UCSF, November 10, 2021

Evaluating benefit-risk, handling missing data and a universal sample size formula for clinical trials

Jeetu Ganju
Ganju Clinical Trials, LLC
jganju@yahoo.com

Ron Yu
Gilead Sciences
ron.yu@gilead.com

As the Title Says…

- Benefit-Risk evaluation
- Handling missing data
- Universal (non-parametric) sample size formula for parallel group trials

  (this one is joint with Lu Tian of Stanford)

Main Idea

- Organize endpoints by importance

  - Compare pairs of patient data on most important endpoint

    Possible outcomes: better (+1), worse (-1), tie (0) (Mann-Whitney U test)

  - Comparison is a tie? Compare next-most important endpoint

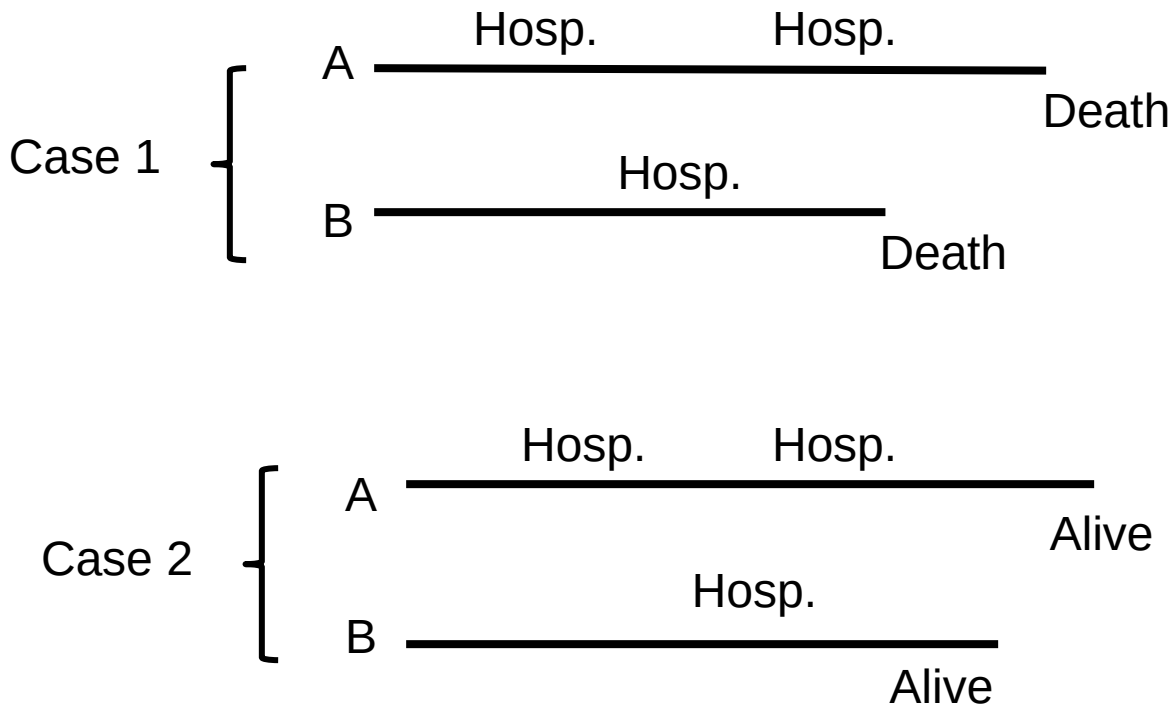
    Comparisons occur at minimum of follow-up times (Finkelstein-Schoenfeld)

- Calculate $\widehat{\boldsymbol{\theta}} = \frac{\widehat{\boldsymbol{Prob}}(A > B)}{\widehat{\boldsymbol{Prob}}(A < B)}$

  (“>” means better than, “<” worse than. A and B are randomly selected patients from Groups 1 and 2)

Examples

Endpoint 1: time to death. Endpoint 2: frequency of hospitalizations



Who did better?

