Penn Medicine

Liver Forum 18

Stretch Exercise in Statistics: Beyond binary outcomes

Michael Harhay, PhD
Assistant Professor of Epidemiology, Medicine, and Statistics and Data Science
Perelman School of Medicine

April 5, 2025

Funding & Disclosures

- Funding
  - NHLBI & PCORI

- Personal fees
  - DSMB for several academic (e.g., NIH and DOD sponsored) trials – nothing private.
  - American Thoracic Society
    - Deputy Editor, *American Journal of Respiratory and Critical Care Medicine*
  - Elsevier
    - Statistical review, *The Lancet Family Journals*

Key take-homes

1. Measuring benefit in a trial doesn't have to be limited to a single outcome.

2. However, while alternative, multi-dimensional or composite outcomes do offer new ways of examining patient outcomes both during and after the trial, they also have their own tradeoffs.

3. Integrating new perspectives from statistics and stakeholders could open a new way of thinking about efficacy! This is your space!

Why consider new endpoints?

- **Composite approaches have a place**: Many successful trials use composite endpoints that combine hard outcomes with ordinal measures of disease severity or improvement.

- **Missed clinical insight is unfortunate**: Most liver disease trials still rely primarily on binary outcomes or time-to-event analyses, potentially missing treatment effects on disease trajectory.

- **Patient-reported outcomes (PROs) are informative**: Inclusion of ordinal PRO measures (like in the POISE trial) captures patient experience in a nuanced way (used PBC-40 questionnaire).

- **>Statistical efficiency for same sample size**: Trials using ordinal outcomes or methods like the win ratio generally have better statistical power than those using purely binary outcomes.

What is the goal of clinical research and trials? The basics

- "Clinical trials are studies intended to discover or verify the effects of one or more investigational medicines."
  - EMA

- "A clinical trial's "endpoints" are measurements of what happens to people in the trial. When a trial is intended to evaluate the efficacy and safety of a new medical product or a new use of an approved product, its endpoints usually measure benefit. Investigators typically use either clinical or surrogate endpoints."
  - FDA

- "Data from the trial must be interpretable with a discernable treatment effect for drug approval."
  - FDA

- "Treatment Effect: An effect attributed to a treatment in a clinical trial. In most clinical trials the treatment effect of interest is a comparison (or contrast) of two or more treatments [on a prespecified endpoint or outcomes measures].
  - EMA: ICH Topic E 9 Statistical Principles for Clinical Trials

Clinical endpoints (from a regulatory vantage)

- Endpoints measure how a patient feels, functions, or survives.
  - Immediately, a dichotomy emerges for serious illnesses, as the green precedes, & may be a surrogate for death.

- Endpoints are meant to represent or characterize a clinical outcome of interest.
  - Objective: survival, disease exacerbation, clinical event (e.g. MI, stroke), etc.
  - Subjective: symptom score, “health-related quality of life” (validated instrument), etc.

- Changes in them should establish a therapeutic clinical benefit that is meaningful to the patient in the context of a given disease.

- Customarily, the basis for approval of new drugs relies on showing an intervention impacts an endpoint.

Note: The term “clinical endpoints” is used here to distinguish from “surrogate endpoints”. This term is used in the FDA’s 2014 Guidance document on Expedited Programs for Serious Conditions. Others may refer to these as “direct”, “true” or “clinically meaningful” endpoints.

Some logic and limitations of the common approach: (1) outcome

- Often, one outcome is targeted as the primary endpoint.
  - Reasons in support of this approach
    - Power calculation can be well informed by past and/or pilot data.
    - Can (maybe, but usually) come to an agreement on an important target outcome, or at least settle on an outcome that, if moved, would be compelling to various stakeholders.
    - Can be straightforward to capture a single outcome in a standardized manner.
    - Cost-effective or logistical reasons to target a specific data capture.

  - Reasons against
    - >1 outcome is important, available, or relevant.
    - Showing a change on a single measure is incomplete in isolation – what is the “effect”?
    - Showing a change on a combined outcome measure is difficult.
      - *small effects, long durations of follow-up required to see change, and/or rare outcomes*.
    - Low statistical power due to these and related reasons. Poorly designed, or **designed-to-fail trials** are unacceptable.
      - Ethically, trials should deliver for those who enrolled and society more broadly.

Empirical and conceptual challenges

- These are summary values. Resolution is not optimized.

Commentary

January 20, 2010

Composite End Points in Randomized Trials
There Is No Free Lunch

George Tomlinson, PhD; Allan S. Detsky, MD, PhD

» Author Affiliations

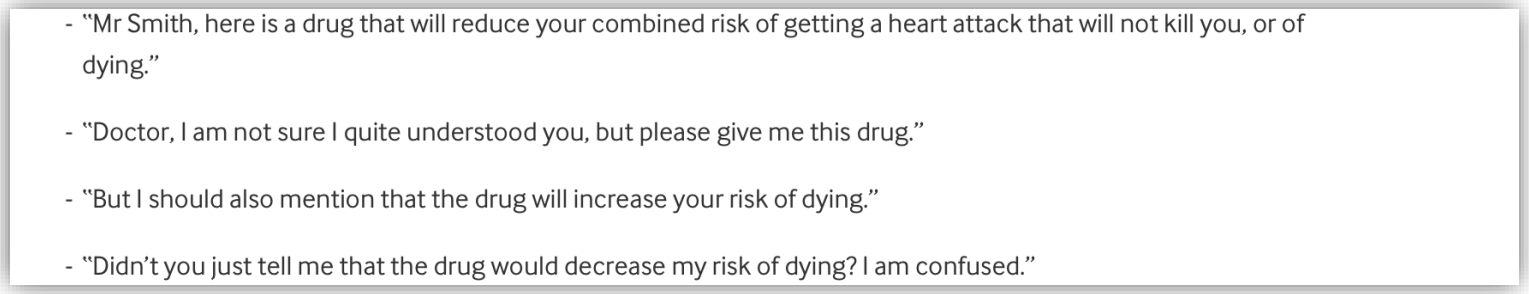
JAMA. 2010;303(3):267-268. doi:10.1001/jama.2009.2017

Non-hierarchical composite endpoints

- In the simplest rendition: a single measure of effect, based on a combination of individual endpoints.
- Statistical power and effect estimate precision increases – more for same concept from before.
- Particularly useful for drugs that can benefit patients in several ways or if component events are infrequent.
- Examples:
  - "MACE" (major adverse cardiac events): cardiovascular death, non-fatal MI, and non-fatal stroke
  - "Clinical Worsening:" may include a categorical decline in functioning, worsening symptoms, the addition of a new medication, hospitalization due to the disease, death, etc.
  - Death or liver transplant
- Often analyzed as a (1) time to the first event or (2) occurrence of any event.

Non-hierarchical composite endpoints: *simplest but hardest*?

