APPENDIX:

Table S1. Selected Clinical Trials for Covid19 with outcomes and ordinal scales

Study	Primary outcome	Ordinal scale	Treatments	Study Design	Population
LOTUS ChiCTR2000029308	Time to clinical improvement Clinical improvement defined as two points improvement on a 7-category ordinal scale or discharge from the hospital, whichever came first.	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 ECMO, invasive mechanical ventilation or both 7) Death.	1) Lopinavir/ritonavir 2) Standard of care	Randomized, controlled, open-label trial. Randomization ratio: 1:1. Final:199 99 Lopinavir-Ritonavir 100 SOC	**Severe Covid-19 patients** hospitalized adult patients with confirmed SARS-CoV-2 infection, and Sao2 <94% while breathing ambient air or Pao2/Fio2 < 300 mm Hg.
ACTT NCT04280705	Time to recovery 28 days from randomization Recovery defined as category 1, 2 or 3.	1) Not hospitalized, no limitations on activities 2) Not hospitalized, limitation on activities and/or requiring home oxygen 3) Hospitalized, not requiring supplemental oxygen - no longer requires ongoing medical care 4) Hospitalized, not requiring supplemental oxygen - requiring	Stage 1: 1) Remdesivir 2) Placebo Stage 2: 1) Remdesivir + baricitinib 2) Remdesivir	Adaptive randomized, double-blind, placebo-controlled platform trial. Randomization ratio: 1:1 Stage 1 sample size: 400 recoveries Stage 2 sample size: 723 recoveries	Hospitalized adults with COVID-19, mild, moderate, and severe patients

		ongoing medical care (COVID-19 related or otherwise) 5) Hospitalized, requiring supplemental oxygen 6) Hospitalized, on non-invasive ventilation or high flow oxygen devices 7) Hospitalized, on invasive mechanical ventilation or ECMO 8) Death			
Remdesivir in Adults with Severe COVID-19 NCT04257656	<u>Time to clinical improvement:</u> Clinical improvement defined as two points improvement on a 6-category ordinal scale or discharge from the hospital, whichever came first.	1) hospital discharge; 2) hospitalized, not requiring supplemental oxygen; 3) hospitalized, requiring supplemental oxygen; 4) Hospitalized, requiring nasal high-flow oxygen therapy, noninvasive mechanical ventilation, or both 6) Hospitalized, requiring ECMO, invasive mechanical ventilation or both 6) death;	1) Remdesivir 2) Placebo	Randomized, double-blind, placebo-controlled Randomization ratio: 2:1 Planned sample size: 325 clinical improvements	Adults (≥18 years) with laboratory confirmed COVID-19 virus infection, and severe pneumonia signs or symptoms, and radiologically confirmed severe pneumonia (severe patients)

Randomized Evaluation of COVID-19 Therapy (RECOVERY) ISRCTN 50189673	All-cause mortality at 28 days after first randomization	None	First randomization: 1) Lopinavir/ritonavir 2) Low-dose Corticosteroid 3) Hydroxychloroquine 4) Azithromycin 5) Standard of care Second randomization (in worsening patients): 1) Tocilizumab 2) Standard of care	Adaptive, randomized, placebo-controlled, multicenter, multi-arm designed, open-label trial Planned sample size: unknown, depending on scale of pandemic	Hospitalized adults with SARS-CoV-2 infection (clinically suspected or laboratory confirmed), mild, moderate, and severe patients
Trial of Treatments for COVID-19 in Hospitalized Adults (DisCoVeRy) NCT04315948	Day 15 subject clinical status on 7-point ordinal scale	1) Not hospitalized, no limitations on activities 2) Not hospitalized, limitation on activities 3) Hospitalized, not requiring supplemental oxygen 4) Hospitalized, requiring supplemental oxygen 5) Hospitalized, on non-invasive ventilation or high flow oxygen devices 6) Hospitalized, on invasive mechanical ventilation or ECMO 7) Death.	1) Remdesivir 2) Lopinavir/ritonavir 3) Lopinavir/ritonavir + Interferon ß-1a 4) Hydroxychloroquine 5) Standard of care	Adaptive, randomized, open-label clinical trial Randomization ratio: participants 1:1:1:1:1 Planned sample size: 3100 participants	Hospitalized adult patients with laboratory-confirmed SARS-CoV-2 infection as determined by PCR, in any specimen < 72 hours prior to randomization Clinical assessment of pneumonia (evidence of rales/crackles on exam) AND SpO2 ≤ 94% on room air OR acute respiratory failure requiring supplemental oxygen, high flow oxygen devices, non-invasive ventilation, and/or mechanical ventilation.
Austrian CoronaVirus Adaptive	Time to clinical improvement	The 7-categories of the World Health Organization proposed scale, as	1) Hydroxychloroquine 2) Lopinavir/ritonavir 3) Standard of care	A multicenter, randomized, open label, controlled platform trial	Laboratory confirmed (i.e. PCR-based assay) infection with SARSCoV-

