Clinical Infectious Diseases

INVITED ARTICLE

INNOVATIONS IN DESIGN, EDUCATION AND ANALYSIS (IDEA): Victor De Gruttola and Scott R. Evans, Section Editors

Evaluating the Efficacy of Therapies in Patients With Coronavirus Disease 2019

Dan-Yu Lin, Donglin Zeng, and Joseph J. Eron

University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

There is a proliferation of clinical trials worldwide to find effective therapies for patients diagnosed with coronavirus disease 2019 (COVID-19). The endpoints that are currently used to evaluate the efficacy of therapeutic agents against COVID-19 are focused on clinical status at a particular day or on time to a specific change of clinical status. To provide a full picture of the clinical course of a patient and make complete use of available data, we consider the trajectory of clinical status over the entire follow-up period. We also show how to combine the evidence of treatment effects on the occurrences of various clinical events. We compare the proposed and existing endpoints through extensive simulation studies. Finally, we provide guidelines on establishing the benefits of treatments.

Keywords. clinical trials; endpoints; severity rating; statistical power; totality of evidence.

Several studies have recently been completed and many more are currently underway or in the planning stages to investigate the efficacy and safety of therapeutic agents in patients diagnosed with coronavirus disease 2019 (COVID-19). A clinical trial of lopinavir/ritonavir (LPV/r) on adult patients hospitalized with severe COVID-19 was completed with unprecedented speed [1]; clinical trials of remdesivir on a spectrum of COVID-19 patients have just concluded or are still ongoing; [2–4] and the World Health Organization and partners recently launched Solidarity, a global megatrial of remdesivir, LPV/r, interferon beta-1a (IFN-β1a), chloroquine, and hydroxychloroquine [5].

Table 1 shows 6 remdesivir trials registered on ClinicalTrials.gov. The Capital Medical University in China has conducted 2 of those trials, 1 in patients with mild/moderate COVID-19 and 1 in patients with severe COVID-19 [2]. Gilead Sciences has also conducted 2 trials, 1 in patients with moderate disease and 1 in patients with severe disease [4]. In addition, the United States National Institute of Allergy and Infectious Diseases (NIAID) has conducted a trial of remdesivir [3] and is now evaluating the combination of baricitinib and remdesivir (ClinicalTrials.gov identifier NCT04401579). Finally, INSERM (the French National Institute of Health and Medical Research) is conducting a trial of remdesivir, LPV/r, IFN-β1a , and hydroxychloroquine.

The efficacy of a therapeutic agent is assessed mainly in terms of the primary endpoint used in a clinical trial. Table 1 shows the primary endpoints adopted by the aforementioned 6 remdesivir trials. The primary endpoints are quite different among these trials, even for patients with similar disease severity at enrollment. Combining data from these trials would enable a more accurate assessment of the efficacy of remdesivir than separate analyses, but data from studies with different endpoints cannot be efficiently combined. Without a common endpoint, it would also be difficult to compare the efficacy of remdesivir with that of other agents in the future.

The currently used primary endpoints are focused on clinical status at a particular day or on the time to a specific improvement of clinical status. To fully represent important clinical outcomes and make efficient use of available data, we propose using the entire clinical course of a patient to assess the efficacy of COVID-19 therapy. Specifically, we evaluate the effect of treatment on the clinical-status trajectory over the follow-up period by regarding daily ratings of clinical status as repeated measures of this trajectory. In addition, we combine the evidence of treatment effects on all levels of improvement of clinical status over time, as well as all levels of deterioration, including critical illness and death. Finally, we demonstrate the advantages of the proposed methods over the existing ones through extensive simulation studies using empirical data from recently completed COVID-19 trials [1–3].

METHODS

The clinical status of a COVID-19 patient is commonly rated on a 7-category ordinal scale: 1, not hospitalized, with resumption of normal activities; 2, not hospitalized, but unable to resume normal activities; 3, hospitalized, not requiring supplemental oxygen; 4, hospitalized, requiring supplemental oxygen; 5, hospitalized, requiring nasal high-flow oxygen therapy, noninvasive mechanical ventilation, or both; 6, hospitalized,

Received 11 June 2020; editorial decision 13 August 2020; published online 21 August 2020.

Correspondence: D. -Y. Lin, Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599-7420 (lin@bios.unc.edu).

Clinical Infectious Diseases® 2021;72(6):1093–100

DOI: 10.1093/cid/ciaa1231

Table 1. Clinical Trials of Remdesivir for Patients With Coronavirus Disease 2019 Registered on ClinicalTrials.gov

Registration Number	Status at Enrollment	Location	Sponsor	Study Size	Primary Endpoint
NCT04252664	Mild or moderate	Wuhan, China	Capital Medical University, China	308	Time to clinical recovery by day 28[a]
NCT04257656	Severe	Wuhan, China	Capital Medical University, China	453	Time to clinical improvement by day 28[b]
NCT04280705	Hospitalized	Global	US NIAID	800	Time to recovery by day 29[c]
NCT04292730	Moderate	Global	Gilead Sciences	600	Clinical status at day 11[d]
NCT04292899	Severe	Global	Gilead Sciences	400	Clinical status at day 14[e]
NCT04315948	Hospitalized	Global	INSERM, France	3100	Clinical status at day 15[f]

Abbreviations: INSERM, French National Institute of Health and Medical Research; US NIAID, United States National Institute of Allergy and Infectious Diseases.

