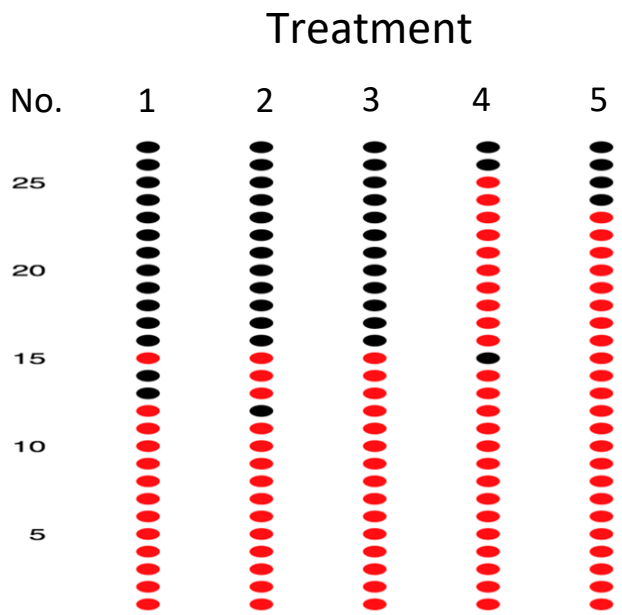
How to define the study endpoint via patient level multiple outcomes

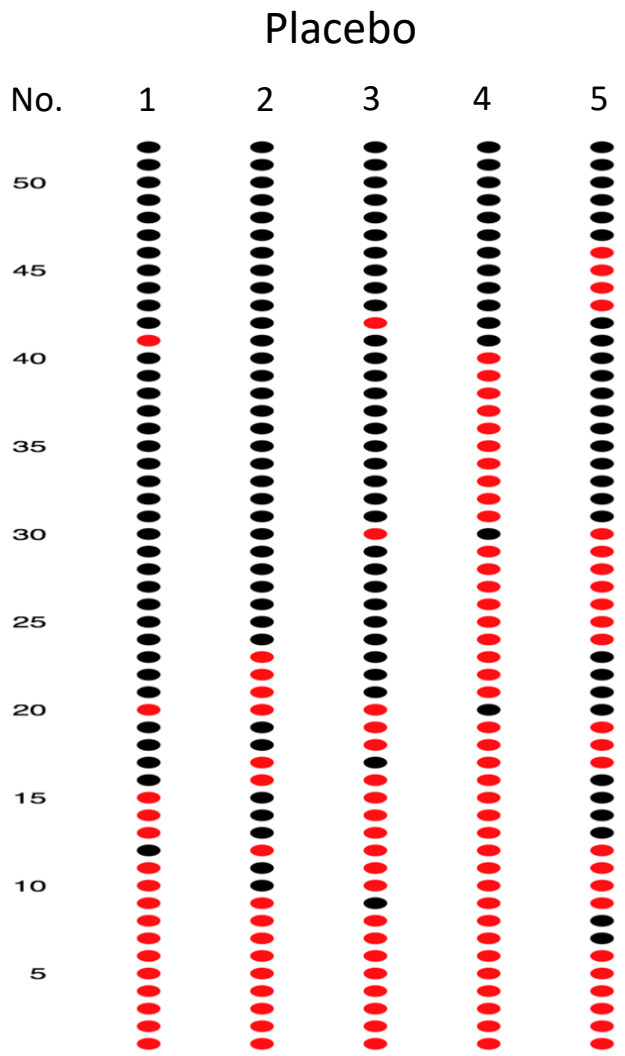
Zack McCaw (Google) and L. J. Wei (Harvard U)

Multiple outcomes (harm and benefit)