Clinical Trial (ACOVACT) NCT04351724	defined as time from randomization to a sustained improvement of at least one category on two consecutive days compared to the status at randomization measured on a seven-category ordinal scale (proposed by WHO).	follows: 1. Not hospitalized, no limitations on activities 2. Not hospitalized, limitation on activities; 3. Hospitalized, not requiring supplemental oxygen; 4. Hospitalized, requiring supplemental oxygen; 5. Hospitalized, on non-invasive ventilation or high flow oxygen devices; 6. Hospitalized, on invasive mechanical ventilation or ECMO; 7. Death.	4) Pooled plasma or IVIG from reconvalescent patients* *Treatment arm will only be opened when product and the respective necessary documents are available.	Randomization ratio for anti-viral treatment arms 1:1:1 Planned sample size is 500 participants The main study is for the comparison of anti-viral treatments Interim analysis after 50 patients in a treatment arm ACOVAT includes further sub-studies with additional randomization on top of the anti-viral treatments	2 (ideally but not necessarily ≤72 hours before randomization for “antiviral” treatments) OR radiological signs of COVID-19 in chest X-ray or computed tomography • Hospitalization due to SARS-CoV-2 infection (for anti-viral treatment arms) • Requirement of oxygen support (due to oxygen saturation <94% on ambient air or >3% drop in case of chronic obstructive lung disease) OR radiological signs of COVID-19

Table S2. Statistical analysis strategies including advantages and disadvantages.

Endpoint	Possible statistical analysis strategy	Advantages	Disadvantages
Binary analyses			
1. Proportion recovered/improved (by one or two categories on an ordinal scale) from baseline to specified time point like 2 weeks.	χ^2-Test, Boschloo's test of proportions, logistic regression	Accounts for baseline, clinically relevant, interpretation	Fixed time, loss of power due to dichotomization and using a binary endpoint
2. Mortality by day 28	χ^2-Test, Boschloo's test of proportions, logistic regression	Clinically meaningful, easy to interpret	Requires large sample sizes when mortality rate low
Ordinal scale analyses			
3. Ordinal outcome such as a 6-point scale at a fixed time point (e.g., 2 weeks),	Wilcoxon	Captures multiple states	Fixed time, no baseline, ties, scale categories should be objective and clinically meaningful, interpretation
4. Change in ordinal scale from baseline to follow-up	t-test or Wilcoxon	Accounts (partly) for baseline	Fixed time, edge effect (little room for improvement/worsening for those at tails/edges), ties, interpretation
5. Ordinal scale at a fixed time point (e.g., 2 weeks).	Proportional odds model	More robust (no normality assumption), score test is asymptotically like Wilcoxon test	Fixed time, Assumption of constant treatment to control odds ratio for each 1 unit change in ordinal scale; efficiency
6. Ordinal outcome at a fixed time point adjusted for baseline value	Generalized proportional odds model, analysis of covariance (ANCOVA)	Accounts for baseline, power, and is equivalent to the analysis of endpoint (4) when using Change in ordinal outcome from baseline to fixed time point with an ANCOVA adjusting for baseline as covariate	Fixed time, edge effect (little room for improvement/worsening for those at tails/edges), ties, interpretation
7. Average of ordinal scale over daily (or at least frequent) measurements during follow-up.	(potential analysis see 4)	Covers a predefined range of days, power	Duration and severity are mixed (e.g., 1-day death equals 7 days healthy), clinical relevance? Diluted effect if treatment effect established later

8. Average of ordinal scale over daily (or at least frequent) measurements during follow-up minus baseline ordinal scale measurement.	(potential analysis see 4)	covers a predefined range of days, power, accounts for baseline	Duration and severity are mixed, clinical relevance? Diluted effect if treatment effect established later
9. Average of ordinal scale over daily (or at least frequent) measurements during follow-up adjusted for baseline	(potential analysis see 6, ANCOVA)	covers a predefined range of days, more power than 8, accounts for baseline	Duration and severity are mixed, clinical relevance, interpretation, Diluted effect if treatment effect established later
10. Area under the curve of ordinal scale over frequent measurements.	Endpoint similar to 7 (potential analyses see 3 or 7)	covers a predefined range of days, area smaller if less time (similar problem for 7, 8, and 9)	Clinical relevance, diluted effect if treatment effect established later
Time-to-event analyses			
11. Time to a specified level of improvement (e.g., time to recovery)	Log-rank test (or Cox Regression) Deaths censored Max follow-up	Captures time element Interpretation: rate of recovery and median days to recovery	Does not consider starting point and individual courses to improvement (For unstratified log-rank test)
12. Time to a specific magnitude of improvement (e.g., 2-point improvement in ordinal scale)	Log-rank test (or Cox Regression) Deaths censored Max folllowup	Captures time element Interpretation: rate of 2-point improvement, median days to 2-point improvement	Improvements are considered equally regardless of starting point (e.g., from 6 to 4 considered equal to 3 to 1) (For Cox proportional hazard assumption)
13. Time to recovery and time to death	Standard Kaplan-Meier & Cox for death. Fine-Gray models for recovery	Provides treatment effects on two different aspects	Unclear how to combine the treatment effects in a single analysis.
Continuous data analyses			
14. Difference in days of oxygen use/intubation/etc	t-test, Wilcoxon	Possible statistical efficiency	May not correlate with eventual outcome
15. Difference in viral loads	t-test, Wilcoxon	Possible statistical efficiency	May not correlate with eventual outcome
16. Various biomarkers	t-test, Wilcoxon	Possible statistical efficiency	May not correlate with eventual outcome

