combinations of the interacting factors for the three groups of two-way interactions considered is given in Supplemental Tables 4-6. From these, we selected a subset from each group of interactions for further investigation based on their effect on power and clinical relevance. We present our findings in Tables 4-6.

Many of the interactions between the treatment effect and each of the other factors were qualitative, that is the direction of the main effect on power changed when an additional factor was altered. For example, Table 4 demonstrates that the effect of deviations from proportional odds on power may change with different placebo group distributions. Power substantially increased (decreased) when treatment effect scenarios were paired with distributions of the placebo group that had more (fewer) patients in categories influenced by the treatment. For example, under treatment effect scenario T2 in which the treatment benefit is only evident for the hospitalization or death categories, having a less skewed placebo group distribution (i.e., more hospitalized or dead patients) raised power from 77.3% to 99.7% (see Tables 1 and 4). On the other hand, having a more skewed placebo group distribution reduced power from 77.3% to 16.8%.

Similarly, Table 5 shows that misclassification may not reduce power when coupled with deviations from proportional odds. In some cases, it may even raise power. Under

treatment effect scenario T2, 20% and 40% misclassification between the non-ICU hospitalized categories and the discharged categories raised power from 77.3% to 86.2% and 92.7%, respectively (see Tables 1 and 5). This is likely because without misclassification the scenario assumes no treatment effect for patients discharged from the hospital. Misclassification, though, evens the proportions between patients who have and have not resumed normal activities, creating the illusion that the treatment has shifted patients into resuming normal activities. Consequently, the log odds ratio for Not Normal or worse versus Normal increased from 0 to raise the average log odds ratio and power.

Additionally, Table 6 demonstrates that for scenarios in which the treatment effect was absent across a range of the ordinal endpoint, collapsing the corresponding categories raised power by increasing the average log odds ratio. Under treatment effect scenarios T1, T2, and T4 (scenarios in which the intervention primarily shows benefit for the hospitalization or death categories), collapsing the discharged categories increased power from 79.1% to 95.8%, 77.3% to 99.5%, and 33.1% to 65.6%, respectively (see Tables 1 and 6). Conversely, collapsing categories over ranges of the ordinal endpoint with a discernible treatment effect reduced power. For example, under treatment effect scenarios T1, T2, and T4, collapsing the four most severe categories reduced power from 79.1% to 67.0%, 77.3% to 76.0%, and 33.1% to 32.0%, respectively.

Comparing Tables 2 and 3, collapsing the non-ICU hospitalized categories and the discharged categories to eliminate potential misclassification yielded greater power than using the 6-level ordinal endpoint when misclassification between both pairs of categories was 40% (65.1% versus 57.7%). When misclassification was limited to the

non-ICU hospitalized categories at the 20% level, collapsing those categories yielded approximately equal power compared to using the 6-level ordinal endpoint. Limiting 20% misclassification to the discharged categories generated greater power for the 6-level endpoint relative to collapsing those categories (70.1% versus 65.6%).

Discussion

To our knowledge, the FLU-IVIG study is the first randomized trial to use an ordinal endpoint to evaluate a novel influenza treatment. Thus, we considered it necessary to thoroughly examine the ordinal endpoint with respect to factors that may affect its statistical power for the trial. Our evaluation has found that the ordinal endpoint yields higher power relative to any collapse of the ordinal endpoint into a binary endpoint. Further, the decisions about the number of categories and assumptions made about the treatment effect, distribution of the placebo group, and the amount of misclassification can have substantial consequences for power. Provided that the overall treatment effect is maintained and other factors are held constant, deviations from proportional odds marginally reduce power. We also found that, holding other factors constant, more skewed placebo group distributions, misclassification of patients among the ordinal categories, and considering fewer ordinal categories decreased power consistent with previous research.[14–16]

However, our analysis has shown that these general conclusions must be qualified as the effect of each of these factors may be reversed when another factor is varied simultaneously. To increase power, if the proportional odds assumption does not hold,

categories in which the treatment is presumed to be effective should be divided as evenly as possible; conversely, categories where the treatment is presumed to be less effective should be collapsed. For IVIG, the treatment may be more beneficial for severe cases. Therefore, we considered an ordinal endpoint which was granular for these patients to attain sufficient power.

In contrast to previous research, [14–16] we explored deviations from proportional odds while holding the overall treatment effect constant. In addition, we studied the joint effect of multiple factors related to design decisions and assumptions about the ordinal endpoint. Previous research has primarily examined the effect of a single factor at a time. Though our results were derived with respect to FLU-IVIG, our novel algorithm and simulation code, which are available for download from GitHub, can be used to evaluate other ordinal endpoints for influenza trials. Other direct measures of patient status, such as those that include complications of influenza (e.g., development of pneumonia while on therapy) and patient-reported outcomes of influenza (e.g., the FLU-PRO instrument),[17] could be used to construct new ordinal endpoints for influenza treatments.

More broadly, ordinal endpoints have been considered for trials studying treatments of vascular disease, streptococcus pneumoniae, and traumatic brain injury.[18–21] In these trials, relative to FLU-IVIG, different parameter values (e.g., treatment effect size) may modify the magnitude of the effect of the four factors evaluated in this paper on power. Furthermore, other statistical methods like the sliding dichotomy, win ratio, and global rank tests may yield different power for detecting a treatment effect along an ordinal endpoint. [22–24] However, we anticipate that our general conclusions will hold when using

the proportional odds model to detect differences in the distribution of an ordinal endpoint across different randomized treatment groups. Moreover, our general approach can be used by future researchers to address concerns about the specification and statistical analysis of ordinal endpoints. In our study, we used the proportional odds model because it remains a standard tool for analyzing ordinal endpoints and will be used in the primary analysis for FLU-IVIG.

Clearly, researchers must consider several factors when designing a clinical trial based on an ordinal endpoint including the number of categories, whether patients can be reliably distinguished between those categories, the anticipated treatment effect, and the distribution of the ordinal endpoint in the placebo group. Simulation studies allow researchers to explore how sensitive power is to decisions and assumptions about these factors. To that end, our general approach for evaluating the FLU-IVIG ordinal endpoint may be useful for examining other ordinal endpoints for influenza trials and other diseases.

Table 1. Main effects of the treatment effect (factor 1) and placebo group distribution (factor 2) on power.

