medRxiv preprint doi: https://doi.org/10.1101/2020.04.19.20069922; this version posted April 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Improving Precision and Power in Randomized Trials for COVID-19 Treatments Using Covariate Adjustment, for Ordinal or Time to Event Outcomes

David Benkeser[1,*], Ivan Diaz[2,*], Alex Luedtke[3,4,*],
Jodi Segal[5], Daniel Scharfstein[6], Michael Rosenblum[6,†]

[*] Co-first authors
[†] Corresponding author: mrosen@jhu.edu, 615 N Wolfe St., Room E3616, Baltimore, MD 21205.
[1] Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University
[2] Division of Biostatistics, Department of Population Health Sciences, Weill Cornell Medicine
[3] Department of Statistics, University of Washington
[4] Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center
[5] Department of Medicine, School of Medicine, Johns Hopkins University
[6] Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Johns Hopkins University

April 20, 2020

1 Abstract

We aim to help inform the choice of estimand (i.e., target of inference) and analysis method to be used in future COVID-19 treatment trials. To this end, we describe estimands for outcome types of particular interest in these trials (ordinal and time-to-event). When the outcome is ordinal, the estimands that we consider are the difference between study arms in the mean outcome, the Mann-Whitney (rank–based) estimand, and the average of the cumulative log odds ratios over the levels of the outcome. For time-to-event outcomes, we consider the difference between arms in the restricted mean survival time, the difference between arms in the cumulative incidence, and the relative risk. Advantageously, the interpretability of these estimands does not rely on a proportional odds or proportional hazards assumptions. For each estimand, we evaluate the potential value added by using estimators that leverage information in baseline variables to improve precision and reduce the required sample size to achieve a desired power. These are called covariate adjusted estimators. To evaluate the performance of the covariate adjusted and unadjusted estimators that we present, we simulate two-arm, randomized trials comparing a hypothetical COVID-19 treatment versus standard of care, where the primary outcome is ordinal or time-to-event. Our simulated distributions are derived from two sources: longitudinal data on over 500 patients hospitalized at Weill Cornell Medicine New York Presbyterian Hospital prior to March 28, 2020, and a CDC

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.19.20069922; this version posted April 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

preliminary description of 2449 cases reported to the CDC from February 12 to March 16, 2020. We focus on hospitalized, COVID-19 positive patients and consider the following outcomes: intubation, ventilator use, and death. We conduct simulations using all three estimands when the outcome is ordinal, but only evaluate the restricted mean survival time when the outcome is time to event. Our simulations showed that, in trials with at least 200 participants, the precision gains due to covariate adjustment are equivalent to requiring 10-20% fewer participants to achieve the same power as a trial that uses the unadjusted estimator; this was the case for each outcome and estimand that we considered.

Paper Organization. Section 2 provides background on the use of covariate adjustment in randomized trials. Section 3 describes estimands and estimation strategies that we find compelling when the outcome is ordinal or a time to event. Section 4 describes the methods behind the simulation study that we conducted to evaluate these estimation strategies. Section 5 presents the results from this simulation. Section 6 presents our recommendations for the estimands and primary efficacy analysis methods that we recommend for future COVID-19 treatment trials.

2 Background on Covariate Adjustment in Randomized Trials, for Ordinal and Time to Event Outcomes

The ICH E9 Guidance on Statistical Methods for Analyzing Clinical Trials (1998) states that "Pretrial deliberations should identify those covariates and factors expected to have an important influence on the primary variable(s), and should consider how to account for these in the analysis to improve precision and to compensate for any lack of balance between treatment groups" [10]. Adjusting for prognostic baseline variables is called covariate adjustment.

Though there appears to be general agreement among regulators [8, 9] on how to adjust for baseline variables when the outcome is continuous valued (i.e., using analysis of covariance, abbreviated as ANCOVA), there is less specific guidance for several outcome types of interest in COVID-19 treatment trials, namely binary, ordinal, or time to event outcomes [1]. We evaluate the performance of covariate adjusted estimators for ordinal and time to event outcomes. Binary outcomes are a special case where the ordinal outcome has two levels, and the standardization approach of [13, 17] can be applied.

3 Estimands and Analysis Methods

Throughout we assume that treatment is randomly assigned independently of baseline covariates. The described methods can easily be adapted to handle the case that treatment assignment depends on (a subset of) the measured baseline covariates, as is the case when stratified randomization is used.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.19.20069922; this version posted April 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

3.1 Ordinal Outcomes

We consider three estimands when the outcome is ordinal. We comment on these estimands briefly here, and refer the reader to Appendix A.1 for analytic expressions for these quantities.

Estimand 1: Difference in means of a prespecified transformation of the ordinal outcome (DIM). In most settings, this transformation will be monotone so that larger values of the ordinal outcome (assumed to be preferable) will result in larger, and therefore preferable, transformed outcomes.

Estimand 2: Mann-Whitney (MW) estimand. This estimand reports the probability that a random individual assigned to treatment will have a better outcome than a random individual assigned to control, with ties broken at random.

Estimand 3: Log-odds ratio (LOR). We consider a nonparametric extension of the log-odds ratio (LOR) [6] defined as the average of the cumulative log odds ratios over levels 1 to $K-1$ of the outcome. In the case that the distribution of the outcome given treatment arm is accurately described by a proportional odds model [12], this estimand corresponds to the coefficient in front of treatment in this model. Otherwise, it represents a projection onto this model.

It follows from Appendix A.1 that all of the above estimands can be written as smooth summaries of the arm-specific cumulative distribution functions (CDFs) of an ordinal outcome. To estimate these quantities, it suffices to estimate these two CDFs and then to evaluate these summaries. This results in so-called plug-in estimators.

We consider two CDF estimators. The first, which we term the unadjusted estimator, corresponds to using the arm-specific empirical CDFs, that is, corresponds to stratifying the sample based on treatment arm and then using the observed empirical distributions within each arm. The resulting estimator for the DIM corresponds precisely to taking the empirical differences of transformed mean outcomes between the two arms. The resulting estimator M for the MW estimand is closely related to the usual Mann-Whitney U-statistic U: in particular, $U = n_0 n_1 M$, where n_0 and n_1 are the total sample sizes in the two treatment groups.

The second CDF estimator, which we term the covariate adjusted estimator, leverages prognostic information that is available at baseline. It uses working models, i.e., models that are fit in the process of computing the estimator but which we do not assume to be correctly specified (and for which consistency and asymptotic normality still hold under arbitrary model misspecification). Specifically, this estimator makes use of a treatment-stratified proportional odds working model that assumes that, within each treatment arm, the CDF of the ordinal outcome satisfies the proportional odds assumption based on a linear combination of the baseline covariates. This yields two working covariate-conditional CDFs (one per treatment arm). Our final treatment-specific marginal CDFs are obtained by averaging these conditional CDFs across the pooled covariates from the two treatment arms. Importantly, the validity of this estimator (e.g., asymptotic unbiasedness, asymptotic normality) does not rely on correct specification of the working treatment-stratified proportional odds model

medRxiv preprint doi: https://doi.org/10.1101/2020.04.19.20069922; this version posted April 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

that this estimator uses. The fact that our CDF estimator is robust to misspecification of the outcome working model is reminiscent of the fact that ANCOVA estimator can be used for covariate adjustment when the outcome is continuous [24, 22]. We also refer the reader to the following references for further discussion of the benefits of covariate adjustment: for the DIM, see [22]; for the MW estimand, see [21]; for the LOR, see [6]; in the case that the ordinal outcome takes only two values, see [14], whose general discussion applies to all three estimands.