Table S3. Demonstration of difference in statistical methods applied to reported study data from the LOTUS lopinavir/ritonavir study

	Based on observed data (n=199)	Hypothetical example each observation included twice (n=398)[+]
Proportional odds model		
Day 7 odds ratio	1.206 (95% CI: 0.710, 2.054) p=0.488	p=0.327
Day 14 odds ratio	1.376(95% CI: -0.835, 2.274) p=0.212	p=0.077
Day 21 odds ratio	1.196 (95%CI: 0.676, 2.124) p=0.539	p=0.386
Day 28 odds ratio	1.370 (95%CI: -0.740, 2.563) p=0.319	p=0.159
Average score (t-test)		
Mean difference	-0.1678 (95%CI: -0.575;0.240) p=0.418	p=0.250
Average score change from baseline (t-test)		
Mean difference, change from baseline	-0.247 (95%CI: -0.628,0.134) p=0.202	p=0.070
Time-to-recovery (log rank test)		
Recovery rate ratio	1.248 (95%CI: 0.899,1.732) p=0.187	p=0.061
Time-to-improvement (log rank test)		
Beneficial ratio	1.307 (95%CI: 0.946;1.807) p=0.105	p=0.022
Mortality (Fisher's exact test)		
Odds ratio	0.786 (95%CI:0.372, 1.644) p=0.602	p=0.390

Details of simulation

Ordinal trajectories for each subject were generated according to a linear random effects model with time index log of the day since randomization Informally, subject i drew a random curve of 'destiny' and foreach day of follow-up, the integer part of the line at that day was given as ordinal score. Except for the lagged effect scenario, the model is given by

$$Y_{id} = B0 + B1 \log(d) + B2\, Z \log(d) + b_{0i} + b_{1i} * \log(d) + W e_{id} \quad (1)$$

with b_{0i} distributed $N(0,1.5^2)$ and b_{1i} distributed $I \times N(-4,.3^2) + (1-I)\, N(7,s^2)$ with I distributed Bernoulli(p=0.10) for placebo and Bernoulli(p=0.05)~Be(.05) for treatment, e_{id} distributed $N(0,.25^2)$, and Z the indicator of the treatment group. Note that there is a treatment effect both on the speed of recovery (as B2 <0) and mortality as I has a different Bernoulli probability for the two groups.

For the lagged effect scenario, the day 1 treatment effect begins at day 8:

$$Y_{id} = B0 + B1 \log(d) + B2\, Z\, I(d>7) * \log(d-7) + b_{0i} + b_{1i} * Z * I(d>7) \log(d-7) + W e_{id} \quad (2)$$

With settings for the random variables as described for equation (1). Table S4 provides the parameter values used for the different scenarios.

Table S4: Table of parameters used for the various model. The feature that is changed relative to the reference case is bolded. All scenarios use equation (1) except for the lagged effect which uses equation (2)

Scenario	**B0**	**B1**	**B2**	**s**	**W**
Reference	0	-.05	-.10	.15	0
Lagged Effect*	0	-.05	-.10	.15	0
Faster Recovery	0	**-.10**	-.10	.15	0
Faster Mortality	0	-.05	-.10	**.30**	0
Only Mortality benefit	0	-.05	**0**	.15	0

Details of simulation enforcing proportional odds assumption at each time point

Random multinomial data were generated corresponding to baseline ordinal scores. Then a trajectory of ordinal scores was applied as method 1 above, except that the trajectories were generated with the same distribution for treatment and control arms. Treatment-arm proportions at observation days were then re-scaled to satisfy a proportional odds assumption with according to a common odds ratio for specific treatment effects each day (as specified in table S5). Additional simulation studies (not shown) demonstrated that blinded (pooled) pilot studies are not very informative for guiding the determination of the optimal time. Blinded (pooled) data provide information about the overall proportions in each category, but simple rules such as selecting the time where there are a certain proportion of good outcomes or when the distribution is the most variable do not seem to improve identification of the optimal time for evaluation. Note the one peculiarity of how these models are set up.

Table S5. Simulated power for different tests under different scenarios.

	True common odds ratio by day					**Empirical Power/Rejection Rates**			
	Days 1-10	**Days 11-13**	**Day 14**	**Day 21**	**Day 28**	**Proportional odds at day 14**	**Proportional odds at day 28**	**Log-rank (time to recovery)**	**Average score**
Scenario A	1	1	1	1.5	1.75	0.052	0.879	0.395	0.271
Scenario B	1	1	1.25	1.5	1.75	0.244	0.884	0.442	0.384
Scenario C	1	1.1	1.15	1.25	1.75	0.126	0.884	0.418	0.254

Figure S1. Stacked bar plots for ordinal scores and Kaplan-Meier curves for time-to-recovery for three scenarios for simulation method 2.