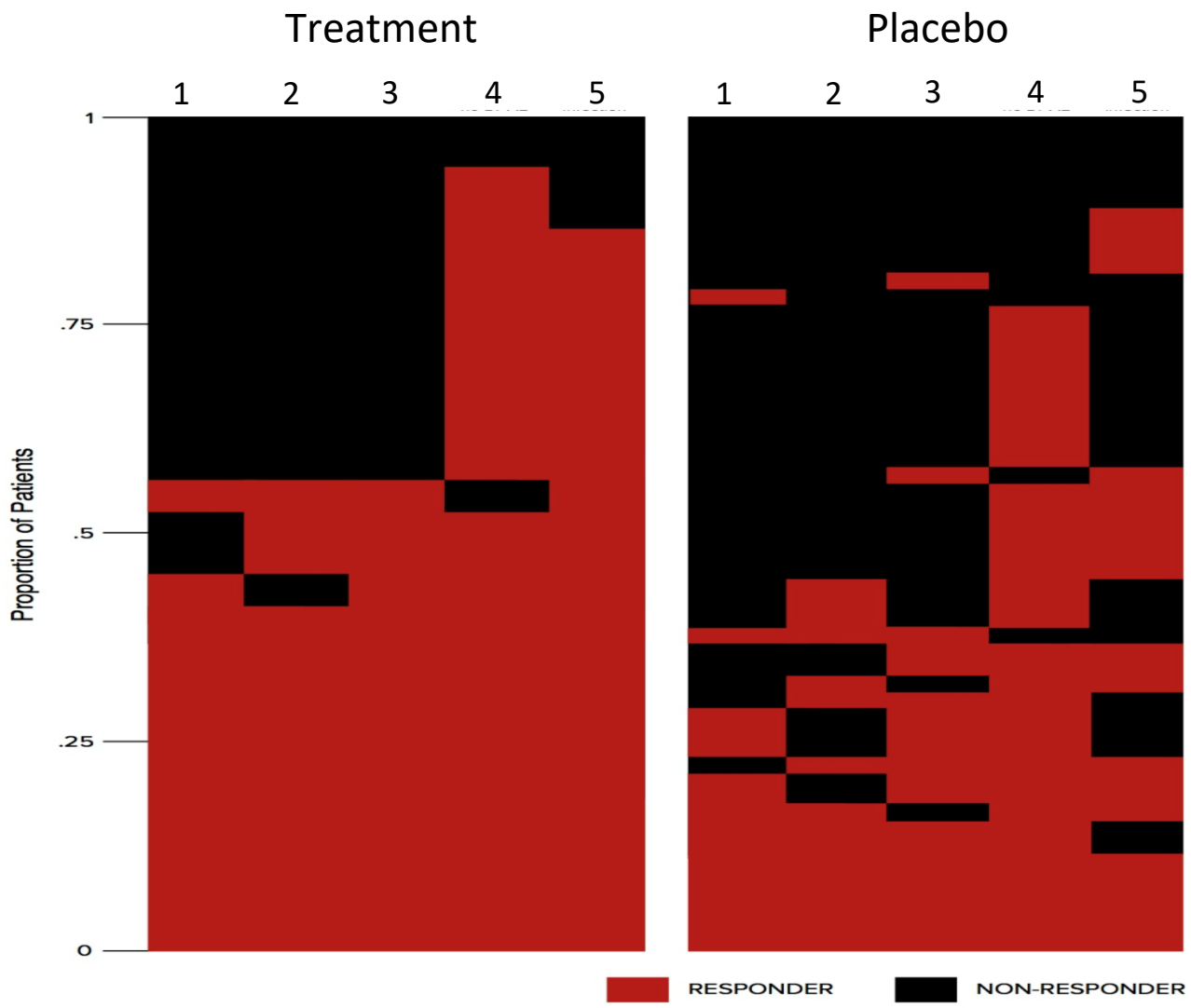
- For each outcome, define a clinically meaningful “response” (or complete response, partial response et al.)

Patient level multiple outcomes





Heat map

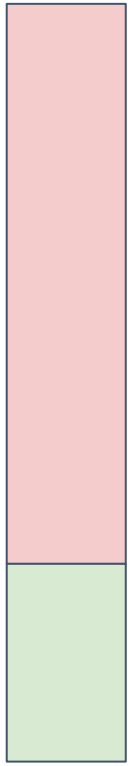

Treatment
Placebo
1
2
3
4
5
1
2
3
4
5
1
.75
.5
.25
0
Proportion of Patients
RESPONDER
NON-RESPONDER

How to utilize multiple outcomes

- What is current practice?
- Define primary endpoints and secondary endpoints
- Define efficacy and toxicity endpoint separately
- Analyzing those endpoints separately with respect to the treatment differences (effects)
- The conventional component-specific analysis – informative missing, censoring or competing risks
- No idea how to interpret those treatment effects at patient level

How about using patient level multiple outcomes

Most COVID-19 Trials Use an Ordinal Categorical Outcome



1. Deceased.
2. Hospitalized, requiring invasive mechanical ventilation.
3. Hospitalized, requiring high-flow oxygen.
4. Hospitalized, requiring low-flow oxygen.
5. Hospitalized, not requiring oxygen but attentive care.
6. Hospitalized, not requiring attentive care.
7. Not hospitalized.