Scenario		Death	In ICU	Hospitalized, not in ICU, on oxygen	Hospitalized, not in ICU, not on oxygen	Discharged, not back to normal activities	Discharged, back to normal activities		
Factor 1: Treatment Effect									
T0: Proportional odds holds (FLU-IVIG design assumption)	% Placebo[a] % IVIG[b] logOR[c]	1.2 0.7 0.57	5.3 3.1 0.57	16.2 10.5 0.57	14.4 10.8 0.57	36.4 36.0 0.57	26.5 39.0	Power (%)[d] Avg. logOR[e]	80.0 0.57
T1: Treatment effect constantly weakens	% Placebo % IVIG logOR	1.2 0.1 2.60	5.3 0.8 2.00	16.2 5.8 1.40	14.4 14.2 0.80	36.4 48.5 0.20	26.5 30.5	Power (%) Avg. logOR	79.1 0.57
T2: Treatment effect limited to the hospitalization or death categories	% Placebo % IVIG logOR	1.2 0.4 1.16	5.3 1.7 1.16	16.2 6.3 1.16	14.4 7.2 1.16	36.4 57.9 0	26.5 26.5	Power (%) Avg. logOR	77.3 0.57
T3: Treatment effect limited to the discharged categories	% Placebo % IVIG logOR	1.2 1.2 0	5.3 5.3 0	16.2 16.2 0	14.4 14.4 0	36.4 9.3 1.16	26.5 53.6	Power (%) Avg. logOR	78.7 0.57
T4: Smaller treatment effect limited to the hospitalization or death categories	% Placebo % IVIG logOR	1.2 0.7 0.57	5.3 3.1 0.57	16.2 10.5 0.57	14.4 10.8 0.57	36.4 48.5 0	26.5 26.5	Power (%) Avg. logOR	33.1 0.31
T5: Smaller treatment effect limited to the discharged categories	% Placebo % IVIG logOR	1.2 1.2 0	5.3 5.3 0	16.2 16.2 0	14.4 14.4 0	36.4 23.9 0.57	26.5 39.0	Power (%) Avg. logOR	23.9 0.25
Factor 2: Distribution of the Placebo Group									
P1: Less skewed placebo group distribution	% Placebo % IVIG logOR	2.0 1.1 0.57	8.3 5.0 0.57	22.4 15.4 0.57	16.6 13.9 0.57	32.7 36.6 0.57	18.0 27.9	Power (%) Avg. logOR	80.9 0.57
P2: Even less skewed placebo group distribution	% Placebo % IVIG logOR	3.2 1.8 0.57	12.7 7.8 0.57	28.5 21.5 0.57	17.1 16.4 0.57	26.7 33.4 0.57	11.7 19.0	Power (%) Avg. logOR	81.9 0.57
P3: More skewed placebo group distribution	% Placebo % IVIG logOR	0.7 0.4 0.57	3.3 1.9 0.57	11.1 6.8 0.57	11.2 7.7 0.57	36.3 31.9 0.57	37.3 51.3	Power (%) Avg. logOR	78.6 0.57
P4: Even more skewed placebo group distribution	% Placebo % IVIG logOR	0.4 0.3 0.57	2.0 1.2 0.57	7.3 4.3 0.57	8.1 5.2 0.57	32.6 25.6 0.57	49.5 63.5	Power (%) Avg. logOR	73.8 0.57

Under factor 1, the treatment effect deviates from proportional odds. Under factor 2, the placebo group distribution deviates from that specified in the FLU-IVIG design. All scenarios assume no misclassification and that the full 6-level ordinal outcome is used.

The placebo group distribution specified in the FLU-IVIG design has been updated from the FLU-IVIG protocol using data from the cohort study. For reference, the category percentages in the FLU-IVIG protocol are 1.8, 3.6, 15.6, 14.1, 39.0, and 25.8% for Death through Discharged, back to normal activities categories, respectively.

[a]% Placebo: percentage of patients in the placebo group for the given ordinal endpoint category.

[b]% IVIG: percentage of patients in the IVIG group for the given ordinal endpoint category.

[c]logOR: (natural) logarithm of the odds ratio of the given ordinal endpoint category or more severe versus less severe between the IVIG and placebo groups.

[d]Power (%): percentage of the 10,000 simulated datasets in which the Wald test statistic for the treatment effect was significant at the two-sided 0.05 level.

[e]Avg. logOR: average of the estimated log odds ratio across the 10,000 simulated datasets from fitting a proportional odds cumulative logistic model.

Table 2. Main effect of misclassification on power (factor 3).

Scenario		Death	In ICU	Hospitalized, not in ICU, on oxygen	Hospitalized, not in ICU, not on oxygen	Discharged, not back to normal activities	Discharged, back to normal activities		
M1: 20% misclassification between the non-ICU hospitalized categories and discharged categories	% Placebo[a]	1.2	5.3	15.8	14.8	34.4	28.5	Power (%)[d]	69.7
	% IVIG[b]	0.7	3.1	10.6	10.7	36.6	38.4		
	logOR[c]	0.57	0.57	0.54	0.57	0.45		Avg. logOR[e]	0.50
M2: 40% misclassification between the non-ICU hospitalized categories and discharged categories	% Placebo	1.2	5.3	15.5	15.1	32.4	30.5	Power (%)	57.7
	% IVIG	0.7	3.1	10.6	10.7	37.2	37.8		
	logOR	0.57	0.57	0.52	0.57	0.33		Avg. logOR	0.44
M3: 20% misclassification between the non-ICU hospitalized categories	% Placebo	1.2	5.3	15.8	14.8	36.4	26.5	Power (%)	79.6
	% IVIG	0.7	3.1	10.6	10.7	36.0	39.0		
	logOR	0.57	0.57	0.54	0.57	0.57		Avg. logOR	0.57
M4: 20% misclassification between the discharged categories	% Placebo	1.2	5.3	16.2	14.4	34.4	28.5	Power (%)	70.1
	% IVIG	0.7	3.1	10.5	10.8	36.6	38.4		
	logOR	0.57	0.57	0.57	0.57	0.45		Avg. logOR	0.51

Under factor 3, patients are misclassified between the non-ICU hospitalized categories and the discharged categories assuming the treatment effect (without misclassification) follows proportional odds, the distribution of the placebo group is as specified in the FLU-IVIG design, and the full 6-level ordinal endpoint is used.