[a]Time (in hours) from initiation of study treatment until normalization of fever, respiratory rate, and oxygen saturation, and alleviation of cough, sustained for at least 72 hours.

[b]Time (in days) from initiation of study treatment until a decline of 2 categories from status at randomization on a 6-category ordinal scale of clinical status or live discharge from the hospital, whichever occurs first.

[c]First day on which the patient reaches 1 of the 3 least severe categories on an 8-point ordinal scale of severity rating.

[d]Distribution of severity rating on a 7-point ordinal scale at day 11.

[e]Distribution of severity rating on a 7-point ordinal scale at day 14.

[f]Distribution of severity rating on a 7-point ordinal scale at day 15.

requiring extracorporeal membrane oxygenation (ECMO), invasive mechanical ventilation, or both; and 7, death [1, 6–8]. This severity-rating system was adopted by the Chinese LPV/r trial [1] and the INSERM trial. It was also adopted by the Chinese remdesivir trial, although the 2 outpatient categories were merged [2]. NIAID also adopted this 7-category scale but divided Category 3 further to indicate whether ongoing medical care is required [3]. Gilead used the severity-rating scale of NIAID but merged the 2 outpatient categories [4].

In the Chinese trials of LPV/r and remdesivir [1, 2], the primary endpoint is time from randomization to clinical improvement, which is defined as a decline of 2 categories of severity (from status at randomization) or live discharge from the hospital, whichever occurs first. The primary endpoint in the INSERM trial is distribution of severity rating at day 15. NIAID also adopted this endpoint but later changed its primary endpoint to time to recovery, which is defined as hospital discharge or not requiring ongoing medical care [3]. The primary endpoints for the 2 Gilead trials were changed from hospital discharge by day 14 and normalization of fever and oxygen saturation by day 14 to distribution of severity rating at days 11 and 14, respectively [4]. Notably, the primary endpoint in each of the 6 trials captures only part of the clinical course of a patient.

Rather than focusing on a specific change in the severity rating over time or the severity rating at a particular day, it is less arbitrary and more comprehensive to consider the severity-rating trajectory over the follow-up period. This endpoint encapsulates the entire clinical course of a patient and represents all available clinical data. To prove the concept, we adopt the 7-category severity rating system used in the Chinese LPV/r trial and the INSERM trial and define "recovery" as hospital discharge.

Figure 1 displays the severity-rating trajectories over a 28-day period for 8 COVID-19 patients. In each of the 4 plots, the even-numbered patient has a higher severity-rating curve than the odd-numbered patient; however, the 2 patients in each plot have the same time to clinical improvement (the primary endpoint used in the Chinese trials of LPV/r and remdesivir) and, with the exception of patients 5 and 6, also have the same time to recovery (the primary endpoint used by NIAID). Clearly, the "clinical improvement" endpoint does not capture all levels of improvement, nor does it directly measure any deterioration of clinical status; "recovery" corresponds to a greater degree of improvement of clinical status for a patient requiring mechanical ventilation at enrollment than a patient requiring only ongoing medical care at enrollment, and it does not provide any information about the changes of clinical status for a patient who has not recovered by the end of follow-up. For instance, patient 7 improves by 1 category whereas patient 8 dies, yet they have the same value for the endpoint of "clinical improvement" or "recovery." The use of clinical status at a single time-point as the primary endpoint is also problematic: the severity rating at day 14 or 15 is the same for patients 1 and 2 and also the same for patients 5 and 6, despite their different severity-rating trajectories. Thus, a true determination of treatment efficacy requires examination of the entire severity-rating trajectory.

We can characterize the treatment effect on severity rating at a particular day by the difference of the mean severity ratings between treatment and control. Thus, we consider average severity rating—the sum of a patient's daily severity ratings over the follow-up period of interest divided by the number of follow-up days (Figure 1)—and characterize the overall treatment effect on severity-rating trajectory by the difference in the mean of average severity rating between treatment and control. We can perform standard linear regression analysis on average severity rating. If there are missing values, then we treat daily severity ratings as incomplete repeated measures and employ the technique of generalized estimating equations (GEEs) [9] (Supplementary Appendix 1).