Because the covariate adjusted approach that we have described is based on a working parametric model, there are many possible approaches to developing valid confidence intervals for the estimand of interest. These approaches include the nonparametric BCa bootstrap [7] or Wald-type confidence intervals. When developing BCa intervals based on trial data, we recommend using a large number of bootstrap replicates to ensure that the final interval does not meaningfully depend on the random number generator.

We recommend handling missing covariates by imputing them based only on data from those covariates that were observed. Importantly, to ensure that treatment assignment can be treated as randomized even conditionally on the imputed covariates, we do not recommend using treatment or outcome information in this imputation. We recommend handling missing ordinal outcomes using doubly robust methods whose validity relies on the outcomes being missing at random conditionally on the covariates and treatment assignment. In terms of implementation, this approach involves simply applying the methods described in Appendix A.2, but with study arm assignment recoded as 0 to indicate that a patient was both randomized to study arm 0 (control) and had their outcome measured, 1 to indicate that a patient was both randomized to study arm 1 (treatment) and had their outcome measured, and -1 to indicate that the outcome is missing.

3.2 Time-to-Event Outcomes

We consider three treatment effect estimands in the time-to-event setting, all of which are interpretable under violations of a proportional hazards assumption. We refer the reader to Appendix B for formal analytic expressions for these quantities.

Estimand 1: Difference in restricted mean survival times (RMSTs). Recall that the RMST corresponds to the expected value of a survival time that is truncated at a specified time [4, 15]. Alternatively, the RMST can be expressed as the area under the survival curve up to the truncation time.

Estimand 2: Cumulative incidence difference (CID). Additive difference in arm-specific incidences of the event by a specified time.

Estimand 3: Relative risk (RR). Ratio of the arm-specific incidences of the event by a specified time

A covariate adjusted targeted minimum loss-based estimator of the RMST is described in Díaz et al. [5]. Analogous to the ordinal outcome case, the estimator of the RMST relies

medRxiv preprint doi: https://doi.org/10.1101/2020.04.19.20069922; this version posted April 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

on an estimate of the survival function at each time point. Estimators obtained through the Kaplan-Meier curve are referred to as unadjusted. Adjusted estimators are obtained by computing time-specific hazard probabilities which are transformed using the product-limit formula and marginalized over covariates. The adjusted approach has two key benefits relative to unadjusted alternatives, for example, those based on an unadjusted Kaplan-Meier estimator [11]. First, under regularity conditions, the adjusted estimator is at least as precise as the unadjusted estimator, asymptotically. Second, the adjusted estimator's consistency depends on an assumption of censoring being independent of the outcome given study arm and baseline covariates, rather than an assumption of censoring in each arm being independent of the outcome marginally; see Eqs. 3 and 4 in Appendix B. The former may be a more plausible assumption. Similar covariate adjusted estimators for the CID and RR are also available [2].

Both [5] and [2] provide approaches that can be used to develop Wald-type confidence intervals and corresponding tests of the null of no effect.

4 Simulation Methods

4.1 Data-Generating Distributions

4.1.1 Ordinal outcomes

We generated data based on [3], which reported outcomes for hospitalized and non-hospitalized individuals with COVID-19. We focus on the hospitalized data and place additional results pertaining to the non-hospitalized population in the Appendix. There were three levels of the ordinal outcome, with level 1 representing death, level 2 representing that ventilation was required, and level 3 representing that ventilation was not required. We adjusted for the following patient age categories: 0–19, 20–44, 45–54, 55–64, 65–74, 75–84, and ≥85. Lower and upper estimates were given for probabilities of outcomes in each age group in [3]; we used the mean of these within each age group.

For the hospitalized COVID-19 positive population, the resulting probabilities of being in each age group and of the outcomes: ventilator use and death are listed in Table 1.

Age	P(age)	P(death | age)	P(ventilation | age)
0–19	0.01	0.00	0.00
20–44	0.09	0.01	0.18
45–54	0.12	0.03	0.32
55–64	0.13	0.08	0.31
65–74	0.18	0.11	0.37
75–84	0.22	0.17	0.47
≥ 85	0.25	0.37	0.35

Table 1: Hospitalized, COVID-19 positive population: Age and conditional outcome distributions used for the control arm in the ordinal outcome simulation study.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.19.20069922; this version posted April 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

We evaluated performance of the methods for both ineffective and effective treatments. We considered trials with 1:1 randomization probabilities and total enrollment of $n = 100$, 200, 500, and 1000. To simulate treatment effects, we assumed that a treatment would have no effect on the probability of death but would decrease the odds of ventilation by the same relative amount in each age category. This relative reduction was selected for each sample size, such that a two-sided t-test that the average outcome between treatment arms were equal rejected the null hypothesis with approximately 80% power (n=200,500,1000) and 50% power (n=100). The discrepancy between $n = 100$ and the other sample sizes occurred because there did not exist a relative reduction that achieved 80% power at a sample size of 100. For each setting, 1000 Monte Carlo replicates were performed.

The adjusted estimators that we implemented in our simulation were based on a proportional odds regression model that included age category as a linear term (0-19, 20-44, 45-54, 55-64, 65-74, 75-84, $\geq$85). We estimated the treatment propensity using the marginal empirical probability of treatment. We used 1000 replicates when developing BCa bootstrap confidence intervals.

4.1.2 Time-to-event outcomes

In order to mimic key predictive features of clinical outcomes in COVID patients, our simulation data generating mechanism is based on the over 500 patients hospitalized at Weill Cornell Medicine New York Presbyterian Hospital prior to March 28, 2020. This ensures we have at least 14 days of follow-up for all individuals. The event of interest in this simulation is time from hospitalization to intubation, and the predictive variables used are sex, age, whether the patient required supplemental oxygen at ED presentation, dyspnea, hypertension, and the presence of bilateral infiltrates in the chest x-ray. We focus on the RMST 14 days after hospitalization. Patient data was re-sampled with replacement to generate 1000 datasets of each of the sizes $n = 100$, 200, 500, and 1000. For each dataset, a hypothetical treatment variable was drawn from a Bernoulli distribution with probability 0.5 independently of all other variables. A positive treatment effect of 1.06 was simulated by adding an independent random draw from a χ^2 distribution with 4 degrees of freedom to the treated group. Five percent of the patients were selected at random to be censored, and their censoring time was drawn from a uniform distribution.

We compare the performance of the implementation of the covariate adjusted estimator described in Section 6 of [5] to the unadjusted Kaplan-Meier-based estimator described in [5]. Wald-type confidence intervals and corresponding tests of the null of no effect are reported.