Comparative COVID-19 Clinical Trials

- **Gilead Remdesivir Study in *NEJM***
  - **Treatment:** 5-day vs. 10-day Remdesivir.
  - **Primary parameter:** Odds ratio on Day 14 from ordinal logistic regression.

- **NIH Adaptive COVID Treatment Trial (ACTT-1) in *NEJM***
  - **Treatment:** Remdesivir vs. placebo.
  - **Primary outcome:** Time to recovery/improvement based on the ordinal outcome.

Gilead Remdesivir Clinical Status Data

	Category											
	≤ 1	> 1	≤ 2	> 2	≤ 3	> 3	≤ 4	> 4	≤ 5	> 5	≤ 6	> 6
10-day	21	176	54	143	64	133	78	119	91	106	94	103
5-day	16	184	32	168	41	159	60	140	71	129	80	120
Odds Ratio 10-day v. 5-day	0.73		0.54		0.54		0.65		0.64		0.73	

Gilead Remdesivir Analysis in *NEJM*

- Even if the model is plausible, a common odds ratio is difficult to interpret without a corresponding "event" probability. Moreover, an average of odds ratios is no long an odds ratio (since odds ratio is not a probability measure).

- Since the proportional odds assumption was not met, a Wilcoxon test (P=0.14) was performed, but no corresponding estimate of treatment efficacy was reported.

- **Lesson we learn**: the prespecified analysis should be interpretable and model-free.

- How do we get a summary for measuring the size of the treatment difference via the ordinal categorical outcome? Pr (A > B) ? Should we assign a numerical value for each category?

ACTT-1 Study in *NEJM*

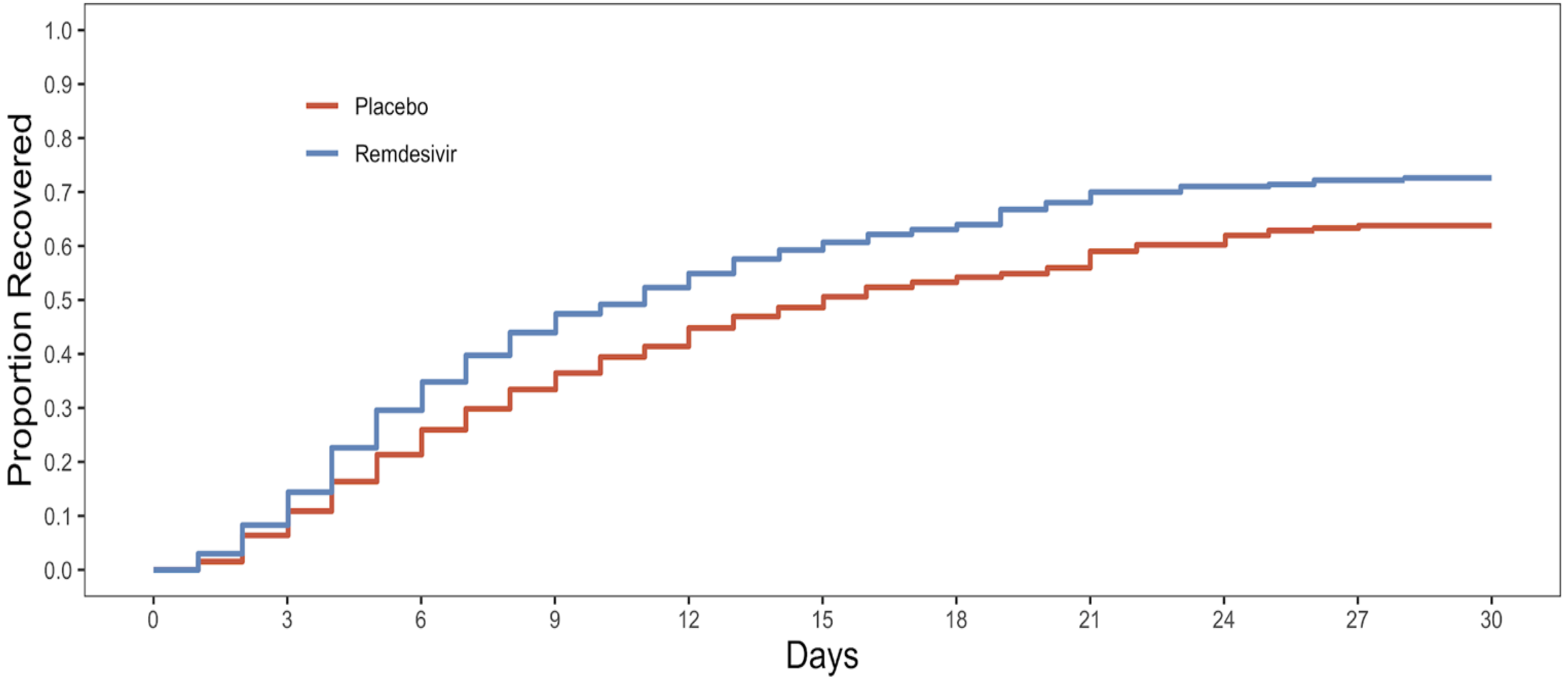
- Compared Remdesivir with Placebo among patients hospitalized with confirmed COVID-19.