✓

✓

Benefit-Risk

Current Approach

- Drug approval depends on benefit-risk evaluation
  - Benefit: efficacy endpoints, formally evaluated
  - Risk: subjective evaluation of adverse events

  Combined evaluation is subjective

Side note: term ‘benefit-risk’ is not symmetric

Benefit: gain that will be accrued

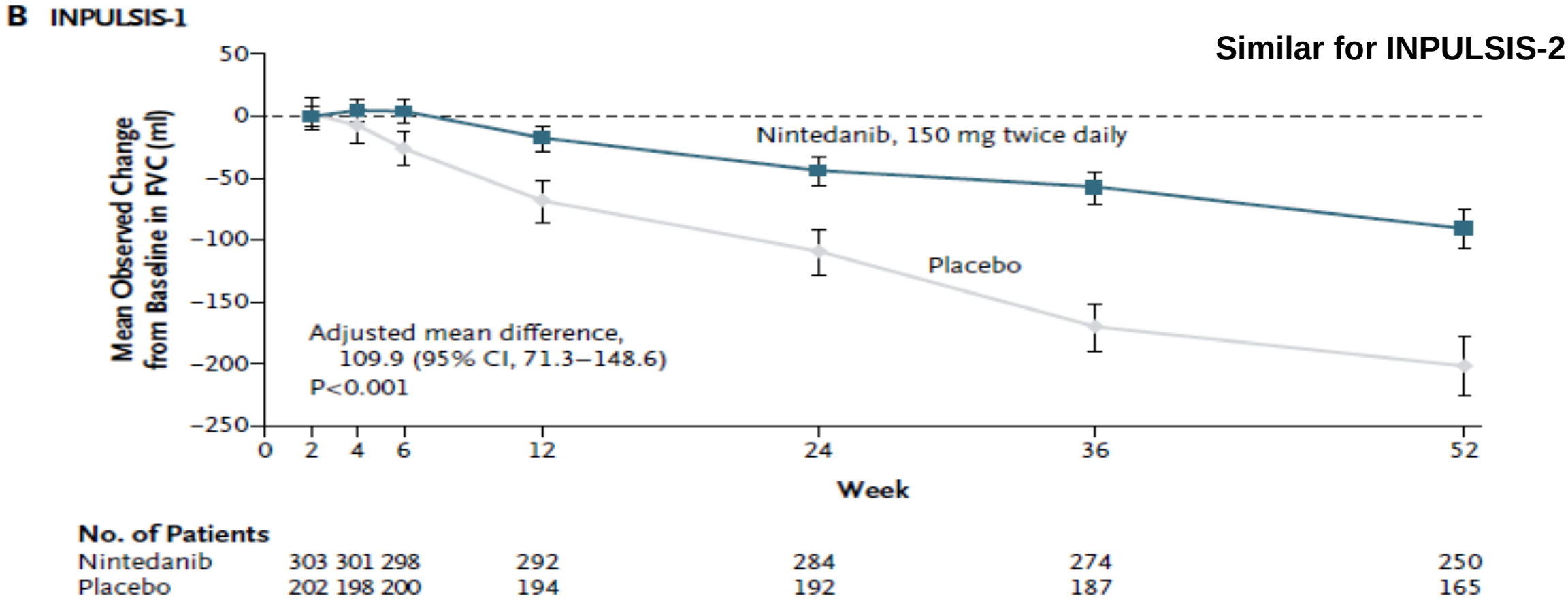
Risk: harm that may be experienced

Neutral term: benefit-harm

Example: Idiopathic Pulmonary Fibrosis (Ofev) 1/5

Primary efficacy endpoint: Annual rate of decline in forced vital capacity (FVC)