4.2 Performance Criteria

We compare the type I error and power of tests of the null hypothesis of no treatment effect based on unadjusted and adjusted estimators, both within and across estimands. For each estimand, we also compare the estimator bias, variance, and mean squared error of the unadjusted and the adjusted approaches. We report the relative efficiency of the unadjusted relative to the adjusted estimator (ratio of variance of adjusted estimator to variance of

medRxiv preprint doi: https://doi.org/10.1101/2020.04.19.20069922; this version posted April 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

unadjusted estimator)— since both of these estimators are asymptotically unbiased, one minus the relative efficiency approximately has the interpretation as the proportion reduction in sample size needed for a covariate adjusted estimator to achieve the same statistical power as the unadjusted estimator, at local alternatives.

5 Simulation Results

5.1 Ordinal outcomes

We present results based on bootstrap-based confidence intervals and hypothesis tests here, while results based on closed-form inference methods are presented in Appendix C. Tables 2, 3, and 4 display results for the difference in means, Mann-Whitney estimand, and log odds ratio, respectively. Type 1 error control of the covariate adjusted methods was comparable to or better than that of the unadjusted methods across most settings. The covariate adjusted methods achieved higher power across all settings. Absolute gains in power varied from 4% to 7% for the DIM, 1% to 7% for the MW estimand, and 3% to 5% for the LOR. The relative efficiency of covariate adjusted methods relative to unadjusted methods varied from 0.83 to 0.89 for the DIM, 0.82 to 0.91 for the MW estimand, and 0.83 to 0.88 for the LOR.

n	Estimator Type	Effect	P(reject H_0)	MSE	Bias	Variance	Rel. Eff.
100	Unadjusted	0.000	0.070	0.023	-0.006	0.023	1.000
100	Adjusted	0.000	0.052	0.019	-0.007	0.019	0.827
100	Unadjusted	0.303	0.504	0.022	-0.008	0.022	1.000
100	Adjusted	0.303	0.576	0.019	-0.004	0.019	0.842
200	Unadjusted	0.000	0.049	0.010	-0.002	0.010	1.000
200	Adjusted	0.000	0.043	0.009	-0.003	0.009	0.863
200	Unadjusted	0.303	0.794	0.012	-0.003	0.012	1.000
200	Adjusted	0.303	0.845	0.010	-0.000	0.010	0.872
500	Unadjusted	0.000	0.064	0.005	-0.001	0.005	1.000
500	Adjusted	0.000	0.056	0.004	-0.000	0.004	0.837
500	Unadjusted	0.195	0.822	0.005	-0.000	0.005	1.000
500	Adjusted	0.195	0.863	0.004	0.001	0.004	0.891
1000	Unadjusted	0.000	0.044	0.002	0.000	0.002	1.000
1000	Adjusted	0.000	0.041	0.002	-0.000	0.002	0.849
1000	Unadjusted	0.136	0.815	0.002	0.000	0.002	1.000
1000	Adjusted	0.136	0.872	0.002	-0.000	0.002	0.889

Table 2: Results for difference in means (DIM) in hospitalized population using BCa bootstrap for inference.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.19.20069922; this version posted April 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

n	Estimator Type	Effect	P(reject H_0)	MSE	Bias	Variance	Rel. Eff.
100	Unadjusted	0.500	0.062	0.003	-0.002	0.003	1.000
100	Adjusted	0.500	0.061	0.002	-0.003	0.002	0.824
100	Unadjusted	0.627	0.713	0.002	-0.002	0.002	1.000
100	Adjusted	0.627	0.780	0.002	-0.002	0.002	0.849
200	Unadjusted	0.500	0.047	0.001	-0.001	0.001	1.000
200	Adjusted	0.500	0.045	0.001	-0.001	0.001	0.865
200	Unadjusted	0.627	0.949	0.001	-0.001	0.001	1.000
200	Adjusted	0.627	0.966	0.001	-0.000	0.001	0.878
500	Unadjusted	0.500	0.058	0.001	-0.000	0.001	1.000
500	Adjusted	0.500	0.053	0.000	-0.000	0.000	0.843
500	Unadjusted	0.582	0.951	0.001	-0.000	0.001	1.000
500	Adjusted	0.582	0.963	0.000	-0.000	0.000	0.905
1000	Unadjusted	0.500	0.043	0.000	0.000	0.000	1.000
1000	Adjusted	0.500	0.049	0.000	-0.000	0.000	0.844
1000	Unadjusted	0.557	0.934	0.000	0.000	0.000	1.000
1000	Adjusted	0.557	0.950	0.000	-0.000	0.000	0.890

Table 3: Results for Mann Whitney (MW) estimand in hospitalized population using BCa bootstrap for inference.

5.2 Time-to-event outcomes

We present two sets of results. Table 5 displays the results for estimators that adjust for age and sex along with the four other variables described in Section 4.1.2. Type 1 error control of the covariate adjusted methods was comparable to or better than that of the unadjusted methods across most settings. The covariate adjusted methods achieved higher power across all settings, with absolute gains varying from 3% to 8%. The relative efficiency of covariate adjusted methods relative to unadjusted methods improved monotonically as a function of sample size, from approximately 0.95 when $n = 100$ to approximately 0.80 when $n = 1000$. Table 6 displays the results of the analysis that only adjusts for age and sex. In this case, the gains of the covariate adjusted methods relative to the unadjusted methods are more modest, with absolute gains in power of approximately 1% and relative efficiency ranging from 0.97 to 1.00. These results suggest that there can be a meaningful benefit from adjusting for prognostic covariates beyond just age and sex.

5.3 Main Findings

We found that adjusting for variables beyond just age and sex can add substantial value. Specifically, for a time-to-event intubation outcome among hospitalized patients at Weill Cornell Medicine, we saw substantial gains from additionally adjusting for hypertension, dyspnea, whether the patient required supplemental oxygen at ED presentation, and the presence of bilateral infiltrates in the chest x-ray. For ordinal outcomes, we saw that the log

medRxiv preprint doi: https://doi.org/10.1101/2020.04.19.20069922; this version posted April 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

n	Estimator Type	Effect	P(reject H_0)	MSE	Bias	Variance	Rel. Eff.
100	Unadjusted	0.000	0.064	–	–	–	–
100	Adjusted	0.000	0.048	–	–	–	–
100	Unadjusted	-0.686	0.322	–	–	–	–
100	Adjusted	-0.686	0.355	–	–	–	–
200	Unadjusted	0.000	0.040	0.080	0.004	0.081	1.000
200	Adjusted	0.000	0.034	0.069	0.006	0.069	0.856
200	Unadjusted	-0.686	0.572	0.111	-0.000	0.111	1.000
200	Adjusted	-0.686	0.628	0.096	-0.004	0.096	0.863
500	Unadjusted	0.000	0.062	0.035	0.002	0.035	1.000
500	Adjusted	0.000	0.064	0.029	0.000	0.029	0.826
500	Unadjusted	-0.408	0.577	0.038	-0.001	0.038	1.000
500	Adjusted	-0.408	0.626	0.033	-0.001	0.033	0.868
1000	Unadjusted	0.000	0.043	0.015	0.000	0.015	1.000
1000	Adjusted	0.000	0.039	0.013	0.000	0.013	0.850
1000	Unadjusted	-0.278	0.584	0.016	-0.000	0.016	1.000
1000	Adjusted	-0.278	0.626	0.014	0.002	0.014	0.878

Table 4: Results for log-odds ratio (LOR) in hospitalized population using BCa bootstrap for inference. When n=100, the estimator exhibited non-robust behavior, as some samples included no observations in the treatment arm of at least one outcome level.

odds ratio estimator has undesirable performance when the sample size is small ($n = 100$). Because the log odds ratio estimand is also hard to interpret clinically in the likely scenario that a proportional odds assumption fails to hold, we recommend the difference of means or Mann Whitney estimands when the outcome is ordinal.