[a]% Placebo: percentage of patients in the placebo group for the given ordinal endpoint category.

[b]% IVIG: percentage of patients in the IVIG group for the given ordinal endpoint category.

[c]logOR: (natural) logarithm of the odds ratio of the given ordinal endpoint category or more severe versus less severe between the IVIG and placebo groups.

[d]Power (%): percentage of the 10,000 simulated datasets in which the Wald test statistic for the treatment effect was significant at the two-sided 0.05 level.

[e]Avg. logOR: average of the estimated log odds ratio across the 10,000 simulated datasets from fitting a proportional odds cumulative logistic model.

Table 3. Main effect of the number of categories on power (factor 4).

Scenario		Death	In ICU	Hospitalized, not in ICU, on oxygen	Hospitalized, not in ICU, not on oxygen	Discharged, not back to normal activities	Discharged, back to normal activities		
C1: Collapse the non-ICU hospitalized categories and discharged categories	% Placebo[a] % IVIG[b] logOR[c]	1.2 0.7 0.57	5.3 3.1 0.57	30.6 21.2 0.57		62.9 75.0		Power (%)[d] Avg. logOR[e]	65.1 0.57
C2: Collapse the non-ICU hospitalized categories	% Placebo % IVIG logOR	1.2 0.7 0.57	5.3 3.1 0.57	30.6 21.2 0.57		36.4 36.0 0.57	26.5 39.0	Power (%) Avg. logOR	79.7 0.57
C3: Collapse the discharged categories	% Placebo % IVIG logOR	1.2 0.7 0.57	5.3 3.1 0.57	16.2 10.5 0.57	14.4 10.8 0.57	62.9 75.0		Power (%) Avg. logOR	65.6 0.57
C4: Collapse the four most severe categories	% Placebo % IVIG logOR	37.1 25.0 0.57				36.4 36.0 0.57	26.5 39.0	Power (%) Avg. logOR	78.8 0.57
C5: Collapse the four most severe categories and discharged categories (binary endpoint)	% Placebo % IVIG logOR	37.1 25.0 0.57				62.9 75.0		Power (%) Avg. logOR	63.9 0.57

Under factor 4, categories of the ordinal endpoint are collapsed assuming that the treatment effect follows proportional odds, the distribution of the placebo group is as specified in the FLU-IVIG design, and no misclassification.

[a]% Placebo: percentage of patients in the placebo group for the given ordinal endpoint category.

[b]% IVIG: percentage of patients in the IVIG group for the given ordinal endpoint category.

[c]logOR: (natural) logarithm of the odds ratio of the given ordinal endpoint category or more severe versus less severe between the IVIG and placebo groups.

[d]Power (%): percentage of the 10,000 simulated datasets in which the Wald test statistic for the treatment effect was significant at the two-sided 0.05 level.

[e]Avg. logOR: average of the estimated log odds ratio across the 10,000 simulated datasets from fitting a proportional odds cumulative logistic model.

Table 4. Effect of interacting the treatment effect (factor 1) and placebo group distribution (factor 2) on power.

Scenario		Death	In ICU	Hospitalized, not in ICU, on oxygen	Hospitalized, not in ICU, not on oxygen	Discharged, not back to normal activities	Discharged, back to normal activities		
T1: Treatment effect constantly weakens P2: Even less skewed placebo group distribution	% Placebo[a] % IVIG[b] logOR[c]	3.2 0.2 2.60	12.7 2.3 2.00	28.5 14.0 1.40	17.1 25.4 0.80	26.7 44.1 0.20	11.7 13.9	Power (%)[d] Avg. logOR[e]	99.6 0.94
T2: Treatment effect limited to the hospitalization or death categories P2: Even less skewed placebo group distribution	% Placebo % IVIG logOR	3.2 1.0 1.16	12.7 4.6 1.16	28.5 14.4 1.16	17.1 13.4 1.16	26.7 54.9 0	11.7 11.7	Power (%) Avg. logOR	99.7 0.96
T3: Treatment effect limited to the discharged categories P2: Even less skewed placebo group distribution	% Placebo % IVIG logOR	3.2 3.2 0	12.7 12.7 0	28.5 28.5 0	17.1 17.1 0	26.7 8.6 1.16	11.7 29.8	Power (%) Avg. logOR	19.6 0.22
T1: Treatment effect constantly weakens P4: Even more skewed placebo group distribution	% Placebo % IVIG logOR	0.4 0.0 2.60	2.0 0.3 2.00	7.3 2.3 1.40	8.1 6.3 0.80	32.6 36.7 0.20	49.5 54.4	Power (%) Avg. logOR	33.1 0.33
T2: Treatment effect limited to the hospitalization or death categories P4: Even more skewed placebo group distribution	% Placebo % IVIG logOR	0.4 0.1 1.16	2.0 0.7 1.16	7.3 2.5 1.16	8.1 3.1 1.16	32.6 44.1 0	49.5 49.5	Power (%) Avg. logOR	16.8 0.21
T3: Treatment effect limited to the discharged categories P4: Even more skewed placebo group distribution	% Placebo % IVIG logOR	0.4 0.4 0	2.0 2.0 0	7.3 7.3 0	8.1 8.1 0	32.6 6.3 1.16	49.5 75.9	Power (%) Avg. logOR	97.4 0.91

Under factors 1 and 2, the treatment effect deviates from proportional odds and the distribution of the placebo group deviates from that specified in the FLU-IVIG design assuming no misclassification and the full 6-level ordinal endpoint is used. The treatment effects were paired with placebo group distributions that had more or fewer patients in categories affected by the treatment.

[a][a]% Placebo: percentage of patients in the placebo group for the given ordinal endpoint category.

[b][b]% IVIG: percentage of patients in the IVIG group for the given ordinal endpoint category.

[c]logOR: (natural) logarithm of the odds ratio of the given ordinal endpoint category or more severe versus less severe between the IVIG and placebo groups.

[d]Power (%): percentage of the 10,000 simulated datasets in which the Wald test statistic for the treatment effect was significant at the two-sided 0.05 level.

[e]Avg. logOR: average of the estimated log odds ratio across the 10,000 simulated datasets from fitting a proportional odds cumulative logistic model.

Table 5. Effect of interacting the treatment effect (factor 1) and misclassification (factor 3) on power.