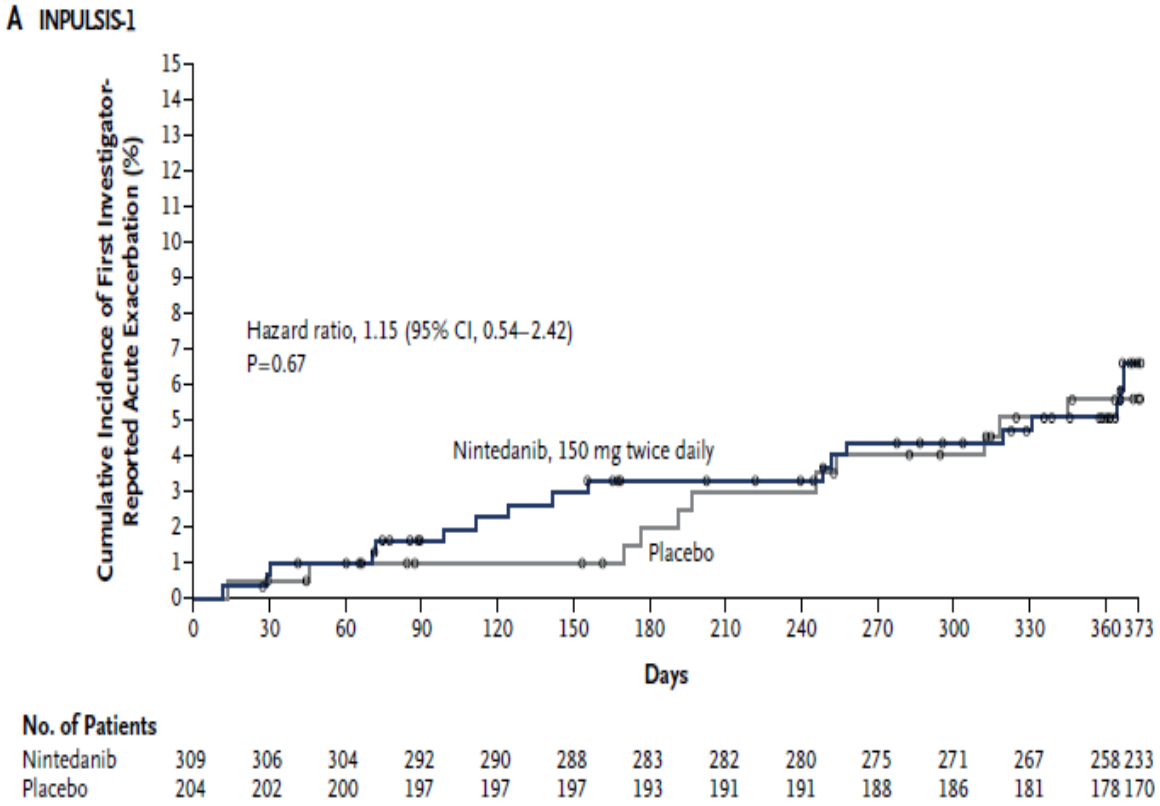
Similar for INPULSIS-2



Richeldi et al. NEJM 2014

Efficacy: IPF (Ofev)

Key secondary: time to first acute exacerbation



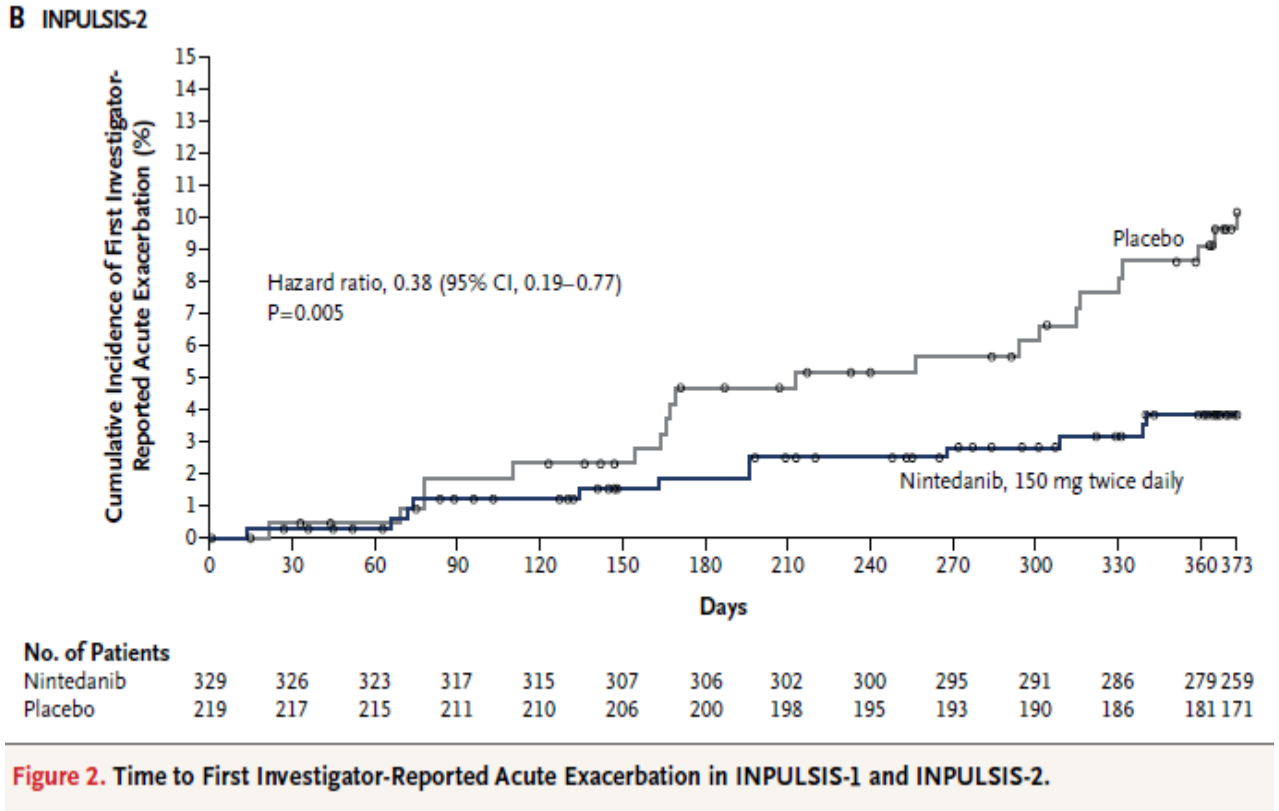


Figure 2. Time to First Investigator-Reported Acute Exacerbation in INPULSIS-1 and INPULSIS-2.