6 Recommendations for Treatment Effect Definition and Primary Efficacy Analysis

1. **Randomization scheme.** Use stratified block randomization, with strata including site and calendar time (which may be strong predictors of the outcome and for which there may be too many to adjust for at the end of the trial). Prognostic baseline variables such as age, disease severity, and comorbidities will be adjusted for in the primary analysis (see point 3 below), rather than being stratified on in the randomization procedure.

2. **Estimand when the outcome is ordinal.** In selecting a treatment effect definition, if a utility function can be agreed upon to transform the outcome to a score with a clinically meaningful scale, then we recommend using the difference between the transformed means in the treatment and control arms. Otherwise, we recommend using the unweighted difference between means or the Mann-Whitney estimand. We recommend against estimating log odds ratios, since these are difficult to interpret.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.19.20069922; this version posted April 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Sample Size	Estimator Type	Effect	P(reject H_0)	MSE	Bias	Variance	Rel. Eff.
100	Unadjusted	0	0.011	0.781	0.018	0.780	1.000
100	Adjusted	0	0.038	0.740	0.006	0.740	0.948
100	Unadjusted	1.06	0.085	0.570	-0.265	0.500	1.000
100	Adjusted	1.06	0.160	0.546	-0.251	0.483	0.957
200	Unadjusted	0	0.048	0.481	-0.013	0.481	1.000
200	Adjusted	0	0.055	0.418	-0.013	0.418	0.869
200	Unadjusted	1.06	0.326	0.328	-0.145	0.307	1.000
200	Adjusted	1.06	0.397	0.293	-0.123	0.278	0.893
500	Unadjusted	0	0.050	0.201	-0.003	0.201	1.000
500	Adjusted	0	0.052	0.164	-0.005	0.164	0.814
500	Unadjusted	1.06	0.729	0.151	-0.070	0.146	1.000
500	Adjusted	1.06	0.810	0.129	-0.065	0.125	0.855
1000	Unadjusted	0	0.048	0.100	0.001	0.100	1.000
1000	Adjusted	0	0.045	0.079	0.001	0.079	0.793
1000	Unadjusted	1.06	0.959	0.079	-0.060	0.076	1.000
1000	Adjusted	1.06	0.986	0.065	-0.058	0.061	0.819

Table 5: Results for difference in restricted mean survival times (RMST) in hospitalized population using Wald-based inference when the estimator adjusts for age, sex, and the four other variables described in Section 4.1.2.

3. **Covariate adjustment.** Adjust for prognostic baseline variables to improve precision and power. We expect improvements to be substantial since there are already several known prognostic baseline variables, e.g., age and co-morbidities. The baseline variables should be specified before the trial is started (or should be selected using only blinded data from the trial using a prespecified algorithm) and the number of variables adjusted for should be no more than (roughly) the total sample size divided by 20.

4. **Confidence intervals and hypothesis testing.** we recommend that the nonparametric bootstrap (BCa method) be used with 10000 replicates for constructing a confidence interval. The entire estimation procedure, including any model fitting, should be repeated in each replicate data set. Hypothesis tests can be conducted either by inverting the confidence interval or by permutation methods — the latter can be especially useful in smaller sample size trials in order to achieve the desired Type I error rate.

5. **Early stopping for efficacy or futility.** Rules for early stopping such as O'Brien-Fleming boundaries can be directly applied, except where z-statistics are constructed using the covariate adjusted estimator described above and the covariance between statistics at different analysis times is estimated using nonparametric bootstrap as described above. The timing of analyses can be based on the information accrued, which has the advantage of reducing the average sample size even in trials with no treatment effect, which results

medRxiv preprint doi: https://doi.org/10.1101/2020.04.19.20069922; this version posted April 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

n	Estimator Type	Effect	P(reject H_0)	MSE	Bias	Variance	Rel. Eff.
100	Unadjusted	0	0.011	0.781	0.018	0.780	1.000
100	Adjusted	0	0.015	0.771	0.014	0.771	0.987
100	Unadjusted	1.06	0.085	0.570	-0.265	0.500	1.000
100	Adjusted	1.06	0.098	0.570	-0.263	0.501	1.000
200	Unadjusted	0	0.048	0.481	-0.013	0.481	1.000
200	Adjusted	0	0.050	0.476	-0.014	0.476	0.989
200	Unadjusted	1.06	0.326	0.328	-0.145	0.307	1.000
200	Adjusted	1.06	0.337	0.326	-0.141	0.306	0.995
500	Unadjusted	0	0.050	0.201	-0.003	0.201	1.000
500	Adjusted	0	0.049	0.196	-0.003	0.196	0.975
500	Unadjusted	1.06	0.729	0.151	-0.070	0.146	1.000
500	Adjusted	1.06	0.742	0.147	-0.069	0.143	0.978
1000	Unadjusted	0	0.048	0.100	0.001	0.100	1.000
1000	Adjusted	0	0.047	0.096	0.001	0.096	0.963
1000	Unadjusted	1.06	0.959	0.079	-0.060	0.076	1.000
1000	Adjusted	1.06	0.963	0.077	-0.060	0.073	0.972

Table 6: Results for difference in restricted mean survival times (RMST) in hospitalized population using Wald-based inference when the estimator only adjusts for age and sex.

from earlier futility stopping.

6. **Plotting the CDF when the outcome is ordinal.** Regardless of which treatment effect definition is used in the primary efficacy analysis, we recommend that the estimated CDF of the primary outcome be plotted for each study arm when the outcome is ordinal. This is analogous to plotting Kaplan-Meier curves for time-to-event outcomes, which can help in interpreting the trial results. For example, if the observed probability of death is similar in the two arms but the probability of ICU admissions is smaller in the treatment arm, then this will be evident in the plots. We also recommend to plot simultaneous confidence bands to represent uncertainty about the estimated curves. To improve precision, covariate adjustment can be used when constructing the estimated CDFs and the corresponding confidence band.

7. **Missing covariates.** We recommend handling missing covariates by imputing these covariates based only on data from those covariates that were observed. Importantly, to ensure that treatment assignment is randomized conditionally on the imputed covariates, no treatment or outcome information should be used in this imputation.

8. **Missing ordinal outcomes.** We recommend handling missing ordinal outcomes using doubly robust methods whose validity relies on the outcomes are missing at random conditionally on the covariates and treatment assignment. One such doubly robust approach can be evaluated by applying the methods described in Appendix A.2, but with study

medRxiv preprint doi: https://doi.org/10.1101/2020.04.19.20069922; this version posted April 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

arm recoded as 0 to indicate that a patient was both randomized to study arm 0 (control) and had their outcome measured, 1 to indicate that a patient was both randomized to study arm 1 (treatment) and had their outcome measured, and -1 to indicate that the outcome is missing. When study arm is recoded in this way and the outcome is not missing completely at random but is missing at random conditionally on covariates, it is important that the model used for $\hat{\pi}$ described in Appendix A.2 actually depend on the baseline covariates, since this recoded treatment is not fully randomized.