We can also characterize the treatment effect on severity rating at a particular day by the odds ratio of lower severity

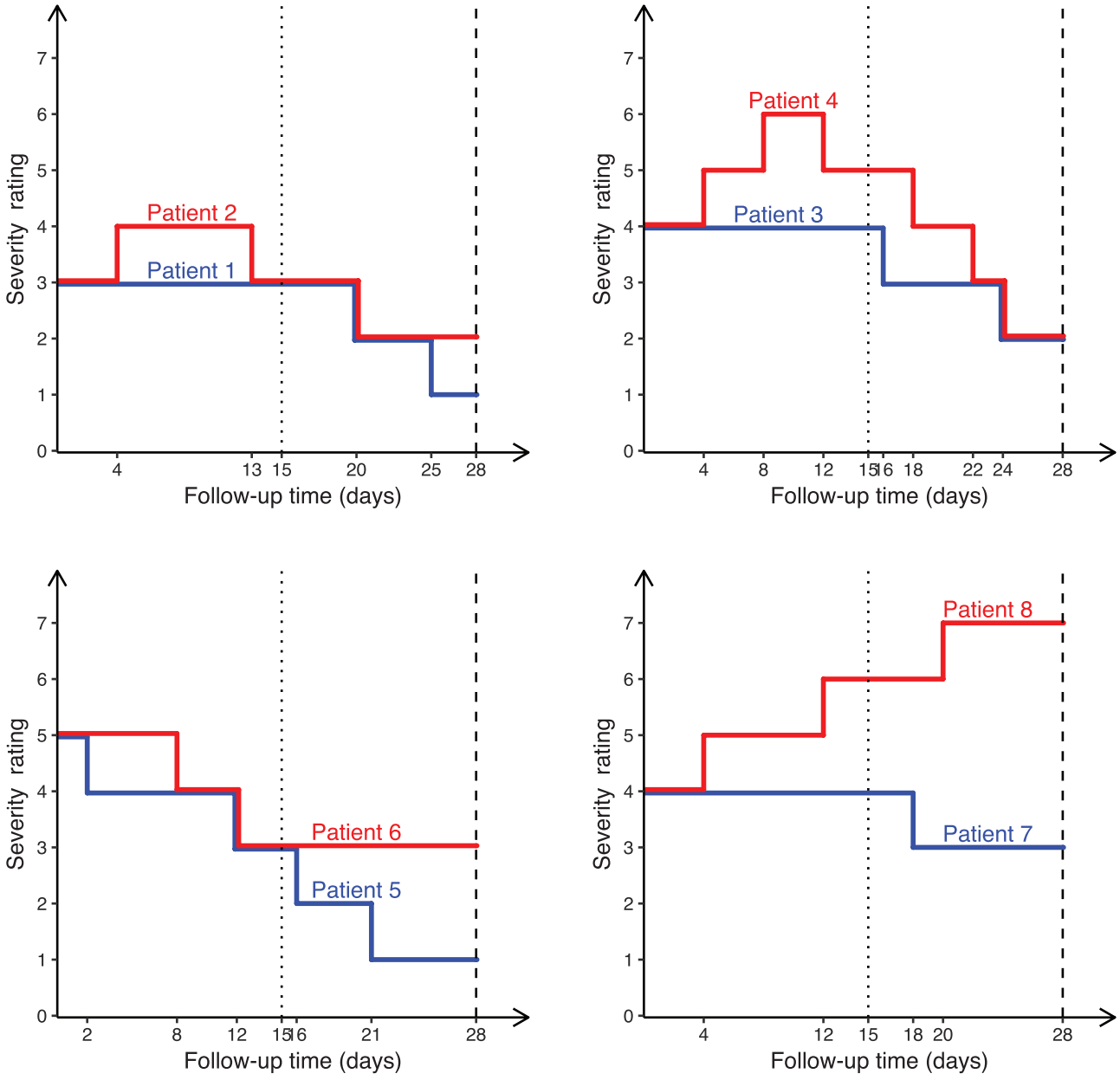

Figure 1. Severity-rating trajectories for 8 patients with coronavirus disease 2019. Severity is rated on a 7-point ordinal scale: 1, not hospitalized with resumption of normal activities; 2, not hospitalized, but unable to resume normal activities; 3, hospitalized, not requiring supplemental oxygen; 4, hospitalized, requiring supplemental oxygen; 5, hospitalized, requiring nasal high-flow oxygen therapy, noninvasive mechanical ventilation, or both; 6, hospitalized, requiring extracorporeal membrane oxygenation, invasive mechanical ventilation, or both; and 7, death. Average severity rating is the sum of daily severity ratings over the follow-up period divided by the duration of follow-up (ie, 28 days). Patients 1–8 acquire average severity ratings of 2.61, 3.04, 3.43, 4.29, 2.82, 3.71, 3.64, and 5.71, respectively, over the 28-day period.

(ie, the odds of falling into or below a severity-rating category vs falling above it for the treatment group divided by that of the control group) under the proportional odds model [10]; see Supplementary Appendix 1. If we assume that the odds ratio of lower severity is constant over the follow-up period of interest, then we can estimate or test this common odds ratio by applying GEE to the ordinal daily severity ratings. Although the assumption of a common odds ratio may not hold when treatment is effective, this formulation yields a nonparametric test of the null hypothesis that treatment has no effect on the distribution of severity rating at any time and an estimator of the overall treatment effect on the severity-rating trajectory. It is worthwhile to examine the odds ratios at various time points. The estimated common odds ratio is not very meaningful when individual odds ratios are in opposite directions.