- Require a lot of considerations:
  - Each component should itself be clinically meaningful.
  - Ideally, each component would be *approximately* equally meaningful.
    - Including death makes it really challenging right away. Death is an absorbing state, and others are not.
    - Death = rehospitalization = transplant?
  - The composite should not include individual components for which a treatment effect is not expected.
  - "Success" should not be concluded if driven by a less meaningful component.
  - May complicate communication of the established benefit of a drug.
    - There may often be inadequate evidence to establish a treatment effect on any of the components individually. (If a trial "wins" on MACE, can you conclude the drug improves survival?).
    - What happens if components go in a different direction? Raises regulatory concerns/questions (rightfully).



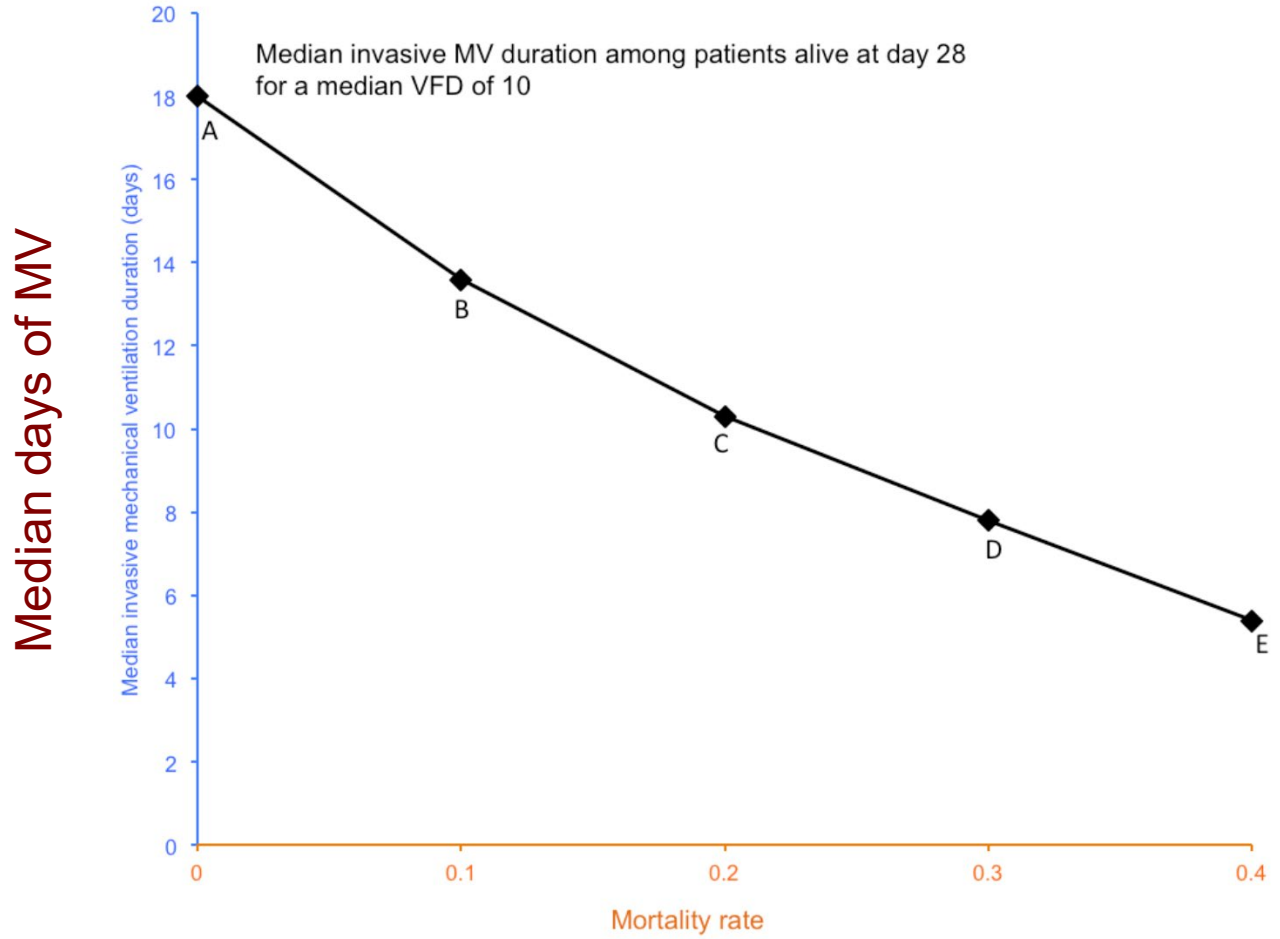
Cordoba G et al. Definition, reporting, and interpretation of composite outcomes in clinical trials: systematic review *BMJ* 2010; 341 :c3920 doi:10.1136/bmj.c3920

Ventilator-Free Day Outcomes Can Be Misleading

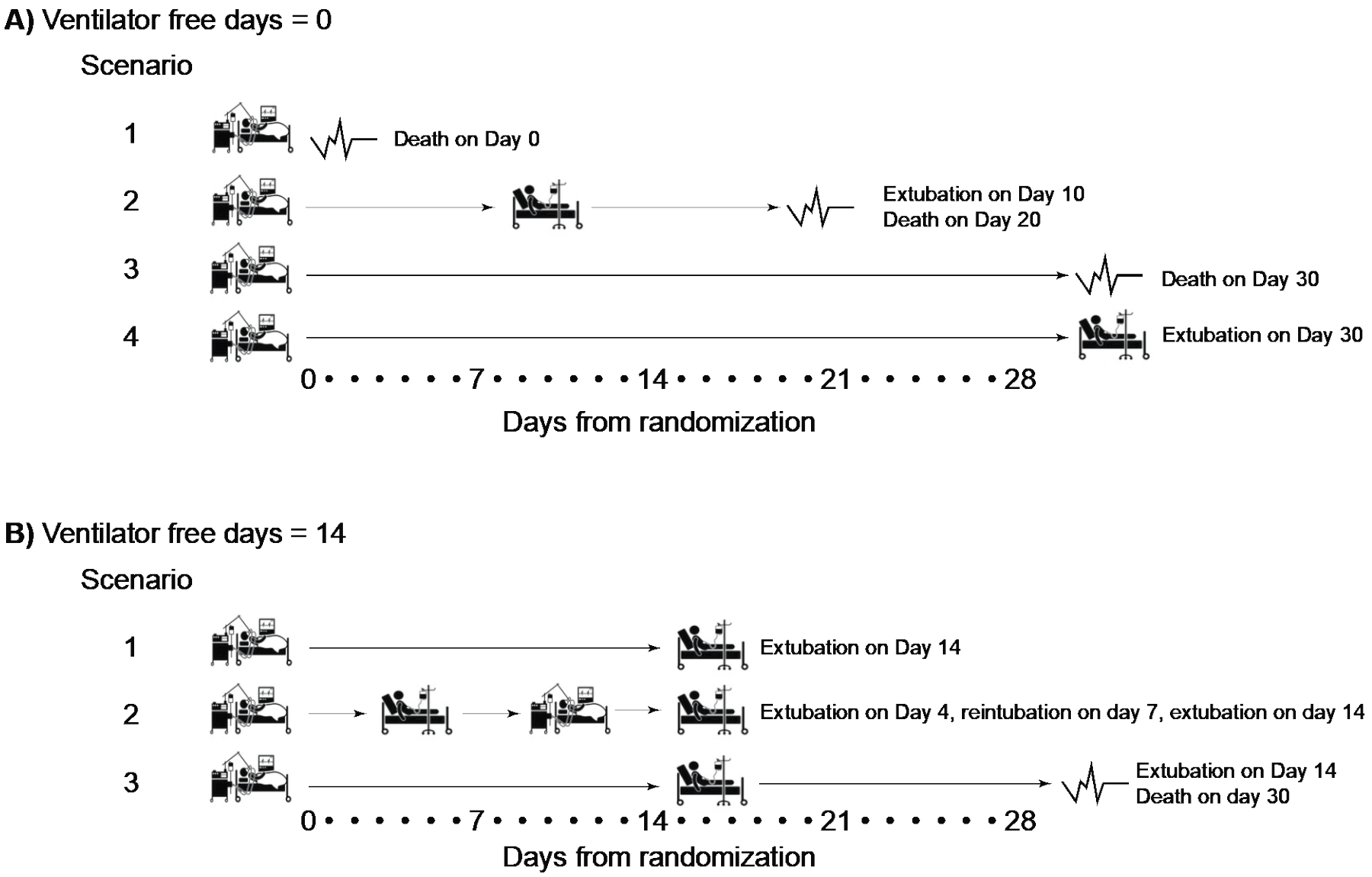
The same number of VFDs in different trials does not equal:

1. Same mortality,
2. Same healthcare utilization and public health impact, nor
3. Caregiver burden.

Many composite endpoints solve a statistical problem with the trade-off of a conceptual problem.



Bodet-Contentin et al, *Critical Care Medicine*, 2018

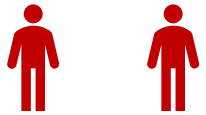


Yehya, Harhay et al. Re-appraisal of Ventilator-free Days in Critical Care Research. *AJRCCM,* 2019

Hierarchical composite endpoints

Win ratio: compares all trial participants in one arm to all trial participants in the other trial arm, and assigns each participant's outcome a win or not.

Arm A | Arm B

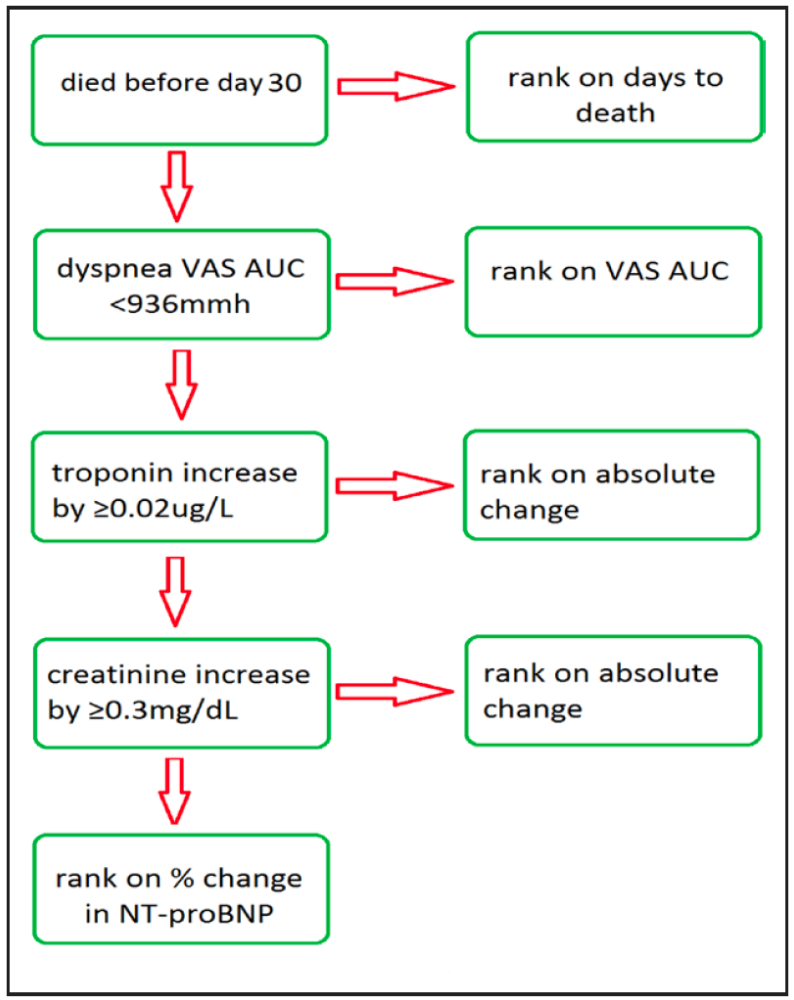
Neither survives hospital stay:  **= both get 0 points**

Both patients live:  **= first discharged in blue gets 1, other 0**

Only one survives:  **= surviving blue patient gets 1, other 0**

The trial arm with the most  (i.e., wins) across all possible comparisons is deemed to have responded better to the intervention.

Global rank score



Derivation of the global rank composite.

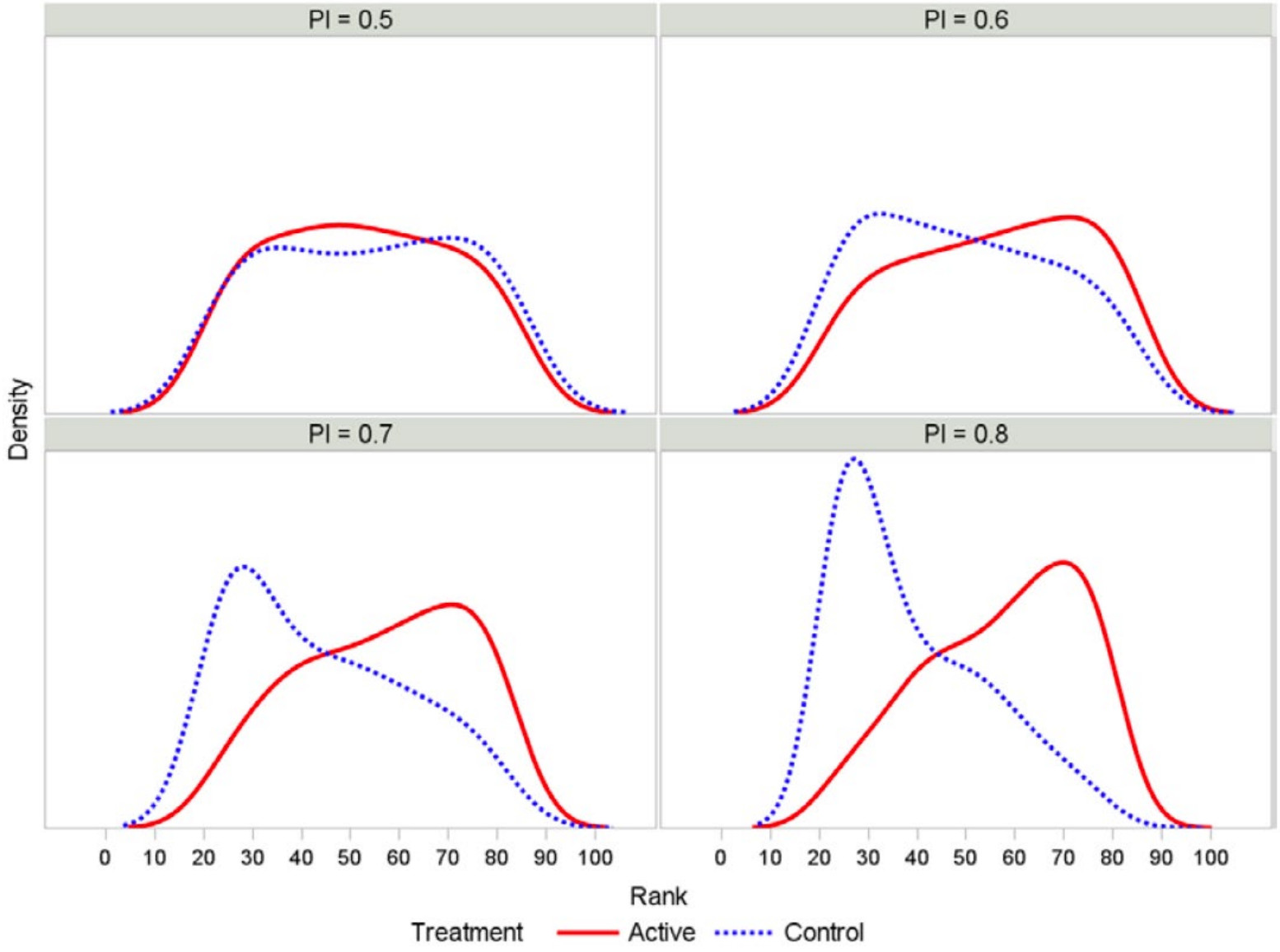
1. Each patient is assigned a rank from 1 to n according to their response on several outcomes.