9. **Loss to follow up with time-to-event outcomes.** We recommend accounting for loss-to-follow-up using doubly robust methods such as those described in [2, 5]. These methods rely on a potentially more plausible condition on the censoring distribution than do unadjusted methods. The covariate adjusted estimator for the restricted mean survival time in the time-to-event setting is double robust under censoring being independent of the outcome given baseline variables and arm assignment.

References

[1] Austin, P. C., A. Manca, M. Zwarenstein, D. N. Juurlink, and M. B. Stanbrook (2010). A substantial and confusing variation exists in handling of baseline covariates in randomized controlled trials: a review of trials published in leading medical journals. *Journal of Clinical Epidemiology 63*(2), 142–153.

[2] Benkeser, D., M. Carone, and P. B. Gilbert (2018). Improved estimation of the cumulative incidence of rare outcomes. *Statistics in medicine 37*(2), 280–293.

[3] CDC COVID-19 Response Team (2020). Severe outcomes among patients with coronavirus disease 2019 (covid-19) — united states, february 12–march 16, 2020. *MMWR Morb Mortal Wkly Rep 69*(12), 343–346.

[4] Chen, P.-Y. and A. A. Tsiatis (2001). Causal inference on the difference of the restricted mean lifetime between two groups. *Biometrics 57*(4), 1030–1038.

[5] Díaz, I., E. Colantuoni, D. F. Hanley, and M. Rosenblum (2019). Improved precision in the analysis of randomized trials with survival outcomes, without assuming proportional hazards. *Lifetime data analysis 25*(3), 439–468.

[6] Díaz, I., E. Colantuoni, and M. Rosenblum (2016). Enhanced precision in the analysis of randomized trials with ordinal outcomes. *Biometrics 72*(2), 422–431.

[7] Efron, B. and R. J. Tibshirani (1994). *An introduction to the bootstrap*. CRC press.

[8] (EMA), E. M. A. (2015). Guideline on adjustment for baseline covariates in clinical trials. Reference number EMA/CHMP/295050/2013. Committee for Medicinal Products for Human Use (CHMP).

medRxiv preprint doi: https://doi.org/10.1101/2020.04.19.20069922; this version posted April 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

[9] FDA (2019). Adjusting for covariates in randomized clinical trials for drugs and biologics with continuous outcomes. draft guidance for industry.

[10] FDA and EMA (1998). E9 statistical principles for clinical trials. *U.S. Food and Drug Administration: CDER/CBER. European Medicines Agency: CPMP/ICH/363/96*.

[11] Kaplan, E. L. and P. Meier (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association 53*(282), 457–481.

[12] McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological) 42*(2), 109–127.

[13] Moore, K. L. and M. J. van der Laan (2009). Covariate adjustment in randomized trials with binary outcomes: Targeted maximum likelihood estimation. *Statistics in Medicine 28*(1), 39.

[14] Rosenblum, M. and M. J. Van Der Laan (2010). Simple, efficient estimators of treatment effects in randomized trials using generalized linear models to leverage baseline variables. *The international journal of biostatistics 6*(1).

[15] Royston, P. and M. K. Parmar (2011). The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Statistics in medicine 30*(19), 2409–2421.

[16] Scharfstein, D. O., A. Rotnitzky, and J. M. Robins (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association 94*(448), 1096–1120.

[17] Steingrimsson, J. A., D. F. Hanley, and M. Rosenblum (2016). Improving precision by adjusting for baseline variables in randomized trials with binary outcomes, without regression model assumptions. *Johns Hopkins University, Dept. of Biostatistics Working Papers. http://biostats.bepress.com/jhubiostat/paper280*.

[18] Van der Laan, M. J., M. Laan, and J. M. Robins (2003). *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media.

[19] Van der Laan, M. J. and S. Rose (2011). *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media.

[20] Van Der Laan, M. J. and D. Rubin (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics 2*(1).

[21] Vermeulen, K., O. Thas, and S. Vansteelandt (2015). Increasing the power of the mann-whitney test in randomized experiments through flexible covariate adjustment. *Statistics in medicine 34*(6), 1012–1030.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.19.20069922; this version posted April 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

[22] Wang, B., E. L. Ogburn, and M. Rosenblum (2019). Analysis of covariance in randomized trials: More precision and valid confidence intervals, without model assumptions. *Biometrics 75*(4), 1391–1400.

[23] Williamson, E. J., A. Forbes, and I. R. White (2014). Variance reduction in randomised trials by inverse probability weighting using the propensity score. *Statistics in medicine 33*(5), 721–737.

[24] Yang, L. and A. A. Tsiatis (2001). Efficiency study of estimators for a treatment effect in a pretest–posttest trial. *The American Statistician 55*(4), 314–321.

Appendix

A Estimands and estimators when the outcome is ordinal

A.1 Estimands

Let Y_0 denote a draw from the distribution of ordinal outcomes in treatment arm 0 and Y_1 denote an independent draw from the distribution of ordinal outcomes in treatment arm 1. The three estimands write as

$$\text{DIM: } E[u(Y_1)] - E[u(Y_0)],$$

$$\text{MW: } P(Y_1 > Y_0) + 0.5P(Y_1 = Y_0),$$

$$\text{LOR: } \underset{\beta \in \mathbb{R}}{\arg\min} \sum_{j=1}^{K} \left\{\text{logit}\, P(Y_1 \leq j) - \text{logit}\, P(Y_0 \leq j) - \beta\right\}^2.$$

Above, u is a prespecified real-valued transformation of an outcome.

A.2 Covariate adjusted estimator

Consider a setting in which we observe n independent draws $(X_1, A_1, Y_1), \ldots, (X_n, A_n, Y_n)$, where each X_i represents baseline covariates, each A_i represents a treatment, and each Y_i represents an outcome. We suppose that $A \perp Y|X$. We now derive an estimator for the CDF that is closely related to an estimator presented in Scharfstein et al. (1999) [16] and to targeted minimum loss-based estimators [20, 19].