We can describe the changes of clinical status over time by a multistate model [11] (Figure 2). We know each patient's initial state, that is, clinical status at randomization. We also

observe the time when each patient transits to a different state (ie, category), provided that the transition occurs before the end of follow-up. An effective treatment should accelerate transitions to less severe categories and slow transitions to more severe categories. We can view these transitions as multiple types of events [11, 12]. There are 9 types of events for hospitalized patients: improvement by 1, 2, 3, 4, or 5 categories and deterioration by 1, 2, 3, or 4 categories (from status at randomization). Each patient can potentially experience 6 events, whose types depend on the initial clinical status: A patient initially in Category 5 can improve by 1, 2, 3, or 4 categories or deteriorate by 1 or 2 categories; a patient initially in Category 4 can improve or deteriorate by 1, 2, or 3 categories; and so on.

As detailed in Supplementary Appendix 2, we formulate the treatment effects on the 5 levels of improvement and the 4 levels of deterioration through 9 Cox proportional hazards models [13]. Suppose that the hazard ratios of treatment vs control for the 5 levels of improvement are the same. Then we can estimate or test this common hazard ratio by the methodology of Wei, Lin and Weissfeld (WLW) [12], which is an extension of GEE to multiple events data. Although the 5 hazard ratios of improvement may not be the same when treatment is effective, this framework provides a valid test of the null hypothesis of no treatment effect on any level of improvement and a summary of the treatment effect on improvement. Likewise, we can use the WLW methodology to estimate or test a common hazard ratio for the 4 levels of deterioration. Furthermore, we can test the global null hypothesis of no treatment benefit on any change of clinical status. We refer to these 3 methods as WLW-imp, WLW-det, and WLW-ben. We suggest reporting the common hazard ratio of improvement and that of deterioration, as well as the 9 constituent hazard ratios.

The "clinical improvement" endpoint used in the Chinese trials of LPV/r and remdesivir pertains to improvement by 1 category for patients initially in Category 3 and to improvement by 2 categories for more ill patients, with time to improvement by 1 category also as a secondary endpoint [1, 2]. By contrast, WLW-imp clearly distinguishes between improvement by 1 vs 2 categories and also includes improvement by >2 categories. In addition, it automatically accounts for multiple comparisons.

The "recovery" endpoint used in the NIAID trial corresponds to a greater degree of improvement for a patient who is more ill at enrollment; however, all "recovery" events are treated the same in the endpoint. By contrast, WLW-imp considers each level of improvement separately and thus makes fuller and more precise use of the available data.

The ultimate goal of any therapy for COVID-19 patients is to prevent death. It would therefore be desirable to use 28-day mortality as the primary endpoint. However, the mortality rates are relatively low for COVID-19 [1–3], so a large number of patients is required to achieve good statistical power for detecting a moderate treatment difference in mortality. On the other hand, the patients who become critically ill (eg, on ECMO or invasive mechanical ventilation) are likely to suffer multiorgan failure, and even if they survive past day 28, they are still at risk of dying. Thus, critical illness is a good surrogate for 28-day mortality. A treatment comparison based on time to critical illness comprises a larger number of events and therefore tends to be more powerful than the mortality difference. We can combine the evidence of treatment effects on time to recovery and time to critical illness or death through the WLW methodology; we refer to this method as WLW-rc or WLW-rm.

RESULTS

To assess the operating characteristics of various endpoints and methods, we conducted a simulation study mimicking the design of the Chinese remdesivir trial [2]. We let 15%, 70%, and 15% of the patients belong to Category 3, 4, and 5, respectively, at enrollment [1, 2]. Within each category, we assigned patients to treatment or placebo at a ratio of 2:1. We simulated the transitions between the 7 categories of severity rating according to the multistate model [11] shown in Figure 2. As described in Supplementary Appendix 3, we chose a set of transition probabilities such that 70% of the placebo patients experienced the "clinical improvement" endpoint and 15% died by day 28 [1, 2].

As detailed in Supplementary Appendix 3, we considered 10 possible scenarios for the treatment effects on the transitions between the 7 categories. Case 1 pertains to the null hypothesis of no treatment effect. Cases 2–9 pertain to alternative hypotheses in which treatment accelerates the transition to a less severe category and slows the transition to a more severe category.

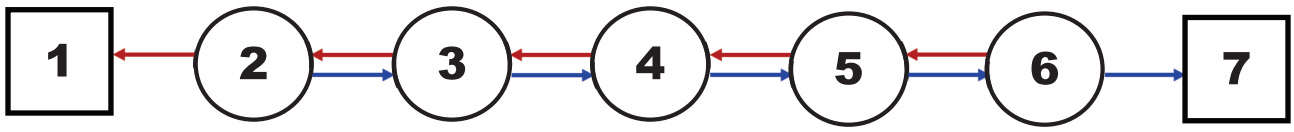

Figure 2. A multistate model for the changes of severity ratings over time in patients with coronavirus disease 2019. The 7 states correspond to the 7 categories of disease severity: 1, not hospitalized with resumption of normal activities; 2, not hospitalized, but unable to resume normal activities; 3, hospitalized, not requiring supplemental oxygen; 4, hospitalized, requiring supplemental oxygen; 5, hospitalized, requiring nasal high-flow oxygen therapy, noninvasive mechanical ventilation, or both; 6, hospitalized, requiring extracorporeal membrane oxygenation, invasive mechanical ventilation, or both; and 7, death. Possible transitions between adjacent states are shown by arrows. Direct transitions over multiple states (eg, 3 to 5, 4 to 1) are possible but are not explicitly indicated.