2. These outcomes are arranged in a hierarchy to ensure:
    1. rank of 1 is assigned to the patient with the most adverse response (an early death).
    2. A higher rank of n is assigned to the patient with the most favorable response on any of the criteria.

3. There may be tied ranks if patients' outcomes are equally favorable.

4. The analysis then compares these ranked values between the treatment groups → **Probability of a beneficial effect**.

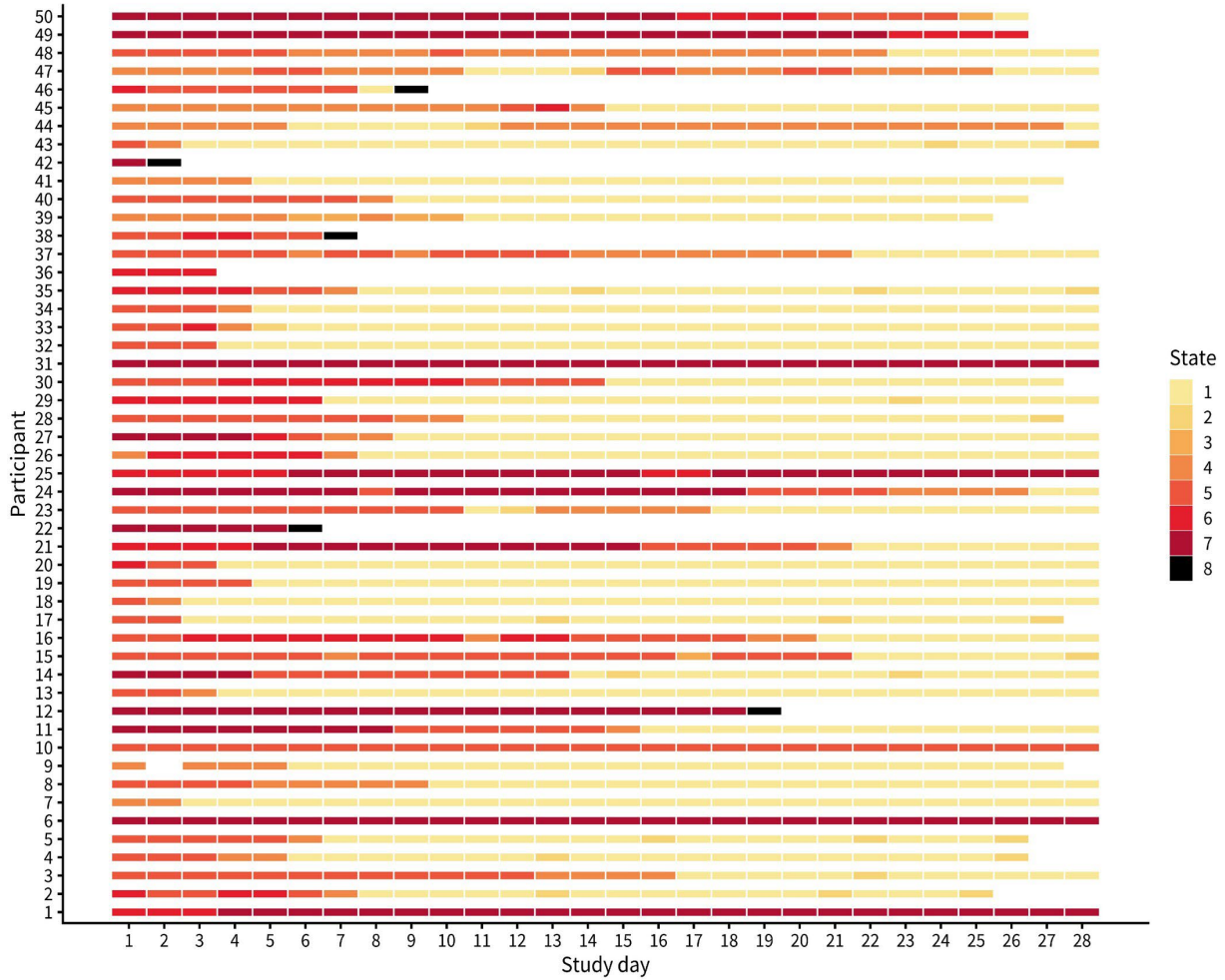
Brown and Ezekowitz, Circ Heart Fail. 2017