For $\alpha := (\alpha(j))_{j=1}^{K-1} \in \mathbb{R}^{K-1}$ and $\beta \in \mathbb{R}^d$, define the following $\mathbb{R}^{K-1} \times \mathbb{R}^d$ function as follows:

$$\text{logit}\, m_{\alpha,\beta}(j, x) = \alpha(j) + \beta^\top x.$$

We will consider the treatment-stratified proportional odds working model for $P\{Y \leq j|A = a, X = x\}$ in which there exist (α_0, β_0) and (α_1, β_1) such that $P\{Y \leq j|A = a, X = x\} = m_{\alpha_a,\beta_a}(j, x)$ for all j, x, a. Importantly, we do not rely on this model being correct.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.19.20069922; this version posted April 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Let $\hat{\pi}(a|x)$ be an estimate of $P(A = a|X = x)$. In the clinical trial setting that we consider, this quantity can be estimated with the empirical marginal treatment probability. At the end of this subsection, we describe alternative approaches for estimating this quantity. Suppose that, for $a \in \{0, 1\}$, $\hat{\alpha}(a)$ and $\hat{\beta}(a)$ are chosen to minimize the following weighted empirical risk in (α, β):

$$-\sum_{j=1}^{K-1}\sum_{i=1}^{n}\frac{I\{A_i = a\}}{\hat{\pi}(A_i|X_i)}\log\left(m_{\alpha,\beta}(j, X_i)^{I\{Y_i\leq j\}}[1 - m_{\alpha,\beta}(j, X_i)]^{I\{Y_i>j\}}\right). \tag{1}$$

Each of these a-specific optimizations can be solved by running a weighted logistic regression on a repeated measures dataset of size $n \times (K - 1)$. Alternatively, they can be fitted using software for a proportional odds model. For both levels of the treatment a, it can be shown that $\hat{\alpha}(a)_1 \leq \hat{\alpha}(a)_2 \leq \ldots \leq \hat{\alpha}(a)_{K-1}$, and so, for any covariate value x, $m_{\hat{\alpha}(a),\hat{\beta}(a)}(\cdot, x)$ is a monotone nondecreasing function. Moreover, if our treatment-stratified proportional odds working model is correct, then $\hat{\alpha}(a)$ and $\hat{\beta}(a)$ are asymptotically linear estimators of the true underlying parameters.

Our covariate adjusted estimate of the CDF $\psi_a(j) := P(Y \leq j|A = a)$ is given by

$$\hat{\psi}_a(j) := \frac{1}{n}\sum_{i=1}^{n} m_{\hat{\alpha}(a),\hat{\beta}(a)}(j, X_i). \tag{2}$$

Because $m_{\hat{\alpha}(a),\hat{\beta}(a)}(\cdot, X_i)$ is monotone nondecreasing for all $i = 1, \ldots, n$, $\hat{\psi}_a(\cdot)$ is also monotone nondecreasing. The above estimator also satisfies the known constraint that $\hat{\psi}_a(j) \in [0, 1]$.

It can also be shown that $\hat{\psi}_a(j)$ is (i) doubly robust and (ii) efficient if both the treatment mechanism and the stratified proportional odds working model are correctly specified. To show this, we now establish that $\hat{\psi}_a(j)$ is in fact an augmented inverse probability weighted estimator. First, note that minimizing (1) to find $\hat{\alpha}_a$ and $\hat{\beta}_a$ implies that the following first-order condition is satisfied for all $j \in \{1, \ldots, K - 1\}$:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{I\{A_i = a\}}{\hat{\pi}(A_i|X_i)}\left[I\{Y_i \leq j\} - m_{\hat{\alpha}_a,\hat{\beta}_a}(j, X_i)\right] = 0.$$

Next, note that adding this to the right-hand side of (2) shows that

$$\begin{aligned}\hat{\psi}_a(j) &= \frac{1}{n}\sum_{i=1}^{n}\frac{I\{A_i = a\}}{\hat{\pi}(A_i|X_i)}\left[I\{Y_i \leq j\} - m_{\hat{\alpha}_a,\hat{\beta}_a}(j, X_i)\right] + \frac{1}{n}\sum_{i=1}^{n} m_{\hat{\alpha}(a),\hat{\beta}(a)}(j, X_i)\\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{I\{A_i = a\}}{\hat{\pi}(A_i|X_i)}I\{Y_i \leq j\} + \frac{1}{n}\sum_{i=1}^{n} m_{\hat{\alpha}(a),\hat{\beta}(a)}(j, X_i)\left[1 - \frac{I\{A_i = a\}}{\hat{\pi}(A_i|X_i)}\right].\end{aligned}$$

The above shows that $\hat{\psi}_a(j)$ is an augmented inverse probability weighted estimator for $\hat{\psi}_a(j)$, with the estimate of the outcome regression $x \mapsto P\{Y \leq j|A = a, X = x\}$ given by $x \mapsto m_{\hat{\alpha}(a),\hat{\beta}(a)}(j, x)$.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.19.20069922; this version posted April 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

We conclude by discussing our estimation of the treatment probability $P(A = a|X = x)$. Though this quantity can always be estimated by the empirical treatment probability in our randomized trial setting, there are generally advantages to estimating this quantity within a richer model. For example, a main-terms logistic regression of treatment on covariates could be used — in a randomized trial setting, this model is correctly specified provided that it includes an intercept term. The advantage of estimating known treatment probabilities via correctly specified parametric models has been discussed elsewhere – see, for example, [23] or, for a general treatment, Section 2.3.7 in [18].

B Estimands and censoring assumptions when the outcome is a time-to-event endpoint

Let T be a time-to-event outcome, C be a right-censoring time, A be a treatment indicator, and X be a collection of baseline covariates. Let τ be an investigator-specified truncation time that will be used to define the RMST, and let t^* be an investigator-specified time at which a comparison between the arm-specific cumulative incidences of the event is of interest.

Let T_1 be a random variable with the distributon of $T|A = 1$ and let T_0 be a random variable with the distribution of $T|A = 0$. The three estimands write as

$$\text{RMST: } E[\min\{T_1, \tau\}] - E[\min\{T_0, \tau\}],$$

$$\text{CID: } P\left(T_1 \leq t^*\right) - P\left(T_0 \leq t^*\right),$$

$$\text{RR: } \frac{P\left(T_1 \leq t^*\right)}{P\left(T_0 \leq t^*\right)}.$$

We conclude by describing the assumptions made by the adjusted and the unadjusted methods on the right censoring time C. Unadjusted methods assume that

$$C \perp\!\!\!\perp T|A. \tag{3}$$

The adjusted methods discussed in the main text assume that

$$C \perp\!\!\!\perp T|A, X, \tag{4}$$

which is often more plausible than (3).

C Additional results for ordinal outcomes

C.1 Results for Wald-style inference for hospitalized patients

C.2 Results for non-hospitalized COVID-19 patients

For the non-hospitalized COVID-19 positive population, the probabilities of being in each age group and of hospitalization and death are listed in Table 10. For this simulation, adjustment

medRxiv preprint doi: https://doi.org/10.1101/2020.04.19.20069922; this version posted April 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

n	Estimator Type	Effect	P(reject H_0)	MSE	Bias	Variance	Rel. Eff.
100	Unadjusted	0.000	0.068	0.023	-0.006	0.023	1.000
100	Adjusted	0.000	0.066	0.019	-0.007	0.019	0.827
100	Unadjusted	0.303	0.502	0.022	-0.008	0.022	1.000
100	Adjusted	0.303	0.592	0.019	-0.004	0.019	0.842
200	Unadjusted	0.000	0.042	0.010	-0.002	0.010	1.000
200	Adjusted	0.000	0.047	0.009	-0.003	0.009	0.863
200	Unadjusted	0.303	0.792	0.012	-0.003	0.012	1.000
200	Adjusted	0.303	0.857	0.010	-0.000	0.010	0.872
500	Unadjusted	0.000	0.060	0.005	-0.001	0.005	1.000
500	Adjusted	0.000	0.057	0.004	-0.000	0.004	0.837
500	Unadjusted	0.195	0.816	0.005	-0.000	0.005	1.000
500	Adjusted	0.195	0.869	0.004	0.001	0.004	0.891
1000	Unadjusted	0.000	0.045	0.002	0.000	0.002	1.000
1000	Adjusted	0.000	0.044	0.002	-0.000	0.002	0.849
1000	Unadjusted	0.136	0.826	0.002	0.000	0.002	1.000
1000	Adjusted	0.136	0.885	0.002	-0.000	0.002	0.889