Scenario		Death	In ICU	Hospitalized, not in ICU, on oxygen	Hospitalized, not in ICU, not on oxygen	Discharged, not back to normal activities	Discharged, back to normal activities		
T1: Treatment effect constantly weakens M1: 20% misclassification between the non-ICU hospitalized categories and discharged categories	% Placebo[a] % IVIG[b] logOR[c]	1.2 0.1 2.60	5.3 0.8 2.00	16.2 7.5 1.14	14.4 12.6 0.80	36.4 44.9 0.26	26.5 34.1	Power (%)[d] Avg. logOR[e]	79.1 0.57
T1: Treatment effect constantly weakens M2: 40% misclassification between the non-ICU hospitalized categories and discharged categories	% Placebo % IVIG logOR	1.2 0.1 2.60	5.3 0.8 2.00	15.5 9.2 0.92	15.1 10.9 0.80	32.4 41.3 0.32	30.5 37.7	Power (%) Avg. logOR	79.6 0.57
T2: Treatment effect limited to the hospitalization or death categories M1: 20% misclassification between the non-ICU hospitalized categories and discharged categories	% Placebo % IVIG logOR	1.2 0.4 1.16	5.3 1.7 1.16	16.2 6.5 1.12	14.4 7.0 1.16	36.4 51.7 0.20	26.5 32.8	Power (%) Avg. logOR	86.2 0.64
T2: Treatment effect limited to the hospitalization or death categories M2: 40% misclassification between the non-ICU hospitalized categories and discharged categories	% Placebo % IVIG logOR	1.2 0.4 1.16	5.3 1.7 1.16	15.5 6.6 1.08	15.1 6.8 1.16	32.4 45.4 0.38	30.5 39.1	Power (%) Avg. logOR	92.7 0.71
T3: Treatment effect limited to the discharged categories M1: 20% misclassification between the non-ICU hospitalized categories and discharged categories	% Placebo % IVIG logOR	1.2 1.2 0	5.3 5.3 0	15.8 15.8 0	14.8 14.8 0	36.4 18.2 0.71	26.5 44.7	Power (%) Avg. logOR	38.0 0.33
T3: Treatment effect limited to the discharged categories M2: 40% misclassification	% Placebo % IVIG logOR	1.2 1.2 0	5.3 5.3 0	15.5 15.5 0	15.1 15.1 0	32.4 27.0 0.24	30.5 35.9	Power (%) Avg. logOR	7.5 0.24

between the non-ICU hospitalized categories and discharged categories									

Under factors 1 and 3, the treatment effect deviates from proportional odds and patients are misclassified between the non-ICU hospitalized categories and the discharged categories. We assume the distribution of the placebo group is as specified in the FLU-IVIG design and the full 6-level ordinal endpoint is used.

[a]% Placebo: percentage of patients in the placebo group for the given ordinal endpoint category.

[b]% IVIG: percentage of patients in the IVIG group for the given ordinal endpoint category.

[c]logOR: (natural) logarithm of the odds ratio of the given ordinal endpoint category or more severe versus less severe between the IVIG and placebo groups.

[d]Power (%): percentage of the 10,000 simulated datasets in which the Wald test statistic for the treatment effect was significant at the two-sided 0.05 level.

[e]Avg. logOR: average of the estimated log odds ratio across the 10,000 simulated datasets from fitting a proportional odds cumulative logistic model.

Table 6. Effect of interacting the treatment effect (factor 1) and number of categories (factor 4) on power.

Scenario		Death	In ICU	Hospitalized, not in ICU, on oxygen	Hospitalized, not in ICU, not on oxygen	Discharged, not back to normal activities	Discharged, back to normal activities		
T1: Treatment effect constantly weakens C3: Collapse the discharged categories	% Placebo[a] % IVIG[b] logOR[c]	1.2 0.1 2.60	5.3 0.8 2.00	16.2 5.8 1.40	14.4 14.2 0.80	62.9 79.0		Power (%)[d] Avg. logOR[e]	95.8 0.89
T1: Treatment effect constantly weakens C4: Collapse the four most severe categories	% Placebo % IVIG logOR	37.1 21.0 0.80				36.4 48.5 0.20	26.5 30.5	Power (%) Avg. logOR	67.0 0.50
T2: Treatment effect limited to the hospitalization or death categories C3: Collapse the discharged categories	% Placebo % IVIG logOR	1.2 0.4 1.16	5.3 1.7 1.16	16.2 6.3 1.16	14.4 7.2 1.16	62.9 84.4		Power (%) Avg. logOR	99.5 1.16
T2: Treatment effect limited to the hospitalization or death categories C4: Collapse the four most severe categories	% Placebo % IVIG logOR	37.1 15.6 1.16				36.4 57.9 0	26.5 26.5	Power (%) Avg. logOR	76.0 0.56
T3: Treatment effect limited to the discharged categories C4: Collapse the four most severe categories	% Placebo % IVIG logOR	37.1 37.1 0				36.4 9.3 1.16	26.5 53.6	Power (%) Avg. logOR	80.5 0.60
T4: Smaller treatment effect limited to the hospitalization or death categories C3: Collapse the discharged categories	% Placebo % IVIG logOR	1.2 0.7 0.57	5.3 3.1 0.57	16.2 10.5 0.57	14.4 10.8 0.57	62.9 75.0		Power (%) Avg. logOR	65.6 0.57
T4: Smaller treatment effect limited to the hospitalization or death categories C4: Collapse the four most severe categories	% Placebo % IVIG logOR	37.1 25.0 0.57				36.4 48.5 0	26.5 26.5	Power (%) Avg. logOR	32.0 0.31
T5: Smaller treatment effect limited to the discharged categories	% Placebo % IVIG logOR	37.1 37.1 0				36.4 23.9 0.57	26.5 39.0	Power (%) Avg. logOR	25.7 0.27

C4: Collapse the four most severe categories						

Under factors 1 and 4, the treatment effect deviates from proportional odds and categories of the ordinal endpoint are collapsed assuming the distribution of the placebo group is as specified in the FLU-IVIG design and no misclassification. Categories were collapsed according to whether or not they contained the treatment effect.

[a]% Placebo: percentage of patients in the placebo group for the given ordinal endpoint category.

[b]% IVIG: percentage of patients in the IVIG group for the given ordinal endpoint category.