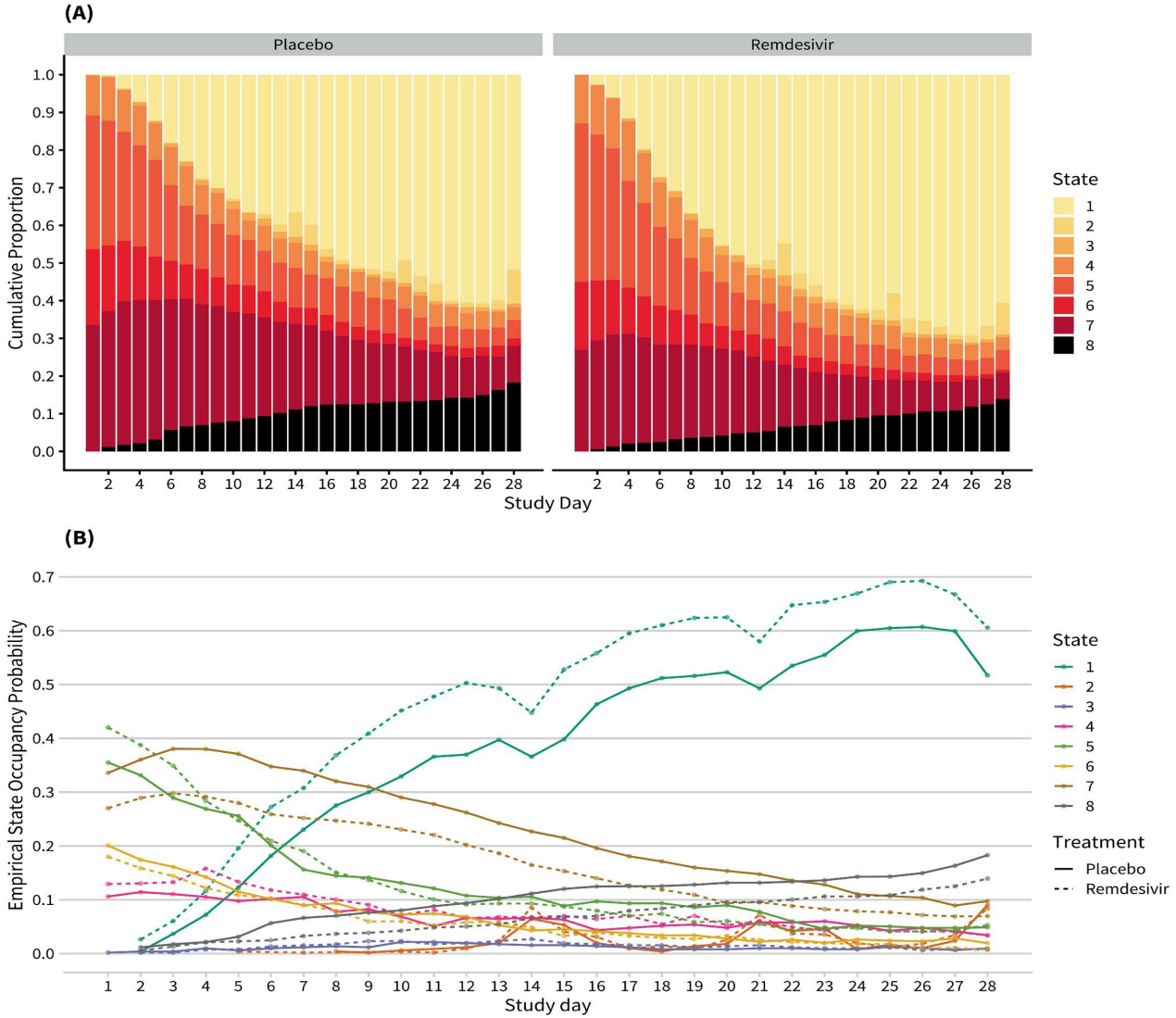
Brown and Ezekowitz, Circ Heart Fail. 2017

Adding the dimension of time to hierarchical outcomes

Ordinal state	Description
State 1	Not hospitalized and no limitations of activities
State 2	Not hospitalized, with limitation of activities, home oxygen requirement, or both
State 3	Hospitalized, not requiring supplemental oxygen and no longer requiring ongoing medical care
State 4	Hospitalized, not requiring supplemental oxygen but requiring ongoing medical care related to Covid-19 or to other medical conditions
State 5	Hospitalized, requiring any supplemental oxygen
State 6	Hospitalized, requiring noninvasive ventilation or use of high-flow oxygen devices
State 7	Hospitalized, receiving invasive mechanical ventilation or extracorporeal membrane oxygenation (ECMO)
State 8	Death

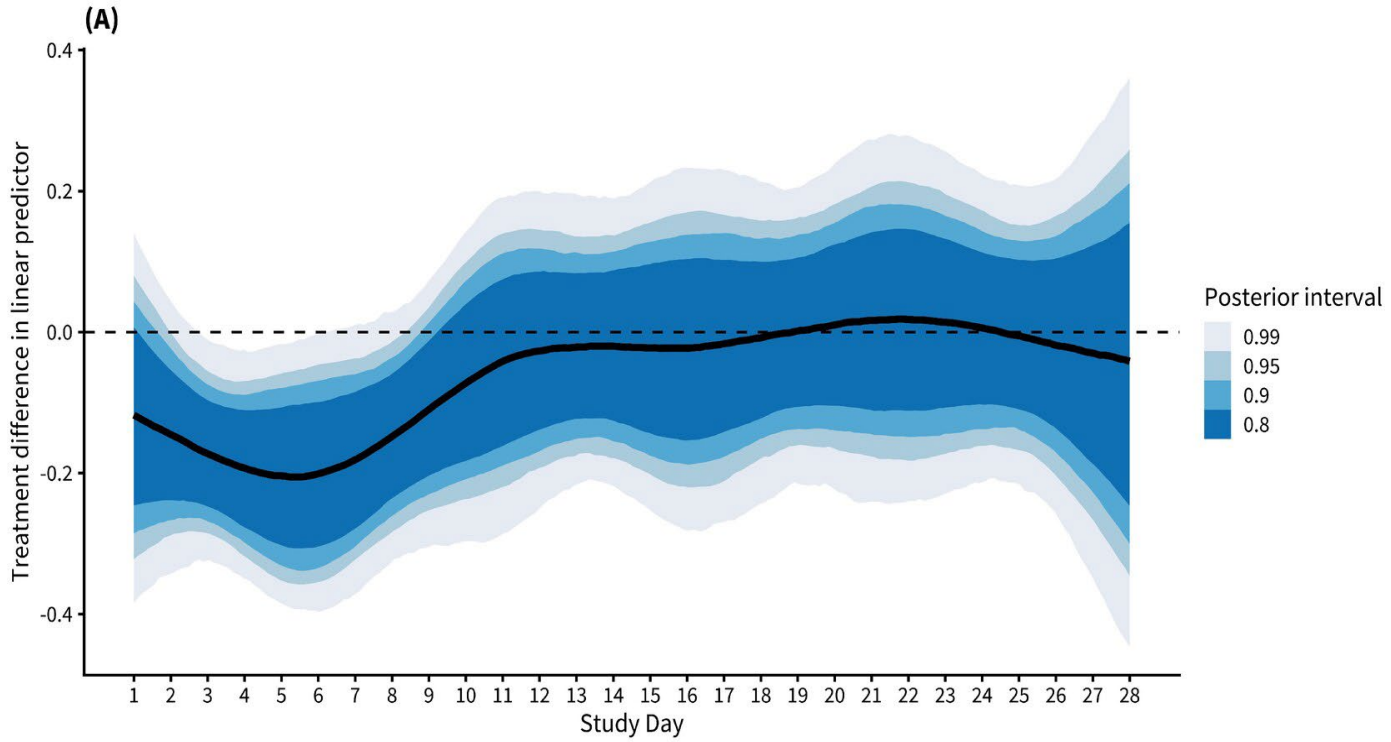


Rohde MD, French B, Stewart TG, Harrell FE Jr. Bayesian transition models for ordinal longitudinal outcomes. Stat Med. 2024 Aug 15;43(18):3539-3561