Richeldi et al. NEJM 2014

Efficacy: IPF (Ofev)

Other secondary: FVC percent predicted and survival

Table 2. Secondary Lung-Function End Points at Week 52.

End Point	INPULSIS-1			
	Nintedanib (N=307)	Placebo (N=204)	Difference, Nintedanib vs. Placebo (95% CI)	P Value
Adjusted absolute mean change from baseline in FVC — ml	−95.1	−205.0	109.9 (71.3 to 148.6)	<0.001
Adjusted absolute mean change from baseline in FVC — % of predicted value	−2.8	−6.0	3.2 (2.1 to 4.3)	<0.001

FVC percent predicted: Similar for INPULSIS-2

Survival

(Table S8 in the Supplementary Appendix). The proportion of patients who died from any cause over the 52-week treatment period was 5.5% in the nintedanib group and 7.8% in the placebo group (hazard ratio in the nintedanib group, 0.70; 95% CI, 0.43 to 1.12; P=0.14) (Fig. S8 in the

Richeldi et al. NEJM 2014

Safety: IPF (Ofev)

The most frequent serious adverse reactions reported in patients treated with OFEV, more than placebo, were bronchitis (1.2% vs. 0.8%) and myocardial infarction (1.5% vs. 0.4%). The most common adverse events leading to death in patients treated with OFEV, more than placebo, were pneumonia (0.7% vs. 0.6%), lung neoplasm malignant (0.3% vs. 0%), and myocardial infarction (0.3% vs. 0.2%). In the predefined category of major adverse cardiovascular events (MACE) including MI, fatal events were reported in 0.6% of OFEV-treated patients and 1.8% of placebo-treated patients.

Adverse reactions leading to discontinuation were reported in 21% of OFEV-treated patients and 15% of placebo-treated patients. The most frequent adverse reactions that led to discontinuation in OFEV-treated patients were diarrhea (5%), nausea (2%), and decreased appetite (2%).

Source: Ofev USPI

Safety: IPF (Ofev)

Summary of Adverse Events (portion of table reproduced)

Multiple occurrences usually not included

Table 1 Adverse Reactions Occurring in ≥5% of OFEV-treated Patients and More Commonly Than Placebo in Studies 1, 2, and 3

Adverse Reaction	OFEV, 150 mg n=723	Placebo n=508
Gastrointestinal disorders		
Diarrhea	62%	18%
Nausea	24%	7%
Abdominal pain[a]	15%	6%
Vomiting	12%	3%
Hepatobiliary disorders		
Liver enzyme elevation[b]	14%	3%

How to quantify benefit-risk?

Source: Ofev USPI

Proposal

- Before unblinding, specify important efficacy and safety endpoints

  Arrange endpoints by priority

  Example

  death, acute exacerbations, SAE1, FVC, SAE2

  (allows for inclusion of multiple occurrences of events)

- First calculate $\widehat{\boldsymbol{\theta}}$ and C.I. for binary version of primary endpoint

- Next calculate $\widehat{\boldsymbol{\theta}}$ and C.I. sequentially

Analysis

- $\widehat{Prob}$(A > B) = $\frac{1}{nm}\sum\sum I(\mathbf{Group1}_j > \mathbf{Group2}_i)$

Similarly get $\widehat{Prob}$(A < B). **$\hat{\theta}$ is ratio of probabilities**