[c]logOR: (natural) logarithm of the odds ratio of the given ordinal endpoint category or more severe versus less severe between the IVIG and placebo groups.

[d]Power (%): percentage of the 10,000 simulated datasets in which the Wald test statistic for the treatment effect was significant at the two-sided 0.05 level.

[e]Avg. logOR: average of the estimated log odds ratio across the 10,000 simulated datasets from fitting a proportional odds cumulative logistic model.

References

1. World Health Organization. Influenza, http://www.who.int/topics/influenza/en (2016, accessed 30 June 2016).
2. Food and Drug Administation. Guidance for industry influenza: developing drugs for treatment and/or prophylaxis. Report, Silver Spring, MD, April 2011.
3. Ison MG, Fraiz J, Heller B, et al. Intravenous peramivir for treatment of influenza in hospitalized patients. *Antivir Ther* 2014; 19: 349–361.
4. Ison MG, Hui DS, Clezy K, et al. A clinical trial of intravenous peramivir compared with oral oseltamivir for the treatment of seasonal influenza in hospitalized adults. *Antivir Ther* 2013; 18: 651–661.
5. Ison MG, De Jong MD, Gilligan KJ, et al. End points for testing influenza antiviral treatments for patients at high risk of severe and life-threatening disease. *J Infect Dis* 2010; 201: 1654–1662.
6. De Jong MD, Ison MG, Monto AS, et al. Evaluation of intravenous peramivir for treatment of influenza in hospitalized patients. *Clin Infect Dis* 2014; 59: 172–185.
7. South East Asia Infectious Disease Clinical Research Network. Effect of double dose oseltamivir on clinical and virological outcomes in children and adults admitted to hospital with severe influenza: double blind randomised controlled trial. *BMJ* 2013; 346: 1–16.
8. INSIGHT FLU005 IVIG Pilot Study Group. INSIGHT FLU005: An anti-influenza virus hyperimmune intravenous immunoglobulin pilot study. *J Infect Dis* 2016; 213: 574–578.
9. National Institute of Allergy and Infectious Diseases. Evaluating the safety and efficacy of anti-influenza intravenous hyperimmune immunoglobulin (IVIG) in adults hospitalized with influenza, https://clinicaltrials.gov/ct2/show/NCT02287467 (2016, accessed 30 June 2016).
10. Royal College of Physicians of London. National early warning score (NEWS), https://www.rcplondon.ac.uk/projects/outputs/national-early-warning-score-news (2015, accessed 30 June 2016).
11. Lynfield R, Davey R, Dwyer DE, et al. Outcomes of influenza A(H1N1)pdm09 virus infection: results from two international cohort studies. *PLoS One* 2014; 9: 1–15.
12. Dwyer DE. Surveillance of illness associated with pandemic (H1N1) 2009 virus infection among adults using a global clinical site network approach: The INSIGHT FLU 002 and FLU 003 studies. *Vaccine* 2011; 29: 56–62.
13. McCullagh P. Regression models for ordinal data. *J R Stat Soc B* 1980; 42: 109–142.
14. Whitehead J. Sample size calculations for ordered categorical data. *Stat Med* 1993; 12: 2257–2271.
15. Jeong JH. A note on asymptotic efficiency of a regression coefficient parameter under ordinal logistic regression model. *Commun Stat - Theory Methods* 2001; 30: 1257–1269.
16. Strömberg U. Collapsing ordered outcome categories: a note of concern. *Am J Epidemiol* 1996; 144: 421–424.
17. Powers JH, Guerrero ML, Leidy NK, et al. Development of the Flu-PRO: a patient-reported outcome (PRO) instrument to evaluate symptoms of influenza.

BMC Infect Dis 2015; 16: 1–11.
18. Roozenbeek B, Lingsma HF, Perel P, et al. The added value of ordinal analysis in clinical trials: an example in traumatic brain injury. *Crit Care* 2011; 15: 1–7.
19. Pédrono G, Thiébaut R, Alioum A, et al. A new endpoint definition improved clinical relevance and statistical power in a vaccine trial. *J Clin Epidemiol* 2009; 62: 1054–1061.
20. Maas AIR, Steyerberg EW, Marmarou A, et al. IMPACT recommendations for improving the design and analysis of clinical trials in moderate to severe traumatic brain injury. *Neurotherapeutics* 2010; 7: 127–134.
21. Bath PMW, Geeganage C, Gray LJ, et al. Use of ordinal endpoints in vascular prevention trials: comparison with binary outcomes in published trials. *Stroke* 2008; 39: 2817–2823.
22. Kromrey JD, Hogarty KY. Analysis options for testing group differences on ordered categorical variables: An empirical investigation of type I error control and statistical power. *Mult Linear Regres Viewpoints* 1998; 25: 70–82.
23. McHugh GS, Butcher I, Steyerberg EW, et al. A simulation study evaluating approaches to the analysis of ordinal endpoint data in randomized controlled trials in traumatic brain injury: results from the IMPACT Project. *Clin Trials* 2010; 7: 44–57.
24. Pocock SJ, Ariti CA, Collier TJ, et al. The win ratio: A new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. *Eur Heart J* 2012; 33: 176–182.
25. Boos DD, Stefanski LA. *Essential statistical inference*. New York, NY: Springer, 2013.

Appendix

Supplemental Table 1. Power to detect a significant treatment effect.

logOR[a]	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Power (%)[b]	2.5	7.2	16.6	31.8	50.8	69.7	84.4	93.4	97.7	99.4	99.9

Power is computed at the 0.05 (2-sided) level with a sample size of 320 as a function of the log odds ratio assuming the treatment effect follows proportion odds, the distribution of the placebo group is as specified in the FLU-IVIG design, no misclassification, and the full 6-level ordinal endpoint is used.

[a]logOR: average of the estimated log odds ratio across the 10,000 simulated datasets from fitting a proportional odds cumulative logistic model assuming that proportional odds holds.

[b]Power (%): percentage of the 10,000 simulated datasets in which the Wald test statistic for the treatment effect was significant at the two-sided 0.05 level.

Supplemental Table 2. Power for detecting a significant treatment effect for all possible ways of dividing the ordinal endpoint into a binary endpoint.