- Primary endpoint was time to recovery across 28 days. However, some patients died before recovery. Thus, the time to recovery was competing with time to death.

- Prespecified a hazard ratio analysis, which is difficult to interpret (Fine and Gray). The HR was 1.32, which does not mean patients receiving Remdesivir were 32% more likely to recover.

- Need a clinically interetable summary to quantify the treatment effect.

ACTT Cumulative Recovery Curves (for survivors)



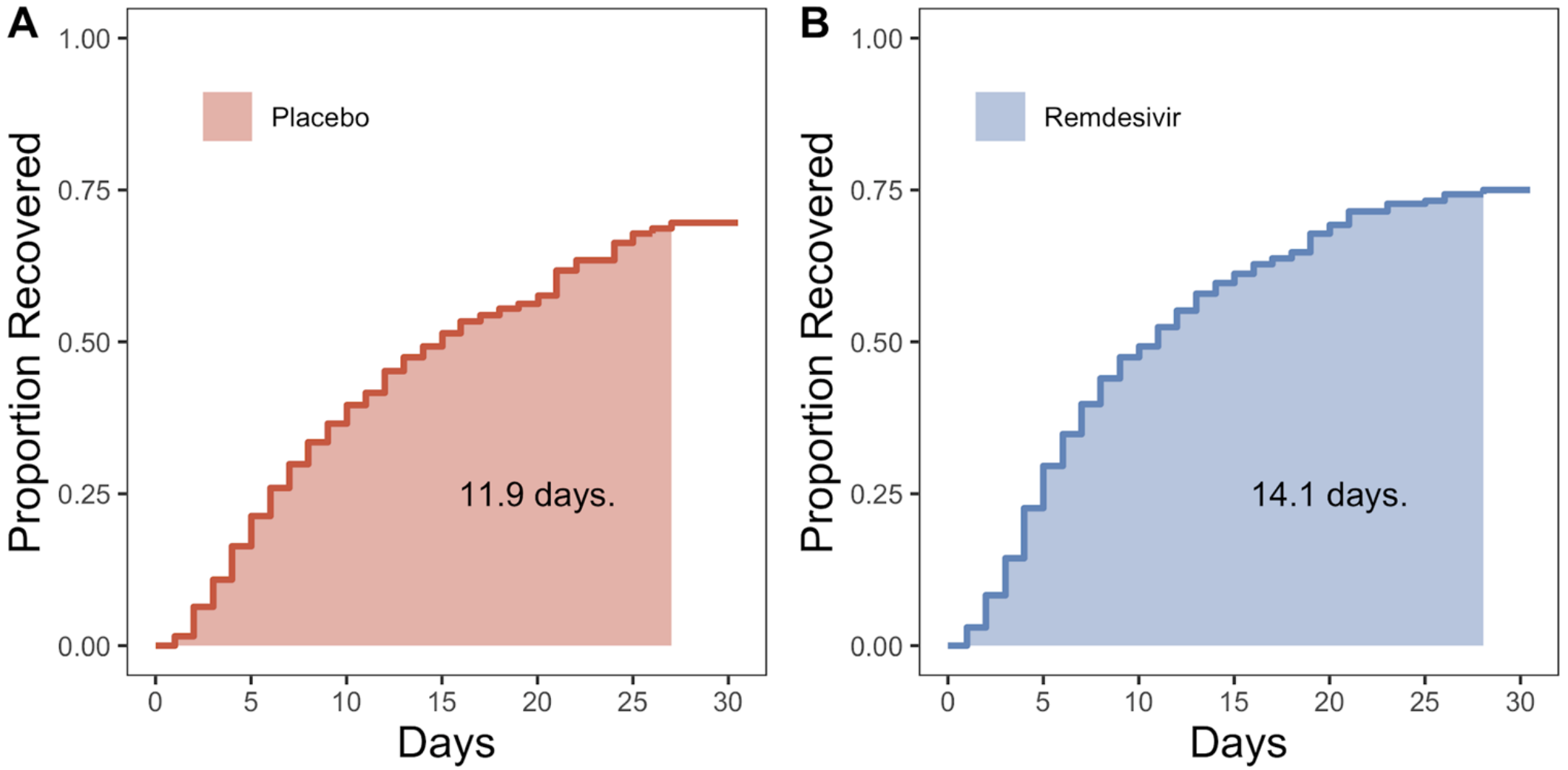
ACTT-1 Analysis

- A model-free and interpretable alternative is the area under the cumulative incidence curve. This is the mean time span post recovery within the 28 days of follow-up.