- **95% C.I. = $exp(\ln\hat{\theta} \mp 1.96\sqrt{var})$**

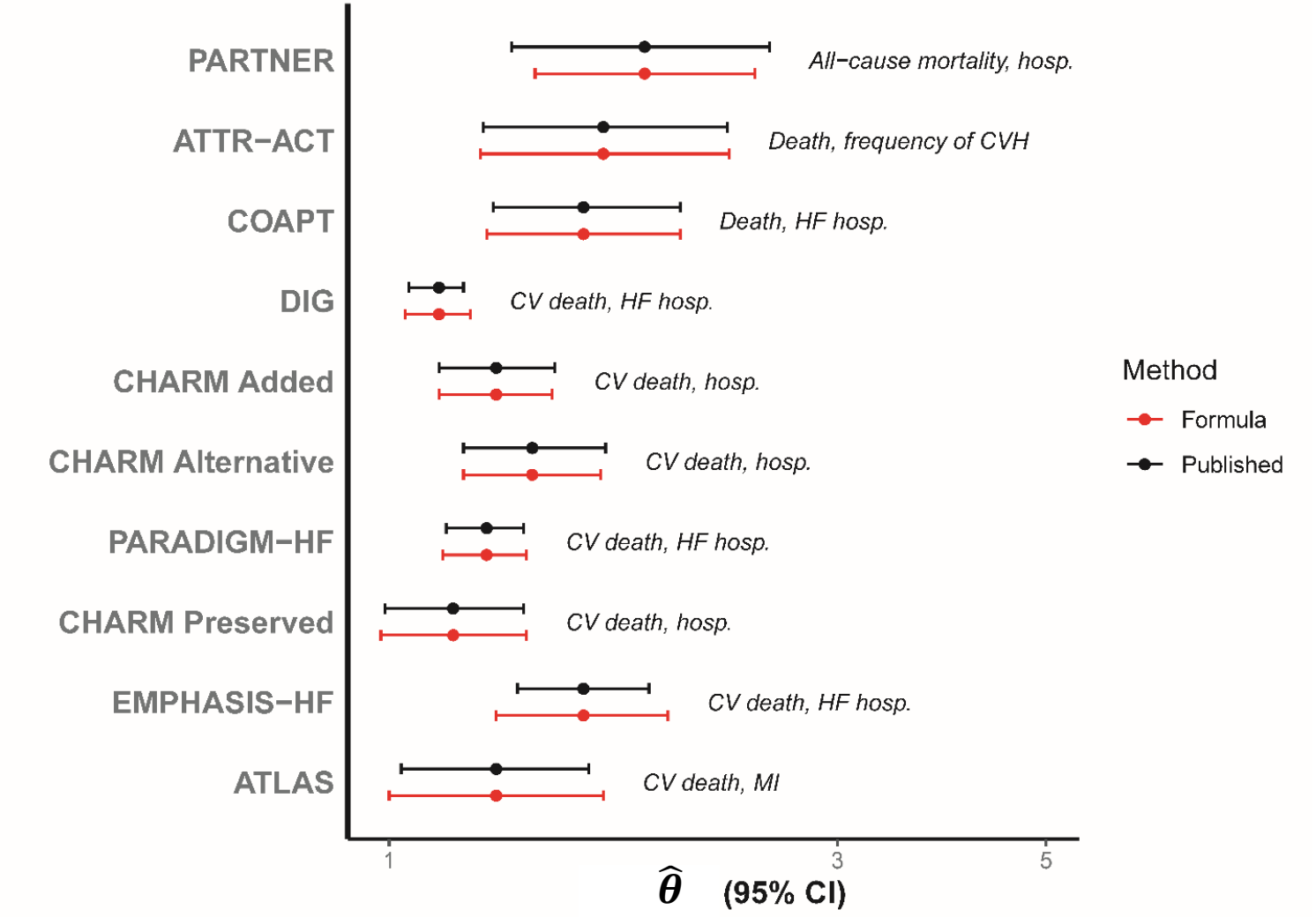
No variance formula.
Bootstrap recommended

- $var \approx \frac{4}{3(n+m)k(1-k)} \times \frac{(1+p_{tie})}{(1-p_{tie})}$
- *n* and *m* are group sizes. *k* is proportion in group 1
- p_{tie} = Prob(tie)

Relevant Literature

1. Mann-Whitney (*Annals Math Stat, 1947*) – key idea
2. Finkelstein and Schoenfeld (*Stat Med 1999*)
3. Buyse (*Stat Med 2007*)
4. Pocock et al (*Eur H J 2012*)
5. Yu and Ganju (*manuscript under review.* Var formula doesn't require individual level data)
6. Evans and Follmann. *Stat Biopharm Res* 2016 (on Benefit-Risk)

Formula Applied to Real Data



Published result required individual level data.

Formula (approximation) requires summary level data

Figure from: *Sample size formula for a win ratio endpoint.* Yu and Ganju, Manuscript under review.