In Case 2, the magnitude of treatment effect is the same for all transitions. In Case 3, the treatment effect is stronger on the transition to a less severe category than to a more severe category; the opposite is true in Case 4. In Case 5, the treatment effect becomes weaker when the current state is more severe; in Case 6, the treatment effect becomes stronger when the current state is more severe. In Case 7, the treatment effect increases as severity at enrollment increases; in Case 8, the treatment effect decreases as severity at enrollment increases. Case 9 is the same as Case 2, but there is a patient-specific random effect to create heterogeneity. Case 10 is the same as Case 2, but treatment accelerates the transition to death. The treatment effects on various endpoints are shown in the top panel of Table 2. Of note, the proportional odds and proportional hazards assumptions do not hold in Cases 2–10; what are shown in Table 2 are the mean estimates of treatment effects based on a large number of simulated data sets.

We implemented linear models for severity rating at day 15 and average severity rating over days 1–28 or 8–28; proportional odds models for odds ratio of lower severity at day 15 and common odds ratio of lower severity over days 1–28 or 8–28; Cox models for time to clinical improvement, time to recovery, time to critical illness, and time to death; and logistic model for 28-day mortality. We stratified each by severity at enrollment. We also implemented the 4 WLW methods. In each scenario, we simulated 100 000 data sets, each with 453 patients. For each of the 15 methods, we tested the null hypothesis of no treatment effect at the 1-sided nominal significance level of 2.5% and estimated the rejection probability.

The results of the simulation study are summarized in the top panel of Table 3. We excluded logistic model for 28-day mortality from the summary because the estimation algorithm did not always converge due to the small number of deaths. Cox models for time to critical illness and time to death and WLW-det have slightly inflated type I error due to the small number of events. The other methods have reasonable type I error.

We now discuss the results in Cases 2–9. Cox models for time to clinical improvement and time to recovery have about 80% and 82% power, respectively, whereas linear models for average severity rating have about 90% power. The power of linear model for severity rating at day 15 is about 5% lower than that of linear models for average severity rating. Proportional odds models have similar power to linear models in most cases. As expected, Cox model for time to critical illness is more powerful than Cox model for time to death. WLW-imp is much more powerful than Cox models for time to clinical improvement and time to recovery. WLW-det is much more powerful than Cox models for time to critical illness and time to death. WLW-ben is nearly as powerful as linear and proportional odds models for severity-rating trajectory. WLE-rc is much more powerful than Cox model for time to critical illness and also tends to be more powerful than Cox model for time to recovery.

In Case 10, treatment has a beneficial effect on improvement and deterioration generally, except that it slightly increases the risk of death. In this case, Cox model for time to death has 1.6% power and WLW-det has 60%, whereas the other methods have 80% or higher power. In such situations, the tests based on composite endpoints should be used with caution, and having low probability to claim a beneficial treatment may be desirable.

We conducted a second simulation study mimicking the design of the NIAID remdesivir trial [3]. We let 15%, 40%, and 45% of the patients belong to Category 3, 4, and 5, respectively, at enrollment. Within each category, we assigned patients to treatment or placebo at a ratio of 1:1. We adopted the same set of transition probabilities and the same 8 scenarios of treatment effects as in the first simulation study but chose smaller effect sizes such that the Cox model for time to recovery has 80% power with 1000 patients. The treatment effects on various endpoints are shown in the bottom panel of Table 2.

The results of this simulation study are summarized in the bottom panel of Table 3. All 15 methods have correct type I error. In Cases 2–9, average severity rating, common odds ratio, and WLW-ben continue to have the highest power; Cox models for time to critical illness and time to death fare worse than before because of reduced treatment effects; and the logistic model for 28-day mortality is slightly more powerful than the Cox model for time to death. In Case 10, WLW-det and WLW-ben have only 35% and 68% power, respectively.

DISCUSSION

COVID-19 trials have rated clinical status on an ordinal scale of severity rating, with 6, 7, or 8 points. The rating system covers a multitude of important clinical outcomes, favorable or unfavorable. We propose 2 approaches for using the daily severity ratings over the follow-up period of interest to capture the totality of evidence on treatment efficacy: Average severity rating and common odds ratio pertain to severity-rating trajectory, and WLW deals with times to changes of clinical status. Severity-rating trajectory is of great interest if there are substantial fluctuations of severity rating over time, whereas times to changes of clinical status are most relevant if severity-rating curves are largely monotone. Average severity rating assigns a specific value or weight to each rating category, whereas common odds ratio and WLW rely only on the ordering of the rating scale. WLW can also accommodate clinical events not derived from a specific rating scale. We recommend using common odds ratio or WLW-ben as the primary analysis, depending on how clinical status changes over time.