Category	Death versus In ICU or better	In ICU or worse versus Hospitalized, not in ICU, on oxygen	Hospitalized, not in ICU, on oxygen or worse versus Hospitalized, not in ICU, not on oxygen or better	Hospitalized, not in ICU, not on oxygen or worse versus Discharged, not back to normal activities or better	Discharged, not back to normal activities or worse versus Discharged, back to normal activities
Power (%)[a]	0.59	12.8	48.9	63.9	66.2

Power is computed at the 0.05 (2-sided) level with a log odds ratio of 0.57 and sample size of 320 assuming the treatment effect follows proportion odds, the distribution of the placebo group is as specified in the FLU-IVIG design, no misclassification, and the full 6-level ordinal endpoint is used.

[a]Power (%): percentage of the 10,000 simulated datasets in which the Wald test statistic for the treatment effect was significant at the two-sided 0.05 level.

Supplemental Table 3. Altering the distribution of the placebo group.

Placebo Group		Death	In ICU	Hospitalized, not in ICU, on oxygen	Hospitalized, not in ICU, not on oxygen	Discharged, not back to normal activities	Discharged, back to normal activities
P0: The placebo group from the FLU-IVIG design	% Placebo[a] Cumulative log odds[b]	1.2 -4.41	5.3 -2.67	16.2 -1.23	14.4 -0.53	36.4 1.02	26.5
P1: Less skewed placebo group distribution	% Placebo Cumulative log odds	2.0 -3.91	8.3 -2.17	22.4 -0.73	16.6 -0.03	32.7 1.52	18.0
P2: Even less skewed placebo group distribution	% Placebo Cumulative log odds	3.2 -3.41	12.7 -1.67	28.5 -0.23	17.1 0.47	26.7 2.02	11.7
P3: More skewed placebo group distribution	% Placebo Cumulative log odds	0.7 -4.91	3.3 -3.17	11.1 -1.73	11.2 -1.03	36.3 0.52	37.3
P4: Even more skewed placebo group distribution	% Placebo Cumulative log odds	0.4 -5.41	2.0 -3.67	7.3 -2.23	8.1 -1.53	32.6 0.02	49.5

Altering the distribution of the placebo group specified in the FLU-IVIG design to be more or less skewed by changing its cumulative log odds. The derivation of the cumulative log odds to the probabilities in each category is given in the section below.
[a]% Placebo: percentage of patients in the placebo group for the given ordinal endpoint category.
[b]Cumulative log odds: (natural) logarithm of the odds of the given category or more severe versus less severe.

Derivation of the cumulative log odds for the distribution of the placebo group

For an ordinal endpoint Y, the cumulative log odds C_j for level j out of J total levels in the distribution of the placebo group is:

$$\log\left(\frac{P(Y \leq j)}{P(Y > j)}\right) = C_j \; for \; j = 1,2, \ldots J-1$$

To return to the cumulative probability, that is, $P(Y \leq j)$, use the expit function on C_j:

$$\frac{e^{C_j}}{1 + e^{C_j}} * 100 = P(Y \leq j)$$

The probability that Y assumes level j, that is $P(Y = j)$, is then derived as:

$$P(Y = j) = P(Y \leq j) - P(Y \leq j - 1)$$

A demonstration using patients in Death or ICU for P0:

$$\log\left(\frac{1.2 + 5.3}{100 - (1.2 + 5.3)}\right) = -2.67$$

$$\frac{e^{-2.67}}{1 + e^{-2.67}} * 100 = 6.5 = 1.2 + 5.3$$

Supplemental Table 4. Effect of all possible interactions of the treatment effects and placebo group distributions on power.

		T1: Treatment effect constantly weakens	T2: Treatment effect limited to the hospitalization or death categories	T3: Treatment effect limited to the discharged categories	T4: Smaller treatment effect limited to the hospitalization or death categories	T5: Smaller treatment effect limited to the discharged categories
P1: Less skewed placebo group distribution	Power (%)[a] Avg. logOR[b]	95.8 0.76	96.7 79.2	47.2 0.38	52.8 0.42	12.4 0.16
P2: Even less skewed placebo group distribution	Power (%) Avg. logOR	**99.6** **0.94**	**99.7** **0.96**	**19.6** **0.22**	67.6 0.48	6.5 0.09
P3: More skewed placebo group distribution	Power (%) Avg. logOR	54.5 0.43	40.8 0.37	93.9 0.77	16.8 0.21	39.9 0.35
P4: Even more skewed placebo group distribution	Power (%) Avg. logOR	**33.1** **0.33**	**16.8** **0.21**	**97.4** **0.91**	8.4 0.13	51.2 0.44

Bolded interactions were chosen for mention in the text of the paper.

[a]Power (%): percentage of the 10,000 simulated datasets in which the Wald test statistic for the treatment effect was significant at the two-sided 0.05 level.

[b]Avg. logOR: average of the estimated log odds ratio across the 10,000 simulated datasets from fitting a proportional odds cumulative logistic model.

Supplemental Table 5. Effect of all possible interactions of the treatment effects and misclassification on power.

		T1: Treatment effect constantly weakens	T2: Treatment effect limited to the hospitalization or death categories	T3: Treatment effect limited to the discharged categories	T4: Smaller treatment effect limited to the hospitalization or death categories	T5: Smaller treatment effect limited to the discharged categories
M1: 20% misclassification between the non-ICU hospitalized categories and discharged categories	Power (%)[a] Avg. logOR[b]	**79.1** **0.57**	**86.2** **0.64**	**38.0** **0.33**	40.4 0.35	11.3 0.15
M2: 40% misclassification between the non-ICU hospitalized categories and discharged categories	Power (%) Avg. logOR	**79.6** **0.57**	**92.7** **0.71**	**7.5** **0.24**	48.0 0.39	4.4 0.49
M3: 20% misclassification between the non-ICU hospitalized categories	Power (%) Avg. logOR	76.2 0.56	76.9 0.57	79.1 0.58	33.1 0.31	23.6 0.25
M4: 20% misclassification between the discharged categories	Power (%) Avg. logOR	82.4 0.60	86.8 0.64	38.0 0.34	40.9 0.36	11.2 0.15

Bolded interactions were chosen for mention in the text of the paper.

[a]Power (%): percentage of the 10,000 simulated datasets in which the Wald test statistic for the treatment effect was significant at the two-sided 0.05 level.

[b]Avg. logOR: average of the estimated log odds ratio across the 10,000 simulated datasets from fitting a proportional odds cumulative logistic model.