Benefit-Risk Evaluation: Made Up Example

FVC (binary) at week 52

(made binary to create ties)

$\hat{\theta} = 1.50$ 95% CI: 1.25, 1.75

Benefit-Risk Assessment

Endpoint	$\hat{\theta}$	95% C.I.
Death	1.03	0.15, 1.91
+ Acute exacerbation	1.10	0.49, 1.57
+ SAE1	1.02	0.62, 1.43
+ FVC at week 52 (binary)	1.35	1.13, 1.57
+ SAE2	1.15	1.03, 1.45

Null: $\theta = 1$

B-R favorable if $\theta > 1$

FVC binary: week 52 value within 5% of baseline value

+ Acute exacerbation, for example, means the hierarchy is death followed by acute exacerbations

Handling of Missing Data

Convention

- One primary method, plus

  ‘Sensitivity’ analyses, sometimes several

  Results will vary

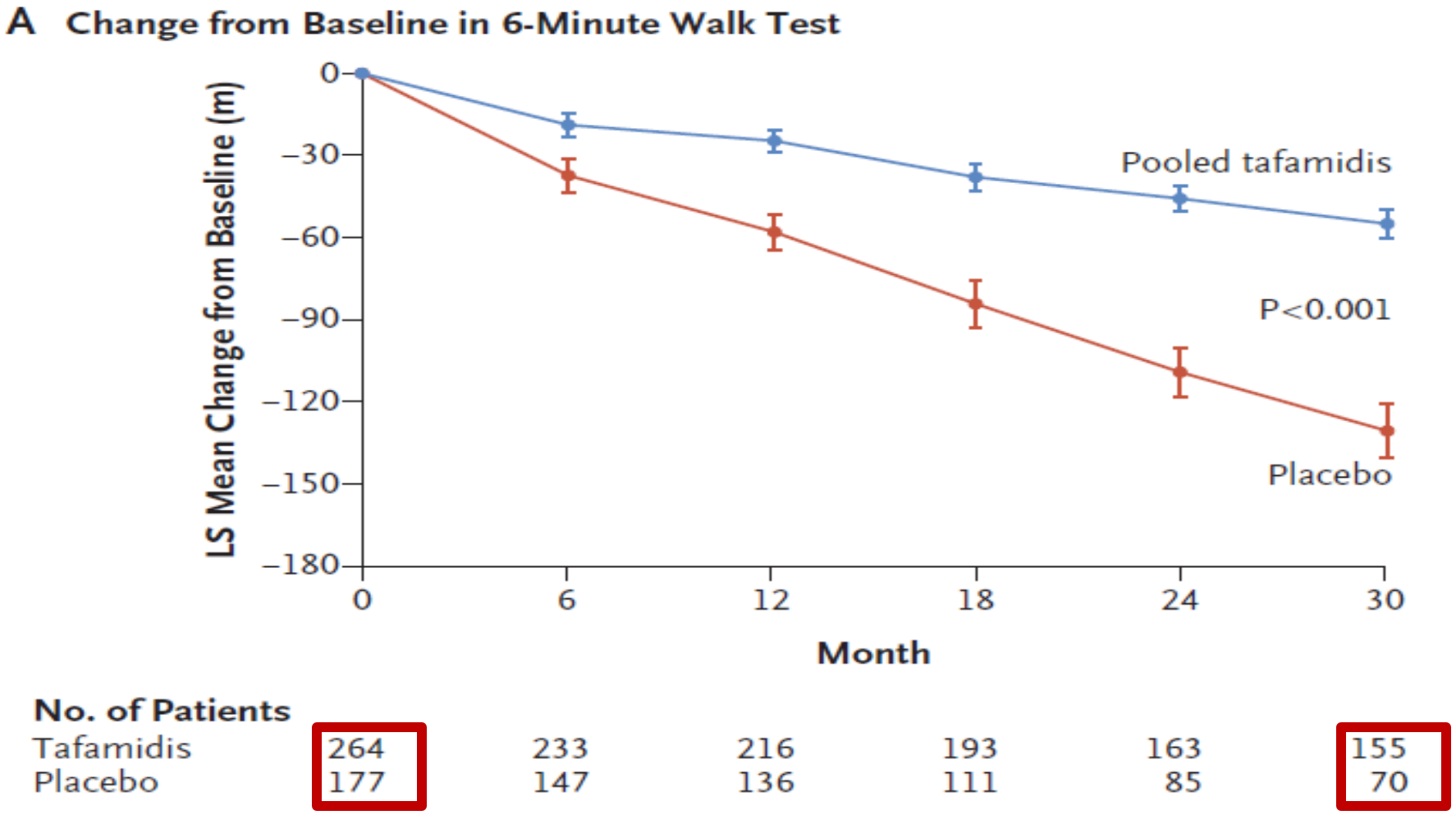
- Unknown: Which result should the public accept?