Time to recovery and 28-day mortality are clinically meaningful endpoints that can be used with common odds ratio or WLW-ben. It is difficult to power a trial on a mortality endpoint, and time to recovery does not measure deterioration of clinical status. We may declare a therapy beneficial if WLW-ben

Table 2. Treatment Effects on 19 Endpoints in the Simulation Studies

Endpoint	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7	Case 8	Case 9	Case 10
Chinese remdesivir trial design										
SR15[a]	0	−0.59	−0.58	−0.61	−0.67	−0.57	−0.61	−0.57	−0.61	−0.57
ASR1–28[b]	0	−0.51	−0.50	−0.52	−0.56	−0.49	−0.52	−0.49	−0.51	−0.47
ASR8–28[b]	0	−0.61	−0.59	−0.63	−0.66	−0.59	−0.62	−0.58	−0.60	−0.55
OR15[c]	1	1.76	1.75	1.78	1.98	1.67	1.77	1.75	1.81	1.76
OR1–28[d]	1	1.65	1.63	1.67	1.79	1.59	1.66	1.63	1.67	1.65
OR8–28[d]	1	1.79	1.77	1.82	1.99	1.71	1.80	1.77	1.80	1.74
TTCI[e]	1	1.36	1.36	1.36	1.36	1.36	1.36	1.36	1.35	1.35
TTR[e]	1	1.37	1.37	1.37	1.38	1.37	1.38	1.37	1.37	1.38
TTC[e]	1	0.66	0.69	0.62	0.67	0.64	0.65	0.68	0.67	0.59
TTD[e]	1	0.54	0.59	0.49	0.61	0.46	0.52	0.59	0.64	1.06
MR28[f]	0	−6.47	−5.81	−7.22	−5.54	−7.65	−6.81	−5.87	−5.11	0.31
TTI1[g]	1	1.24	1.24	1.24	1.21	1.27	1.24	1.25	1.23	1.22
TTI2[g]	1	1.38	1.38	1.38	1.41	1.37	1.37	1.40	1.38	1.38
TTI3[g]	1	1.50	1.50	1.49	1.63	1.44	1.54	1.43	1.52	1.53
TTI4[g]	1	1.65	1.66	1.65	1.73	1.65	2.37	1.29	1.77	1.57
TTD1[h]	1	0.75	0.77	0.72	0.72	0.77	0.75	0.75	0.73	0.70
TTD2[h]	1	0.63	0.66	0.59	0.63	0.62	0.64	0.63	0.65	0.64
TTD3[h]	1	0.52	0.56	0.47	0.56	0.47	0.59	0.50	0.62	0.88
TTD4[h]	1	0.48	0.49	0.47	0.48	0.47	0.54	0.42	0.53	0.58
NIAID remdesivir trial design										
SR15[a]	0	−0.37	−0.36	−0.38	−0.40	−0.36	−0.38	−0.35	−0.39	−0.34
ASR1–28[b]	0	−0.32	−0.31	−0.33	−0.34	−0.32	−0.33	−0.30	−0.33	−0.28
ASR8–28[b]	0	−0.39	−0.37	−0.40	−0.41	−0.39	−0.40	−0.36	−0.38	−0.33
OR15[c]	1	1.39	1.38	1.40	1.46	1.37	1.39	1.38	1.43	1.34
OR1–28[d]	1	1.34	1.33	1.35	1.39	1.33	1.35	1.33	1.36	1.31
OR8–28[d]	1	1.40	1.39	1.42	1.47	1.39	1.41	1.39	1.41	1.34
TTCI[e]	1	1.21	1.20	1.21	1.19	1.22	1.20	1.21	1.20	1.19
TTR[e]	1	1.23	1.23	1.23	1.23	1.23	1.23	1.23	1.23	1.23
TTC[e]	1	0.81	0.83	0.79	0.82	0.80	0.81	0.83	0.81	0.77
TTD[e]	1	0.72	0.76	0.69	0.78	0.66	0.70	0.76	0.82	1.11
MR28[f]	0	−4.75	−4.20	−5.37	−3.87	−5.80	−5.11	−4.09	−3.07	1.38
TTI1[g]	1	1.14	1.14	1.14	1.12	1.16	1.14	1.15	1.13	1.11
TTI2[g]	1	1.22	1.22	1.22	1.22	1.22	1.21	1.24	1.21	1.21
TTI3[g]	1	1.28	1.28	1.28	1.32	1.27	1.31	1.24	1.30	1.29
TTI4[g]	1	1.35	1.35	1.34	1.40	1.33	1.52	1.19	1.44	1.34
TTD1[h]	1	0.85	0.86	0.83	0.84	0.85	0.84	0.85	0.83	0.81
TTD2[h]	1	0.76	0.78	0.73	0.77	0.74	0.77	0.75	0.80	0.88
TTD3[h]	1	0.68	0.71	0.64	0.69	0.66	0.80	0.57	0.83	0.87
TTD4[h]	1	0.71	0.74	0.68	0.70	0.70	0.87	0.55	0.91	0.90

Abbreviations: ASR, average severity rating; MR, mortality rate; NIAID, National Institute of Allergy and Infectious Diseases; OR, odds ratio; SR, severity rating; TTC, time to critical illness; TTCI, time to clinical improvement; TTD, time to death; TTR, time to recovery.