Table 7: Results for difference in means (DIM) in hospitalized population using using Wald-based inference.

via the proportional odds model resulted in numerical instabilities in many simulated data sets. Therefore, we considered adjusting for age using a stratified estimator. That is, the conditional probability of each outcome in each age group was estimated using the empirical proportion of the sample with each outcome in each age group. As with the hospitalized simulation, we assumed that a treatment would have no effect on the probability of death but would decrease the odds of ventilation by the same relative amount in each age category. The relative reduction was selected for each sample size, such that a two-sided t-test that the average outcome between treatment arms were equal rejected the null hypothesis with approximately 80% power (n=200,500,1000) and 50% power (n=100).

medRxiv preprint doi: https://doi.org/10.1101/2020.04.19.20069922; this version posted April 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

n	Estimator Type	Effect	P(reject H_0)	MSE	Bias	Variance	Rel. Eff.
100	Unadjusted	0.500	0.072	0.003	-0.002	0.003	1.000
100	Adjusted	0.500	0.063	0.002	-0.003	0.002	0.824
100	Unadjusted	0.627	0.607	0.002	-0.002	0.002	1.000
100	Adjusted	0.627	0.695	0.002	-0.002	0.002	0.849
200	Unadjusted	0.500	0.048	0.001	-0.001	0.001	1.000
200	Adjusted	0.500	0.047	0.001	-0.001	0.001	0.865
200	Unadjusted	0.627	0.917	0.001	-0.001	0.001	1.000
200	Adjusted	0.627	0.959	0.001	-0.000	0.001	0.878
500	Unadjusted	0.500	0.060	0.001	-0.000	0.001	1.000
500	Adjusted	0.500	0.054	0.000	-0.000	0.000	0.843
500	Unadjusted	0.582	0.926	0.001	-0.000	0.001	1.000
500	Adjusted	0.582	0.950	0.000	-0.000	0.000	0.905
1000	Unadjusted	0.500	0.044	0.000	0.000	0.000	1.000
1000	Adjusted	0.500	0.047	0.000	-0.000	0.000	0.844
1000	Unadjusted	0.557	0.916	0.000	0.000	0.000	1.000
1000	Adjusted	0.557	0.940	0.000	-0.000	0.000	0.890

Table 8: Results for Mann Whitney (MW) estimand in hospitalized population using Wald-based inference.

n	Estimator Type	Effect	P(reject H_0)	MSE	Bias	Variance	Rel. Eff.
100	Unadjusted	0.000	0.054	–	–	–	–
100	Adjusted	0.000	0.059	–	–	–	–
100	Unadjusted	-0.686	0.306	–	–	–	–
100	Adjusted	-0.686	0.373	–	–	–	–
200	Unadjusted	0.000	0.042	0.080	0.004	0.081	1.000
200	Adjusted	0.000	0.044	0.069	0.006	0.069	0.856
200	Unadjusted	-0.686	0.563	0.111	-0.000	0.111	1.000
200	Adjusted	-0.686	0.633	0.096	-0.004	0.096	0.863
500	Unadjusted	0.000	0.060	0.035	0.002	0.035	1.000
500	Adjusted	0.000	0.065	0.029	0.000	0.029	0.826
500	Unadjusted	-0.408	0.574	0.038	-0.001	0.038	1.000
500	Adjusted	-0.408	0.640	0.033	-0.001	0.033	0.868
1000	Unadjusted	0.000	0.041	0.015	0.000	0.015	1.000
1000	Adjusted	0.000	0.047	0.013	0.000	0.013	0.850
1000	Unadjusted	-0.278	0.577	0.016	-0.000	0.016	1.000
1000	Adjusted	-0.278	0.642	0.014	0.002	0.014	0.878

Table 9: Results for log-odds ratio (LOR) in hospitalized population using Wald-based inference. When n=100, the estimator exhibited non-robust behavior, as some samples included no observations in the treatment arm of at least one outcome level.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.19.20069922; this version posted April 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Age	P(age)	P(hospitalization | age)	P(death | age)
0–19	0.05	0.02	0.00
20–44	0.29	0.18	0.00
45–54	0.18	0.25	0.01
55–64	0.18	0.25	0.02
65–74	0.17	0.36	0.04
75–84	0.09	0.45	0.07
≥ 85	0.06	0.51	0.19

Table 10: Non-hospitalized, COVID-19 positive population: Age and conditional outcome distributions used for the control arm in the ordinal outcome simulation study.

n	Estimator Type	Effect	P(reject H_0)	MSE	Bias	Variance	Rel. Eff.
100	Unadjusted	0.500	0.054	0.002	-0.001	0.002	1.000
100	Adjusted	0.500	0.084	0.002	-0.001	0.002	1.003
100	Unadjusted	0.594	0.644	0.001	-0.003	0.001	1.000
100	Adjusted	0.594	0.671	0.001	-0.002	0.001	0.985
200	Unadjusted	0.500	0.064	0.001	-0.000	0.001	1.000
200	Adjusted	0.500	0.038	0.001	-0.000	0.001	0.891
200	Unadjusted	0.594	0.930	0.001	0.001	0.001	1.000
200	Adjusted	0.594	0.866	0.001	0.000	0.001	0.905
500	Unadjusted	0.500	0.053	0.000	-0.000	0.000	1.000
500	Adjusted	0.500	0.030	0.000	-0.000	0.000	0.884
500	Unadjusted	0.561	0.887	0.000	-0.000	0.000	1.000
500	Adjusted	0.561	0.725	0.000	-0.000	0.000	0.878
1000	Unadjusted	0.500	0.053	0.000	0.000	0.000	1.000
1000	Adjusted	0.500	0.055	0.000	0.000	0.000	0.900
1000	Unadjusted	0.544	0.909	0.000	-0.000	0.000	1.000
1000	Adjusted	0.544	0.903	0.000	-0.000	0.000	0.918