The NEW ENGLAND JOURNAL of MEDICINE

HIGHLIGHTS OF PRESCRIBING INFORMATION
These highlights do not include all the information needed to use

What principle should guide us?

Cardiovascular Trial: ATTR-ACT

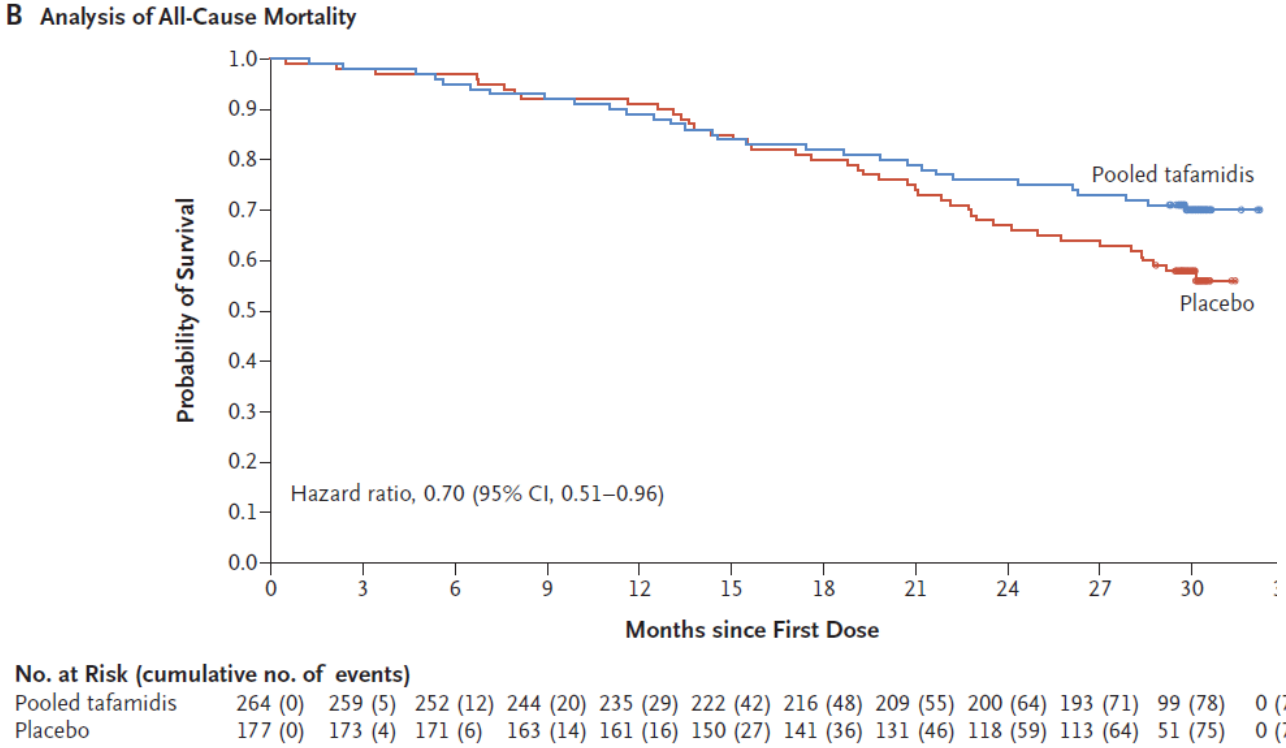


- 6MWT, secondary endpoint
- **Missing at M30: 49%**
- MMRM primary
- Sensitivity analyses (pattern mixture models)