[a]Difference in the mean severity rating at day 15.

[b]Difference in the mean of average severity rating over days 1–28 (ASR1–28) or days 8–28 (ASR8–28).

[c]Odds ratio of lower severity at day 15.

[d]Common odds ratio of lower severity over days 1–28 (OR1–28) or days 8–28 (OR8–28).

[e]Hazard ratio for TTCI, TTR, TTC, or TTD.

[f]Difference in mortality rate (%) at day 28.

[g]Hazard ratio for time to improvement by k categories (TTIk) (k = 1, . . ., 4).

[h]Hazard ratio for time to deterioration by k categories (TTDk) (k = 1, . . ., 4).

or the treatment effect on common odds ratio is statistically significant and the treatment effects on 28-day mortality and time to recovery are in the right directions. WLW-imp and WLW-det can also serve as secondary endpoints.

Most COVID-19 trials were designed to follow patients for only 3–4 weeks. However, patients with severe illness, especially those experiencing prolonged ventilation or developing acute respiratory distress syndrome with a fibrotic component, may have unfavorable long-term outcomes. In addition, patients may require intensive care well beyond the end of the study, and some may die of COVID-19 in several months. Thus, we recommend that patients be followed

Table 3. Type I Error and Power (%) for Detecting Treatment Effects

Endpoint	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7	Case 8	Case 9	Case 10
Chinese remdesivir trial design										
SR15[a]	2.52	84.5	82.1	86.7	91.1	81.1	85.9	81.4	85.6	79.3
ASR1–28[b]	2.50	89.6	87.6	91.5	93.8	88.0	90.7	87.0	89.2	82.0
ASR8–28[b]	2.50	90.3	88.2	92.2	94.1	89.0	91.5	87.2	88.4	80.5
OR15[c]	2.68	85.4	84.4	86.2	95.2	77.9	86.2	83.8	88.7	85.1
OR1–28[d]	2.53	89.6	88.6	90.5	96.3	84.3	90.5	87.6	91.3	89.9
OR8–28[d]	2.58	91.1	90.2	92.1	97.3	86.3	92.1	89.1	91.6	88.4
TTCI	2.38	80.2	80.0	80.0	80.1	80.1	80.1	80.0	80.2	79.8
TTR	2.37	81.8	81.8	81.5	83.0	80.4	83.1	81.0	82.4	83.0
TTC	2.77	70.6	61.7	80.1	68.4	74.3	73.8	64.0	68.0	85.7
TTD	2.73	55.9	46.3	66.8	42.4	72.8	61.1	47.0	37.2	1.6
WLW-imp	2.38	86.3	86.4	85.9	90.8	84.8	92.7	79.3	88.2	85.2
WLW-det	2.81	78.8	69.8	87.4	79.2	80.6	72.2	78.9	72.2	60.2
WLW-ben	2.69	87.7	84.9	90.6	90.4	87.7	90.0	84.0	86.9	80.4
WLW-rc	2.63	83.2	79.3	87.0	82.7	84.1	84.7	80.3	82.1	89.7
NIAID remdesivir trial design										
SR15[a]	2.57	80.6	77.9	83.2	86.2	79.9	81.9	77.2	83.8	70.5
ASR1–28[b]	2.58	86.9	84.5	89.3	90.3	87.4	87.9	83.8	86.7	73.6
ASR8–28[b]	2.58	87.8	85.3	90.2	91.0	88.5	89.2	84.0	85.9	72.2
OR15[c]	2.58	80.6	78.6	82.4	89.5	77.4	81.1	78.9	86.2	71.2
OR1–28[d]	2.46	86.3	84.5	88.2	92.1	85.1	87.4	83.8	88.6	78.7
OR8–28[d]	2.53	87.8	86.1	89.6	93.9	86.3	88.7	85.4	88.7	75.3
TTCI	2.55	75.3	74.7	75.5	70.1	79.2	73.8	76.5	71.6	68.7
TTR	2.57	80.1	79.9	80.2	80.0	80.0	80.2	80.2	80.2	80.1
TTC	2.57	60.8	51.4	71.0	56.4	65.7	64.0	53.3	60.1	79.6
TTD	2.47	53.2	42.8	64.6	37.0	71.8	60.5	40.3	25.4	0.4
MR28[e]	2.49	54.7	44.6	66.0	38.8	72.9	61.1	43.0	25.7	0.6
WLW-imp	2.51	84.9	84.8	84.6	87.2	86.1	90.5	75.7	87.0	79.4
WLW-det	2.25	66.3	55.8	77.2	66.3	69.4	50.5	79.2	45.9	35.4
WLW-ben	2.47	84.3	80.9	87.6	85.6	86.2	84.0	83.2	79.4	68.4
WLW-rc	2.57	79.4	75.1	83.4	77.2	81.3	80.5	76.4	78.8	86.7