- For example, a patient who recovered on Day 15 had a post recovery time span was 28 – 15 = 13 days; the longer, the better.

- Graphically, the average post recovery time span is the area under the cumulative recovery curve.

Area Under the Cumulative Incidence Curve



ACTT-1 Reanalysis in *Annals of Internal Medicine*

- AUCs were 14.1 days for Remdesivir, 11.9 days for Placebo. The difference of 2.2 days (95% CI, 0.89 to 3.52, P<0.001) significantly favored Remdesivir.

- This time scale summary of treatment efficacy is easier to interpret than a 32% increase in the “hazard” of recovery.

- The mean time to recovery is not well-defined since we don’t know the time of recovery for a patient died during the hospitalization.

Discussions

- If the prespecified analysis is model based, but the model does not fit the data, the results can be difficult to interpret.

- We need model-free and clinically interpretable summaries for the treatment difference.

Cardiovascular disease clinical study in *NEJM*

COMPASS Trial

- The COMPASS trial compared combination rivaroxaban/aspirin with aspirin alone among patients with chronic coronary disease and/or peripheral artery disease.

- The primary analysis demonstrated that rivaroxaban/aspirin reduced the risk of cardiovascular events, albeit at an increased risk for major bleeding.

Efficacy vs. Safety

- Conventionally, trialists perform separate efficacy and safety analyses, resulting in separate summaries of benefit and risk.

- However, this approach is suboptimal as we do not know whether the efficacy and safety events occurred within the same patients.

- Moreover, conducting separate efficacy and safety analyses does not reflect clinical practice, in which clinicians must simultaneously weigh risks and benefits.

Net Clinical Benefit in *Circulation*

- In a prespecified analysis, Steffel *et al* compared rivaroxaban/aspirin and aspirin alone with respect to a net clinical benefit (NCB) endpoint, which combined efficacy and safety outcomes at the individual patient level.
  - These were CV-death, stroke, and MI for efficacy, and fatal bleeding or bleeding into a critical organ for safety.

- The HR was 0.80 (95% CI, 0.70-0.91; P=0.0005), demonstrating superiority of rivaroxaban/aspirin with respect to efficacy and safety. The 36-month NCB event-free times were 34.6 and 34.3 months, with a difference of 10.5 (95% CI, 4.9-16.2; P=0.0003) days in favor of rivaroxaban/aspirin.

THALES Trial in *NEJM*

- The THALES trial compared combination ticagrelor/aspirin with aspirin alone among patients with acute noncardioembolic cerebral ischemia.

- The primary outcome was time to stroke or death, whichever occurred first, across 30 days of follow-up.

Efficacy vs. Safety

- For time to stroke/death, the HR was 0.83 (95% CI, 0.71-0.96; P=0.02), suggesting superiority for ticagrelor/aspirin.

- However, severe bleeding was substantially more common with ticagrelor/aspirin (HR: 3.99; 95% CI, 1.7-9.1; P=0.0001).

- It is difficult to tell from these separate summaries whether there was a net clinical benefit with ticagrelor/aspirin.

Net Clinical Benefit

- By constructing a net clinical benefit measure of time to stroke, death, or severe bleeding, the event rates were 6.0% for ticagrelor/aspirin and 6.7% for aspirin alone.

- The difference of 0.7% (95% CI, -0.2% to 1.6%; P=0.12) did not provide significant evidence of benefit from ticagrelor/aspirin.

Need more novel analytic methods

- Ordinal categorical endpoint with patient level multiple outcomes is clinically interpretable
- However, it is not clear how to estimate the size of the treatment difference (beyond testing hypothesis)
- One may use pr(A better than B) as a metric, which is not clinically interesting entirely
- One usually ends up with a binary endpoint from this ordinal outcome, which is easier to interpret (this may lack of statistical power)
- How can we utilize the ordinal outcome for making inference about the treatment effect?