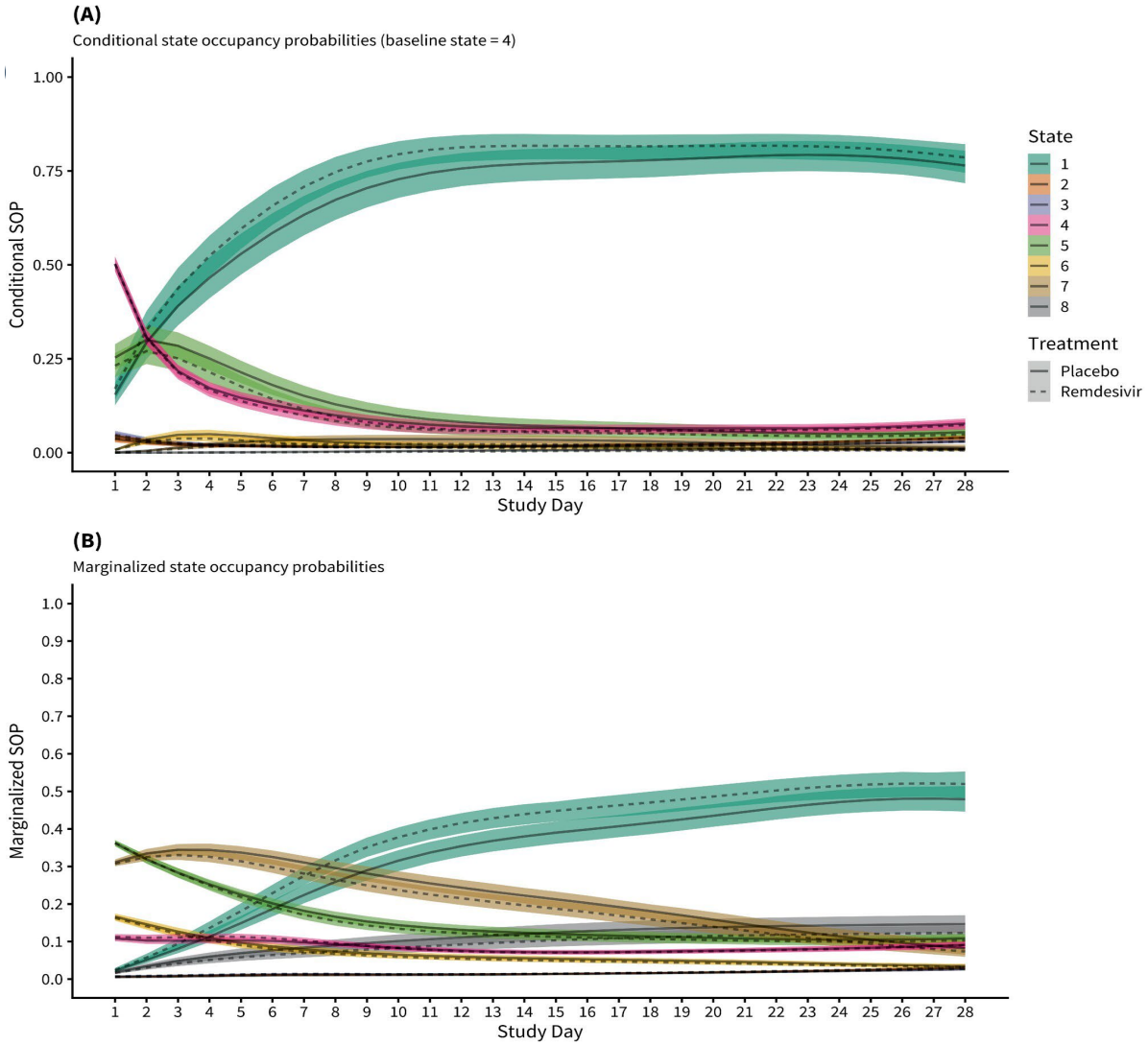


Rohde MD, French B, Stewart TG, Harrell FE Jr. Bayesian transition models for ordinal longitudinal outcomes. Stat Med. 2024 Aug 15;43(18):3539-3561

Ordinal models – some visuals – time scale is of course adaptable



Ordinal mod


(A)
Conditional state occupancy probabilities (baseline state = 4)
Conditional SOP
1.00
0.75
0.50
0.25
0.00
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
Study Day
State
1
2
3
4
5
6
7
8
Treatment
Placebo
Remdesivir
(B)
Marginalized state occupancy probabilities
Marginalized SOP
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
Study Day

Act 2: Some ideas for liver disease trials.

Some examples of composite and ordinal outcomes for liver disease

REGENERATE Trial

- Time to first occurrence of any of the following adjudicated events:
  1. Death (all cause);
  2. MELD score ≥ 15;
  3. Liver transplantation;
  4. Hospitalisation for onset of variceal bleed, hepatic encephalopathy, spontaneous bacterial peritonitis;
  5. Ascites secondary to cirrhosis and requiring medical intervention (e.g. diuretics or paracentesis);
  6. Histological progression to cirrhosis

- HALT-C Trial
  - Used ordinal Ishak fibrosis staging (0-6) to assess progression or regression of fibrosis over time

- STELLAR-3 and STELLAR-4
  - Used a dichotomized endpoint based on an underlying ordinal scale - improvement in fibrosis ≥1 stage without worsening of NASH

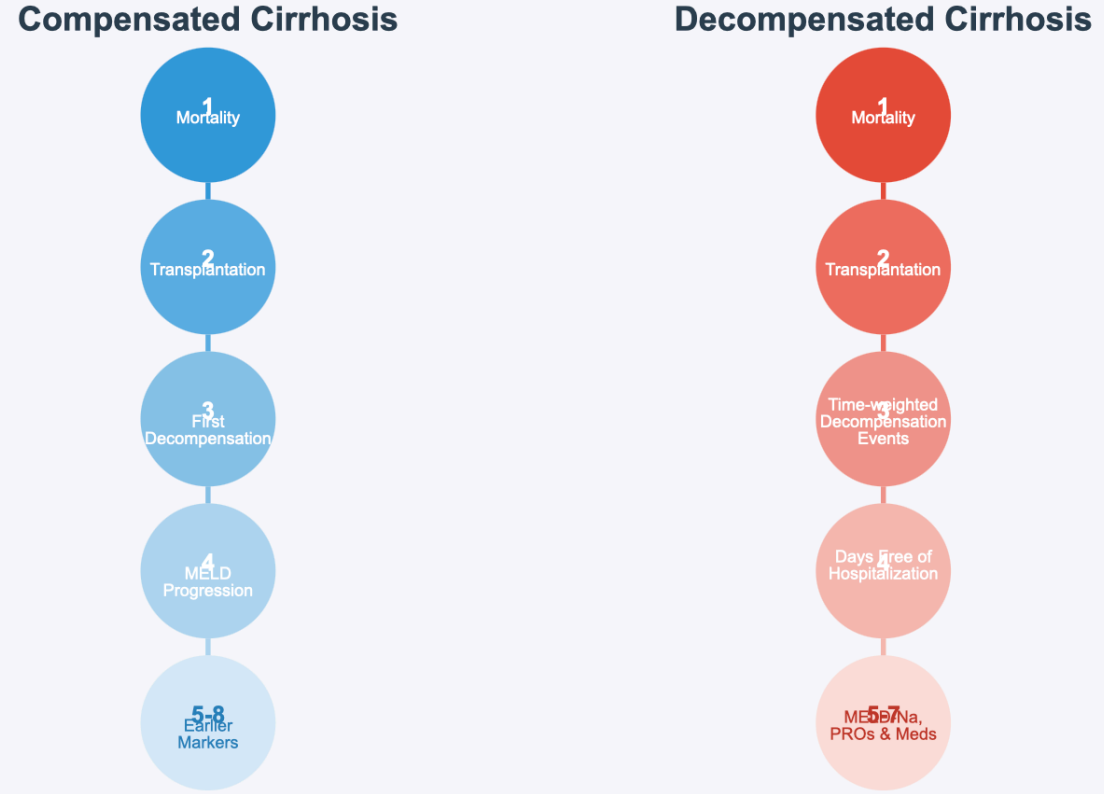
Composite Outcomes in Liver Disease Clinical Trials