Abbreviations: ASR, average severity rating; MR, mortality rate; NIAID, National Institute of Allergy and Infectious Diseases; OR, odds ratio; SR, severity rating; TTC, time to critical illness; TTCI, time to clinical improvement; TTD, time to death; TTR, time to recovery; WLW-ben, Wei-Lin-Weissfeld method for benefit; WLW-det, Wei-Lin-Weissfeld method for deterioration; WLW-imp, Wei-Lin-Weissfeld method for improvement; WLW-rc, Wei-Lin-Weissfeld method for recovery and critical illness.

[a]Mean severity rating at day 15.

[b]Mean of average severity rating over days 1–28 (ASR1–28) or days 8–28 (ASR8–28).

[c]Odds of lower severity at day 15.

[d]Odds of lower severity over days 1–28 (OR1–28) or days 8–28 (OR8–28).

[e]Mortality rate at day 28.

as long as possible in order to evaluate long-term treatment effects.

Given the positive findings from the NIAID trial [3], any trials that compare remdesivir to placebo will likely be terminated, and any trials that compare multiple agents to placebo will likely switch placebo patients to active agents. Combining the data that have been collected thus far on all patients who have received remdesivir or placebo will enable a more accurate assessment of the effects of remdesivir (relative to placebo) on mortality and other outcomes than individual trials. Meta-analysis of summary statistics (ie, estimated treatment effects and standard errors) will be logistically simpler than, but statistically as efficient as, joint analysis of patient-level data [14]. It is difficult to establish the benefits of other treatments beyond that of remdesivir, so having common endpoints that are clinically relevant and statistically powerful is critically important to future COVID-19 trials.

Supplementary Data

Supplementary materials are available at *Clinical Infectious Diseases* online. Consisting of data provided by the authors to benefit the reader, the posted materials are not copyedited and are the sole responsibility of the authors, so questions or comments should be addressed to the corresponding author.

Notes

Acknowledgments. The authors are grateful to Ms Yu Gu for computational assistance and to 2 referees for constructive comments. The first author also thanks Drs Myron Cohen, Thomas Fleming, Dean Follmann, David Harrington, Xiaoying Gu, Lei Liu, Sheng Luo, Yeming Wang, L. J. Wei, Ying Yuan, and Hui Zhang for helpful discussions.

Financial support. This work was supported by the National Institutes of Health.

Potential conflicts of interest. The authors: No reported conflicts of interest. All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest.

References

1. Cao B, Wang Y, Wen D, et al. A trial of lopinavir-ritonavir in adults hospitalized with severe COVID-19. New Engl J Med **2020**; 382:1787–99.
2. Wang Y, Zhang D, Du G, et al. Remdesivir in adults with severe COVID-19: a randomised, double-blind, placebo-controlled, multicentre trial. Lancet **2020**; 395:1569–78.
3. Beigel JH, Tomashek KM, Dodd LE, et al. Remdesivir for the treatment of Covid-19—preliminary report [manuscript published online ahead of print 22 May 2020]. New Engl J Med **2020**. doi:10.1056/NEJ-Moa2007764.
4. Goldman JD, Lye DCB, Hui DS, et al. Remdesivir for 5 or 10 days in patients with severe Covid-19. New Engl J Med **2020**. doi:10.1056/NEJMoa2015301.
5. World Health Organization. "Solidarity" clinical trial for COVID-19 treatments. **2020**. Available at: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/global-research-on-novel-coronavirus-2019-ncov/solidarity-clinical-trial-for-covid-19-treatments.
6. International Severe Acute Respiratory and Emerging Infections Consortium. Home page. Available at: https://isaric.tghn.org/.
7. World Health Organization. Coronavirus disease (COVID-2019) R&D. Available at: http://www.who.int/blueprint/priority-diseases/key-action/novel-coronavirus/en/.
8. Royal College of Physicians. National Early Warning Score (NEWS) 2: Standardising the assessment of acute-illness severity in the NHS. **2017**. Available at: https://www.rcplondon. ac.uk/projects/outputs/national-early-warning-score-news-2.
9. Diggle P, Heagerty P, Liang K-Y, Zeger S. Analysis of longitudinal data. 2nd ed. Oxford, UK: Oxford University Press, **2002**.
10. McCullagh P. Regression models for ordinal data. J Royal Stat Soc B **1980**; 42:109–27.
11. Kalbfleisch JD, Prentice RL. The statistical analysis of failure time data. 2nd ed. Hoboken, NJ: John Wiley & Sons, **2011**.
12. Wei LJ, Lin DY, Weissfeld L. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. J Am Stat Assoc **1989**; 84:1065–73.
13. Cox DR. Regression models and life-tables. J Royal Stat Soc B **1972**; 34:187–202.
14. Lin DY, Zeng D. On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. Biometrika **2010**; 97:321–32.