Table 11: Results for Mann Whitney (MW) estimand in non-hospitalized population using BCa bootstrap for inference.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.19.20069922; this version posted April 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

n	Estimator Type	Effect	P(reject H_0)	MSE	Bias	Variance	Rel. Eff.
100	Unadjusted	0.500	0.060	0.002	-0.001	0.002	1.000
100	Adjusted	0.500	0.089	0.002	-0.001	0.002	1.003
100	Unadjusted	0.594	0.538	0.001	-0.003	0.001	1.000
100	Adjusted	0.594	0.635	0.001	-0.002	0.001	0.985
200	Unadjusted	0.500	0.068	0.001	-0.000	0.001	1.000
200	Adjusted	0.500	0.063	0.001	-0.000	0.001	0.891
200	Unadjusted	0.594	0.901	0.001	0.001	0.001	1.000
200	Adjusted	0.594	0.932	0.001	0.000	0.001	0.905
500	Unadjusted	0.500	0.050	0.000	-0.000	0.000	1.000
500	Adjusted	0.500	0.053	0.000	-0.000	0.000	0.884
500	Unadjusted	0.561	0.861	0.000	-0.000	0.000	1.000
500	Adjusted	0.561	0.910	0.000	-0.000	0.000	0.878
1000	Unadjusted	0.500	0.054	0.000	0.000	0.000	1.000
1000	Adjusted	0.500	0.058	0.000	0.000	0.000	0.900
1000	Unadjusted	0.544	0.885	0.000	-0.000	0.000	1.000
1000	Adjusted	0.544	0.917	0.000	-0.000	0.000	0.918

Table 12: Results for Mann Whitney (MW) estimand in non-hospitalized population using closed-form inference.

n	Estimator Type	Effect	P(reject H_0)	MSE	Bias	Variance	Rel. Eff.
100	Unadjusted	0.000	–	–	–	–	–
100	Adjusted	0.000	–	–	–	–	–
100	Unadjusted	-0.647	–	–	–	–	–
100	Adjusted	-0.647	–	–	–	–	–
200	Unadjusted	0.000	–	–	–	–	–
200	Adjusted	0.000	–	–	–	–	–
200	Unadjusted	-0.647	–	–	–	–	–
200	Adjusted	-0.647	–	–	–	–	–
500	Unadjusted	0.000	0.064	0.108	0.006	0.108	1.000
500	Adjusted	0.000	0.065	0.097	0.009	0.097	0.899
500	Unadjusted	-0.354	0.190	0.115	0.003	0.115	1.000
500	Adjusted	-0.354	0.188	0.103	0.006	0.103	0.897
1000	Unadjusted	0.000	0.054	0.052	0.001	0.052	1.000
1000	Adjusted	0.000	0.052	0.047	0.001	0.047	0.900
1000	Unadjusted	-0.246	0.184	0.055	0.002	0.055	1.000
1000	Adjusted	-0.246	0.195	0.050	0.002	0.050	0.908

Table 13: Results for log-odds ratio (LOR) in non-hospitalized population using BCa bootstrap for inference. When n=100, the estimator exhibited non-robust behavior, as some samples included no observations in the treatment arm of at least one outcome level.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.19.20069922; this version posted April 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

n	Estimator Type	Effect	P(reject H_0)	MSE	Bias	Variance	Rel. Eff.
100	Unadjusted	0.000	–	–	-	–	–
100	Adjusted	0.000	–	–	-	–	–
100	Unadjusted	-0.647	–	–	-	–	–
100	Adjusted	-0.647	–	–	–	–	–
200	Unadjusted	0.000	–	–	–	–	–
200	Adjusted	0.000	–	–	–	–	–
200	Unadjusted	-0.647	–	–	–	–	–
200	Adjusted	-0.647	–	–	–	–	–
500	Unadjusted	0.000	0.031	0.108	0.006	0.108	1.000
500	Adjusted	0.000	0.037	0.097	0.009	0.097	0.899
500	Unadjusted	-0.354	0.184	0.115	0.003	0.115	1.000
500	Adjusted	-0.354	0.204	0.103	0.006	0.103	0.897
1000	Unadjusted	0.000	0.045	0.052	0.001	0.052	1.000
1000	Adjusted	0.000	0.054	0.047	0.001	0.047	0.900
1000	Unadjusted	-0.246	0.180	0.055	0.002	0.055	1.000
1000	Adjusted	-0.246	0.212	0.050	0.002	0.050	0.908

Table 14: Results for log-odds ratio (LOR) in non-hospitalized population using Wald bootstrap for inference. When n=100, the estimator exhibited non-robust behavior, as some samples included no observations in the treatment arm of at least one outcome level.

n	Estimator Type	Effect	P(reject H_0)	MSE	Bias	Variance	Rel. Eff.
100	Unadjusted	0.000	0.059	0.011	-0.003	0.011	1.000
100	Adjusted	0.000	0.076	0.011	-0.001	0.011	0.994
100	Unadjusted	0.193	0.507	0.009	-0.006	0.009	1.000
100	Adjusted	0.193	0.562	0.008	-0.002	0.008	0.962
200	Unadjusted	0.000	0.065	0.006	0.001	0.006	1.000
200	Adjusted	0.000	0.044	0.005	-0.000	0.005	0.878
200	Unadjusted	0.193	0.821	0.005	0.003	0.005	1.000
200	Adjusted	0.193	0.825	0.004	0.002	0.004	0.887
500	Unadjusted	0.000	0.056	0.002	-0.000	0.002	1.000
500	Adjusted	0.000	0.034	0.002	-0.000	0.002	0.869
500	Unadjusted	0.125	0.791	0.002	-0.000	0.002	1.000
500	Adjusted	0.125	0.755	0.002	-0.001	0.002	0.862
1000	Unadjusted	0.000	0.052	0.001	0.000	0.001	1.000
1000	Adjusted	0.000	0.046	0.001	0.000	0.001	0.884
1000	Unadjusted	0.092	0.811	0.001	-0.001	0.001	1.000
1000	Adjusted	0.092	0.838	0.001	-0.001	0.001	0.902

Table 15: Results for difference in means (DIM) in non-hospitalized population using BCa bootstrap for inference.

medRxiv preprint doi: https://doi.org/10.1101/2020.04.19.20069922; this version posted April 23, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

n	Estimator Type	Effect	P(reject H_0)	MSE	Bias	Variance	Rel. Eff.
100	Unadjusted	0.000	0.057	0.011	-0.003	0.011	1.000
100	Adjusted	0.000	0.087	0.011	-0.001	0.011	0.994
100	Unadjusted	0.193	0.501	0.009	-0.006	0.009	1.000
100	Adjusted	0.193	0.594	0.008	-0.002	0.008	0.962
200	Unadjusted	0.000	0.062	0.006	0.001	0.006	1.000
200	Adjusted	0.000	0.066	0.005	-0.000	0.005	0.878
200	Unadjusted	0.193	0.832	0.005	0.003	0.005	1.000
200	Adjusted	0.193	0.878	0.004	0.002	0.004	0.887
500	Unadjusted	0.000	0.050	0.002	-0.000	0.002	1.000
500	Adjusted	0.000	0.054	0.002	-0.000	0.002	0.869
500	Unadjusted	0.125	0.790	0.002	-0.000	0.002	1.000
500	Adjusted	0.125	0.850	0.002	-0.001	0.002	0.862
1000	Unadjusted	0.000	0.052	0.001	0.000	0.001	1.000
1000	Adjusted	0.000	0.053	0.001	0.000	0.001	0.884
1000	Unadjusted	0.092	0.808	0.001	-0.001	0.001	1.000
1000	Adjusted	0.092	0.853	0.001	-0.001	0.001	0.902

Table 16: Results for difference in means (DIM) in non-hospitalized population using closed-form inference.