Trial	Citation	Composite Outcome Used
NASH/NAFLD Trials		
REGENERATE Obeticholic acid for NASH	Younossi ZM, et al. Lancet. 2019;394(10215):2184-2196	• Fibrosis improvement ≥1 stage with no worsening of NASH • NASH resolution with no worsening of fibrosis
PIVENS Pioglitazone/vitamin E for NASH	Sanyal AJ, et al. N Engl J Med. 2010;362(18):1675-1685	• Improvement in NAS ≥2 points without worsening of fibrosis
FLINT Obeticholic acid for NASH	Neuschwander-Tetri BA, et al. Lancet. 2015;385(9972):956-965	• Improvement in NAS ≥2 points without worsening of fibrosis • NASH resolution without worsening fibrosis (secondary)
Cirrhosis/Portal Hypertension Trials		
PREDESCI Beta-blockers for portal hypertension	Villanueva C, et al. Lancet. 2019;393(10181):1597-1608	• Time to decompensation (ascites, bleeding, encephalopathy) • Ordinal grading of ascites and encephalopathy severity
ANSWER Albumin for decompensated cirrhosis	Caraceni P, et al. Lancet. 2018;391(10138):2417-2429	• Primary: 18-month mortality • Secondary: composite of paracentesis, refractory ascites, SBP, bacterial infection, renal dysfunction, encephalopathy, or liver-related hospital admissions
DASIMAR Simvastatin for portal hypertension	Abraldes JG, et al. Gastroenterology. 2016;150(5):1043-1053	• Time to liver transplantation, variceal bleeding, or death • Development of other complications of portal hypertension
Viral Hepatitis Trials		
HALT-C Peginterferon for chronic hepatitis C	Di Bisceglie AM, et al. N Engl J Med. 2008;359(23):2429-2441	• Ordinal Ishak fibrosis staging (0-6) to assess progression or regression of fibrosis over time
GLOBE Telbivudine vs lamivudine for HBV	Lai CL, et al. Gastroenterology. 2007;132(6):2164-2175	• HBV DNA suppression + ALT normalization + HBeAg seroconversion for HBeAg-positive patients • HBV DNA suppression + ALT normalization for HBeAg-negative
Other Liver Disease Trials		
POISE Obeticholic acid for PBC	Nevens F, et al. N Engl J Med. 2016;375(7):631-643	• ALP < 1.67× ULN + total bilirubin ≤ ULN + ALP decrease ≥15% • PBC-40 questionnaire for symptom assessment
NGM282 Phase 2 NGM282 for PSC	Hirschfield GM, et al. Hepatology. 2019;70(3):788-801	• ALP reduction + other liver biochemistry improvements + changes in Enhanced Liver Fibrosis (ELF) score + PROs
STOPAH Steroids/pentoxifylline for AH	Thursz MR, et al. N Engl J Med. 2015;372(17):1619-1628	• Primary: mortality • Secondary: composite of mortality + liver transplant
HCC and Transplant Trials		
RESORCE Regorafenib for HCC	Bruix J, et al. Lancet. 2017;389(10064):56-66	• Primary: Overall survival • Secondary: RECIST criteria (ordinal: CR, PR, SD, PD)

Claude content

Generated by Claude.ai – needs close check

Win Ratio Hierarchical Structures for Liver Disease



Win Ratio: Treatment wins ÷ Control wins based on hierarchical comparison

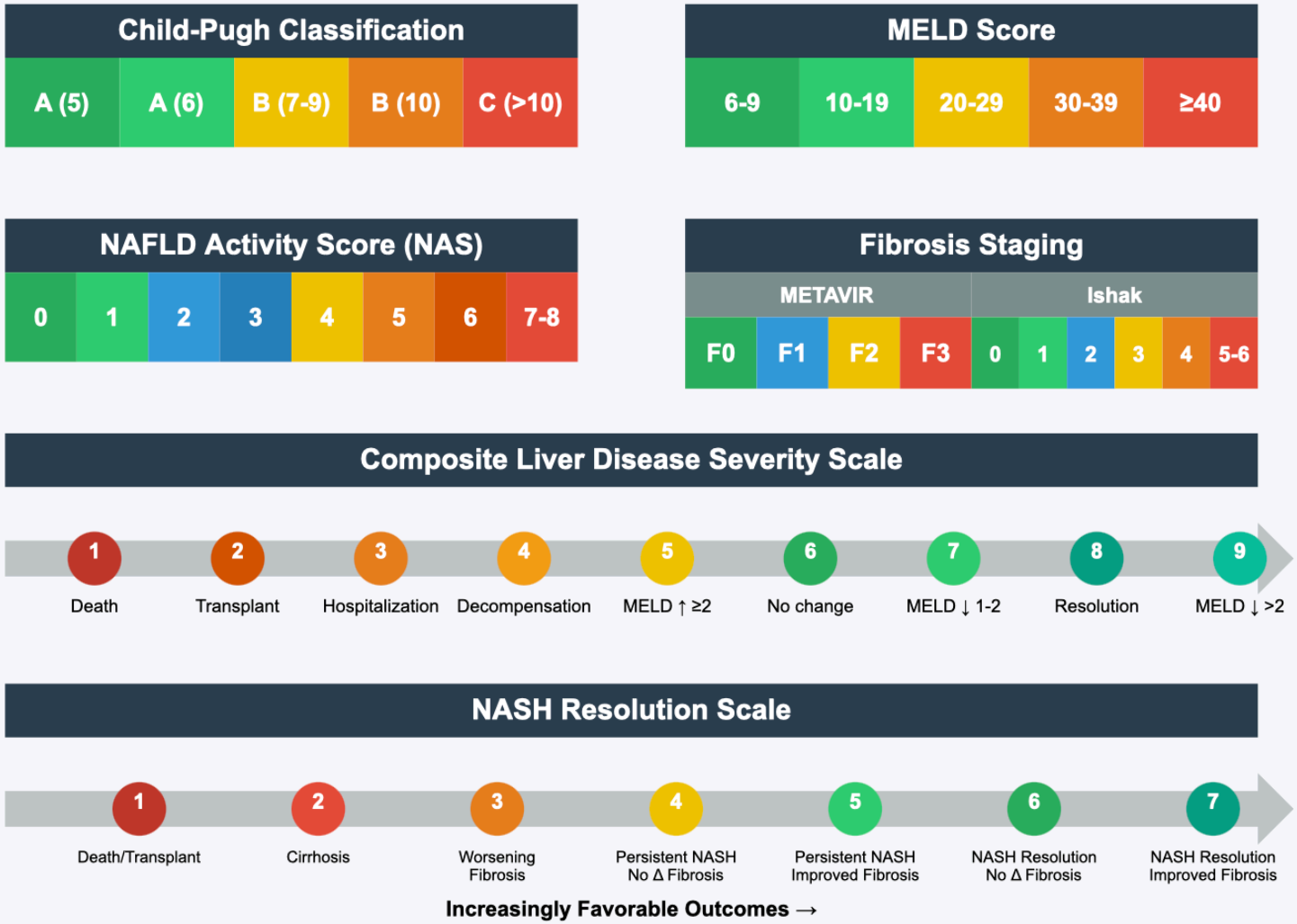
Higher levels in hierarchy take precedence in pairwise patient comparisons