Supplemental Table 6. Effect of all possible interactions of the treatment effects and number of categories on power.

		T1: Treatment effect constantly weakens	T2: Treatment effect limited to the hospitalization or death categories	T3: Treatment effect limited to the discharged categories	T4: Smaller treatment effect limited to the hospitalization or death categories	T5: Smaller treatment effect limited to the discharged categories
C1: Collapse the non-ICU hospitalized categories and discharged categories	Power (%)[a] Avg. logOR[b]	92.3 0.85	99.4 1.18	0.05 0	65.4 0.57	0.05 0
C2: Collapse the non-ICU hospitalized categories	Power (%) Avg. logOR	72.4 0.53	76.4 0.57	80.0 0.59	32.4 0.31	24.2 0.26
C3: Collapse the discharged categories	Power (%) Avg. logOR	**95.8** **0.89**	**99.5** **1.16**	0.05 0	**65.6** **0.57**	0.05 0
C4: Collapse the four most severe categories	Power (%) Avg. logOR	**67.0** **0.50**	**76.0** **0.56**	**80.5** **0.60**	**32.0** **0.31**	**25.7** **0.27**
C5: Collapse the four most severe categories and discharged categories (binary endpoint)	Power (%) Avg. logOR	88.9 0.81	99.4 1.18	0.05 0	64.5 0.57	0.05 0

Bolded interactions were chosen for mention in the text of the paper.

[a]Power (%): percentage of the 10,000 simulated datasets in which the Wald test statistic for the treatment effect was significant at the two-sided 0.05 level.

[b]Avg. logOR: average of the estimated log odds ratio across the 10,000 simulated datasets from fitting a proportional odds cumulative logistic model.

Deviations from proportional odds while maintaining the same overall treatment effect

For the *i*th patient, assume that we have a 3-level ordinal endpoint Y_i and define:

- p_1, p_2, p_3 are the true probabilities in the first, second, and third levels of the ordinal endpoint for the placebo group, respectively.
- q_1, q_2, q_3 are the corresponding true probabilities in the treatment group.
- A_i is an indicator variable for whether or not the *i*th patient is randomized to the treatment group.
- Z_i, V_i are indicator variables for whether or not $Y_i = 1$ and $Y_i = 2$, respectively, for the *i*th patient.

Note that $p_3 = 1 - p_1 - p_2$ and $q_3 = 1 - q_1 - q_2$; therefore, the distribution of the ordinal endpoint in the placebo and treatment groups is uniquely determined by the four parameters p_1, p_2, q_1, and q_2. If we assume a proportional odds model, we can express p_1, p_2, q_1, and q_2 in terms of three model parameters. Let α_1 and α_2 represent the log odds of being in level 1 and in level 1 or level 2, respectively, for subjects randomized to the placebo group and let β represent the log odds ratio of the treatment group to the control group. Assuming proportional odds, $\alpha_1 = \log\left(\frac{p_1}{1-p_1}\right)$, $\alpha_2 = \log\left(\frac{p_1+p_2}{1-p_1-p_2}\right)$, and $\beta = \log\left(\frac{q_1*(1-p_1)}{p_1*(1-q_1)}\right) = \log\left(\frac{(q_1+q_2)*(1-p_1-p_2)}{(p_1+p_2)*(1-q_1-q_2)}\right)$.

Under this model, the log likelihood for $\alpha_1, \alpha_2, \beta$ is given by:

$$\log(L(\alpha_1, \alpha_2, \beta))$$

$$= \sum_{i=1}^{n}\{(1 - A_i)[Z_i \log\left(\frac{e^{\alpha_1}}{e^{\alpha_1}+1}\right) + V_i \log\left(\frac{e^{\alpha_2}-e^{\alpha_1}}{(e^{\alpha_1}+1)(e^{\alpha_2}+1)}\right)$$

$$+(1 - Z_i - V_i)\log\left(\frac{e^{\alpha_1}+1}{(e^{\alpha_1}+1)(e^{\alpha_2}+1)}\right)]\}$$

$$+ \sum_{i=1}^{n}\{A_i[Z_i \log\left(\frac{e^{\alpha_1+\beta}}{e^{\alpha_1+\beta}+1}\right) + V_i \log\left(\frac{e^{\alpha_2}-e^{\alpha_1}}{(e^{\alpha_1+\beta}+1)(e^{\alpha_2}+e^{-\beta})}\right)$$

$$+ (1 - Z_i - V_i)\log\left(\frac{e^{\alpha_1}+e^{-\beta}}{(e^{\alpha_1+\beta}+1)(e^{\alpha_2}+e^{-\beta})}\right)]\}$$

Regardless of whether or not the proportional odds model is correctly specified, we can obtain maximum likelihood estimates for α_1, α_2, and β (i.e., $\widehat{\alpha_1}$, $\widehat{\alpha_2}$, $\hat{\beta}$, the values of α_1, α_2, and β which maximize the log likelihood above). If the proportional odds assumption is not correct, $\hat{\beta}$ is still an estimate of the treatment effect across all levels of the ordinal endpoint but cannot be interpreted as the constant log odds ratio of the treatment group to the placebo group across all binary divisions of the ordinal scale.

As noted in the main text, we sought to derive distributions of the treatment group that deviated from proportional odds while maintaining the same overall treatment effect specified in the design of FLU-IVIG. By overall treatment effect, we mean the average (across repeated experimentation) estimated log odds ratio for the effect of the intervention relative to placebo from fitting a proportional odds cumulative logistic regression model to the data (i.e., $E(\hat{\beta})$).