Ordinal Outcomes in Liver Disease Clinical Trials

Trial	Citation	Ordinal Outcome Used
NASH/NAFLD Trials		
REGENERATE Obeticholic acid for NASH	Younossi ZM, et al. Lancet. 2019;394(10215):2184-2196	• Fibrosis improvement ≥1 stage with no worsening of NASH • NASH resolution with no worsening of fibrosis
STELLAR-3 and STELLAR-4 Selonsertib for NASH	Harrison SA, et al. N Engl J Med. 2023;388(16):1470-1484	• Improvement in fibrosis ≥1 stage without worsening of NASH
CENTAUR Cenicriviroc for NASH	Friedman SL, et al. Hepatology. 2018;67(5):1754-1767	• Improvement in NAS ≥2 points with no worsening of fibrosis • Ordinal changes in fibrosis stage
Cirrhosis/Portal Hypertension Trials		
PREDESCI Beta-blockers for portal hypertension	Villanueva C, et al. Lancet. 2019;393(10181):1597-1608	• Primary: Time to decompensation • Secondary: Ordinal grading of ascites and hepatic encephalopathy severity
CANONIC Study (ACLF CLIF-SOFA Validation)	Moreau R, et al. Gastroenterology. 2013;144(7):1426-1437	• Developed and validated ACLF grades 0-3 • CLIF-SOFA score (ordinal measure of organ dysfunction across 6 systems) • 28-day mortality prediction
ATTIRE Albumin for decompensated cirrhosis	China L, et al. JAMA. 2021;326(2):130-142	• Primary: Composite of infection, kidney dysfunction, or death • Secondary: Ordinal changes in CLIF-SOFA score
Other Liver Disease Trials		
HALT-C Peginterferon for chronic hepatitis C	Di Bisceglie AM, et al. N Engl J Med. 2008;359(23):2429-2441	• Ordinal Ishak fibrosis staging (0-6) to assess progression or regression of fibrosis over time
POISE Obeticholic acid for PBC	Nevens F, et al. N Engl J Med. 2016;375(7):631-643	• Binary response definition for primary endpoint • PBC-40 questionnaire (ordinal assessment of symptom severity across multiple domains)

Generated by Claude.ai – needs close check

Ordinal Outcome Scales for Liver Disease Trials



Generated by Claude.ai – needs close check

Proportional Odds Models

- **Proportional Odds Models**
  - **Approach**: Extension of logistic regression for ordinal data
  - **Analysis**: Estimates cumulative odds ratios for treatment effect
  - **Assumptions**: Proportional odds assumption (effect is consistent across all levels)
  - **Implementation**:
    - Model: $\text{logit}[P(Y \leq j)] = \alpha j - \beta X$, where j = outcome categories
  - **Interpretation**: Treatment effect as odds of being in a better category
  - **Software**: Available in standard statistical packages (R: MASS::polr(), SAS: PROC LOGISTIC)

- **2. Sliding Dichotomy Analysis**
  - **Approach**: Creates binary outcomes based on baseline risk
  - **Implementation**:
    - Stratify patients by baseline severity
    - Define "success" thresholds differently for each stratum
    - For severe patients: success = avoiding worst outcomes
    - For moderate patients: success = achieving moderate improvement
    - For mild patients: success = achieving complete resolution
  - **Advantages**: Accounts for heterogeneity in treatment effects
  - **Example**: In a cirrhosis trial, "success" for Child-Pugh C patients might be survival without transplant, while for Child-Pugh A it might be maintaining normal liver function tests

- **3. Win Ratio with Ordinal Component**
  - **Approach**: Incorporates ordinal scales within win ratio framework
  - **Implementation**:
    - Use conventional win ratio for definitive events (death, transplant)
    - For tied pairs, compare degree of change in ordinal scale
    - Calculate wins based on greater improvement in ordinal measure
  - **Example**: After comparing mortality, compare magnitude of MELD improvement

- **4. Global Rank Test**
  - **Approach**: Ranks all patients from worst to best outcome regardless of treatment
  - **Implementation**:
    - Assign ranks based on hard outcomes first (death, transplant)
    - For tied patients, use ordinal scale to assign further ranks
    - Compare mean or median ranks between treatment groups
  - **Advantages**: Non-parametric, no distributional assumptions

Hypothetical examples

- **Hypothetical NASH Trial**
  - **Ordinal Outcome**: 7-level scale from death to resolution with fibrosis improvement
  - **Primary Analysis**: Proportional odds model
  - **Result Interpretation**: "Treatment X resulted in 2.1 times higher odds of being in a better category on the NASH Resolution Scale compared to placebo (odds ratio 2.1, 95% CI 1.5-2.8, p<0.001)”

- **Hypothetical Cirrhosis Trial**
  - **Ordinal Outcome**: Composite incorporating survival, decompensation events, and MELD changes
  - **Primary Analysis**: Global rank test with sliding dichotomy
  - **Result Interpretation**: "Patients receiving Treatment Y had significantly better global ranks than those receiving placebo (p=0.003), with the most pronounced benefits observed in patients with baseline MELD scores >15"

Final ideas

- Consider developing a core outcome set.
- Engaging stakeholders – is what matters to clinicians the same as what matters to patients?
- Special considerations.
  - Handling liver transplantation as an endpoint
  - Accounting for MELD exceptions
  - Regional variations in transplant availability
  - Impact of alcohol use/abstinence
- Explore non-invasive biomarkers to avoid missing data (avoid liver biopsies).
- Assess techniques (MRE, VCTE, MRI-PDFValidateF) as primary endpoints.
- Validate symptom assessment tools for liver-specific manifestations.

Thanks!

mharhay@pennmedicine.upenn.edu

Commentary

January 20, 2010

Composite End Points in Randomized Trials

There Is No Free Lunch

George Tomlinson, PhD; Allan S. Detsky, MD, PhD

» Author Affiliations

JAMA. 2010;303(3):267-268. doi:10.1001/jama.2009.2017


Penn Medicine