Let $\alpha_{10} = E(\widehat{\alpha_1}),\ \ \alpha_{20} = E(\widehat{\alpha_2})$, and $\beta_0 = E(\hat{\beta})$ represent the average estimated cumulative log odds and log odds ratio. Asymptotically, α_{10}, α_{20}, and β_0 are the values of α_1, α_2, and β for which the expected score is equal to zero.[25] That is for a fixed sample size, α_{10}, α_{20}, and β_0 are (approximately) the values which solve the following system of equations (1) $E\left[\frac{d}{d\alpha_1}\log(L(\alpha_1,\alpha_2,\beta))\right] = 0$; (2) $E[\frac{d}{d\alpha_2}\log(L(\alpha_1,\alpha_2,\beta))] = 0$; and (3) $E[\frac{d}{d\beta}\log(L(\alpha_1,\alpha_2,\beta))] = 0$. Note that:

$$E[\frac{d}{d\beta}\log(L(\alpha_1,\alpha_2,\beta))]$$

$$= E\{A_i[Z_i\frac{e^{\alpha_1+\alpha_2+2\beta}+2e^{\alpha_2+\beta}+1}{(e^{\alpha_1+\beta}+1)(e^{\alpha_2+\beta}+1)} + V_i\frac{1}{e^{\alpha_1+\beta}+1} - \frac{e^{\alpha_2+\beta}}{e^{\alpha_2+\beta}+1}]\}$$

$$= E[E\{A_i[Z_i\frac{e^{\alpha_1+\alpha_2+2\beta}+2e^{\alpha_2+\beta}+1}{(e^{\alpha_1+\beta}+1)(e^{\alpha_2+\beta}+1)} + V_i\frac{1}{e^{\alpha_1+\beta}+1} - \frac{e^{\alpha_2+\beta}}{e^{\alpha_2+\beta}+1}]\}\big|A_i]$$

$$= E\{A_i[E(Z_i|A_i)\frac{e^{\alpha_1+\alpha_2+2\beta}+2e^{\alpha_2+\beta}+1}{(e^{\alpha_1+\beta}+1)(e^{\alpha_2+\beta}+1)} + E(V_i|A_i)\frac{1}{e^{\alpha_1+\beta}+1} - \frac{e^{\alpha_2+\beta}}{e^{\alpha_2+\beta}+1}]\}$$

Note that $E(Z_i|A_i) = A_iq_1 + (1-A_i)p_1$ and $E(V_i|A_i) = A_iq_2 + (1-A_i)p_2$. We then have:

$$E[\frac{d}{d\beta}\log(L(\alpha_1,\alpha_2,\beta))]$$

$$= E\{A_iq_1\frac{e^{\alpha_1+\alpha_2+2\beta}+2e^{\alpha_2+\beta}+1}{(e^{\alpha_1+\beta}+1)(e^{\alpha_2+\beta}+1)} + A_iq_2\frac{1}{e^{\alpha_1+\beta}+1} - A_i\frac{e^{\alpha_2+\beta}}{e^{\alpha_2+\beta}+1}\}$$

Because A_i is the only random variable in the above equation with $E(A_i) = 0.5$ (due to the 1:1 allocation ratio between the randomized groups), we have:

$$= q_1\frac{e^{\alpha_1+\alpha_2+2\beta}+2e^{\alpha_2+\beta}+1}{(e^{\alpha_1+\beta}+1)(e^{\alpha_2+\beta}+1)} + q_2\frac{1}{e^{\alpha_1+\beta}+1} - \frac{e^{\alpha_2+\beta}}{e^{\alpha_2+\beta}+1} = 0$$

A similar analysis can be used to simplify $E[\frac{d}{d\alpha_1}\log(L(\alpha_1,\alpha_2,\beta))]]$ and $E[\frac{d}{d\alpha_2}\log(L(\alpha_1,\alpha_2,\beta))]$ which will be functions of α_1, α_2, β, and the true probabilities in each level of the treatment (q_1 and q_2) and control group (p_1 and p_2).

To derive distributions of the treatment group that deviated from proportional odds while maintaining the same overall treatment effect, we can fix β (the overall treatment effect), p_1 and p_2 (the probabilities in the first two categories of the control group), and q_1 (the probability in the first category of the treatment group) to solve the system of three (nonlinear) equations for q_2, α_1 and α_2. This approach generalizes to ordinal endpoints with any number of outcome levels.

Code to implement this algorithm in the programming language R is available as a GitHub repository (https://github.com/RPeterson4/Supplementary-Code-for-Evaluating-the-Ordinal-Endpoint-for-FLU-IVIG).

Misclassification among the categories of the ordinal endpoint

Scenario		Death	In ICU	Hospitalized, not in ICU, on oxygen	Hospitalized, not in ICU, not on oxygen	Discharged, not back to normal activities	Discharged, back to normal activities
M0: No Misclassification	% Placebo[a] % IVIG[b]	1.2 0.7	5.3 3.1	16.2 10.5	14.4 10.8	36.4 36.0	26.5 39.0

[a]% Placebo: percentage of patients in the placebo group for the given ordinal endpoint category.
[b]% IVIG: percentage of patients in the IVIG group for the given ordinal endpoint category.

Misclassification is added between the oxygen and discharged categories by exchanging fixed percentages of patients between the respective categories for each pair for both randomized groups. An example for M1, which adds 20% misclassification between the non-ICU hospitalized categories and the discharged categories.

M1 Placebo Hospitalized, not in ICU, on oxygen = 16.2 * 0.8 + 14.4 * 0.2 = 15.8
M1 Placebo Hospitalized, not in ICU, not on oxygen = 16.2 * 0.2 + 14.4 * 0.8 = 14.8
M1 Placebo Discharged, not back to normal activities = 36.4 * 0.8 + 26.5 * 0.2 = 34.4
M1 Placebo Discharged, back to normal activities = 36.4 * 0.2 + 26.5 * 0.8 = 28.5

M1 IVIG Hospitalized, not in ICU, on oxygen = 10.5 * 0.8 + 10.8 * 0.2 = 10.6
M1 IVIG Hospitalized, not in ICU, not on oxygen = 10.5 * 0.2 + 10.8 * 0.8 = 10.7
M1 IVIG Discharged, not back to normal activities = 36.0 * 0.8 + 39.0 * 0.2 = 36.6
M1 IVIG Discharged, back to normal activities = 36.0 * 0.2 + 39.0 * 0.8 = 38.4

This yields the placebo and IVIG group distributions for M1:

Scenario		Death	In ICU	Hospitalized, not in ICU, on oxygen	Hospitalized, not in ICU, not on oxygen	Discharged, not back to normal activities	Discharged, back to normal activities
M1: 20% misclassification between the non-ICU hospitalized categories and discharged categories	% Placebo[a] % IVIG[b]	1.2 0.7	5.3 3.1	15.8 10.6	14.8 10.7	34.4 36.6	28.5 38.4

[a]% Placebo: percentage of patients in the placebo group for the given ordinal endpoint category.
[b]% IVIG: percentage of patients in the IVIG group for the given ordinal endpoint category.