Maurer et al NEJM 2018

CV Trial: ATTR-ACT

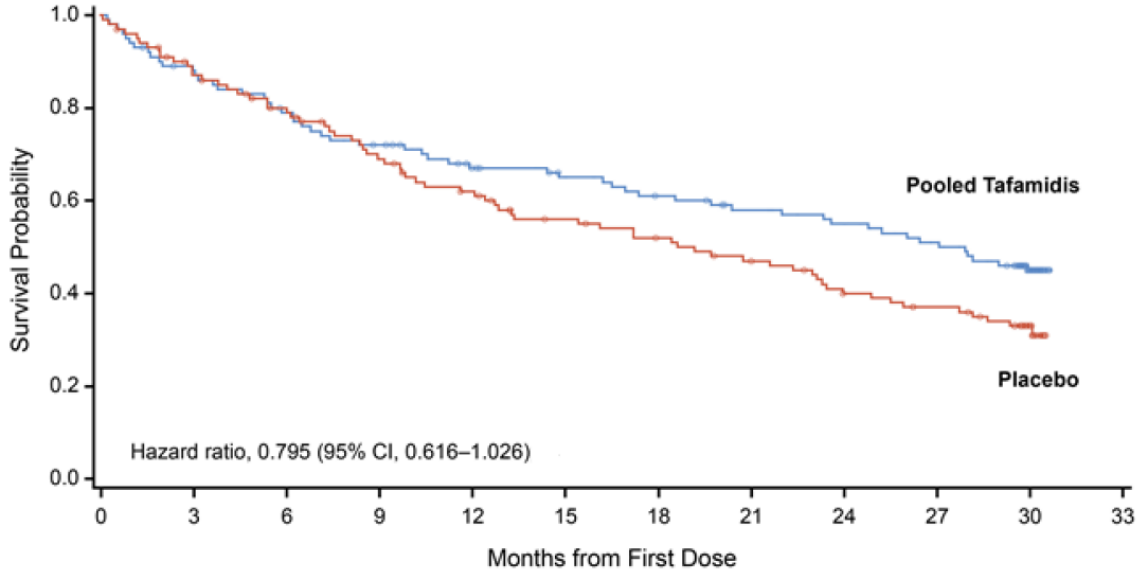
Mortality



No. at Risk (cumulative no. of events)

Pooled tafamidis	264 (0)	259 (5)	252 (12)	244 (20)	235 (29)	222 (42)	216 (48)	209 (55)	200 (64)	193 (71)	99 (78)	0 (78)
Placebo	177 (0)	173 (4)	171 (6)	163 (14)	161 (16)	150 (27)	141 (36)	131 (46)	118 (59)	113 (64)	51 (75)	0 (76)

CV hosp.



Benefit on mortality and # of CV hosp. (primary endpoint). Contributes to missing and ‘missing’ 6MWT at M30.

Maurer et al NEJM 2018

CV Trial: ATTR-ACT

FDA statistical review

Table 8: Comparison of Main Analysis and Sensitivity Analysis Results on 6MWD (ITT)

	LS means	95% CI
Main analysis	75.7	(57.6, 93.8)
Pattern mixture analysis	61.5	(44.4, 78.5)

[Source: Reviewer's table]

Variance of analysis? ☺

Suppose the mortality benefit was even greater than what was observed.
Would the pattern mixture result move closer to the MMRM result or away from it?

https://www.accessdata.fda.gov/drugsatfda_docs/nda/2019/211996Orig1s000,%20212161Orig1s000StatR.pdf

Estimands

ICH E9 (R1) addendum

Composite variable strategies

This relates to the variable of interest (see A.3.3.). An intercurrent event is considered in itself to be informative about the patient's outcome and is therefore incorporated into the definition of the variable. For example, a patient who discontinues treatment because of toxicity may be considered not to have been successfully treated. If the outcome variable was already success or failure,

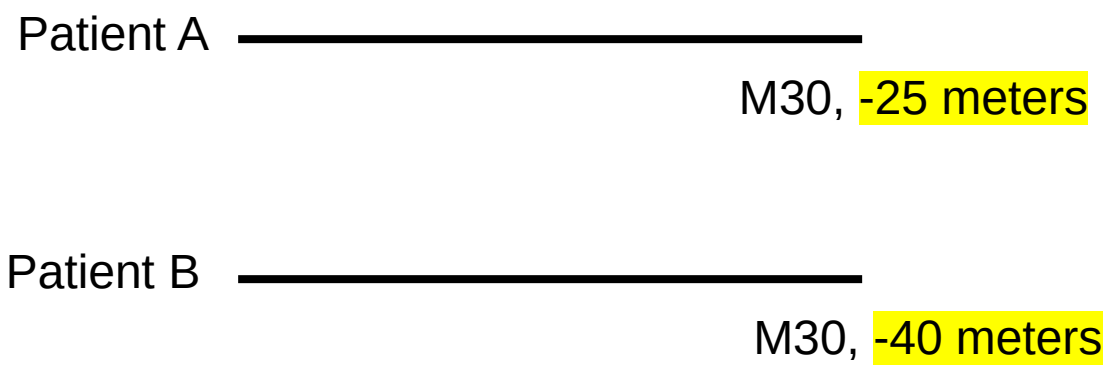
Proposed method: prioritize reason for missing data and include time when missing data occurred

Prioritization of Reasons

- Both patients have month 30 6MWT data

  Compare changes from baseline

Who did better?

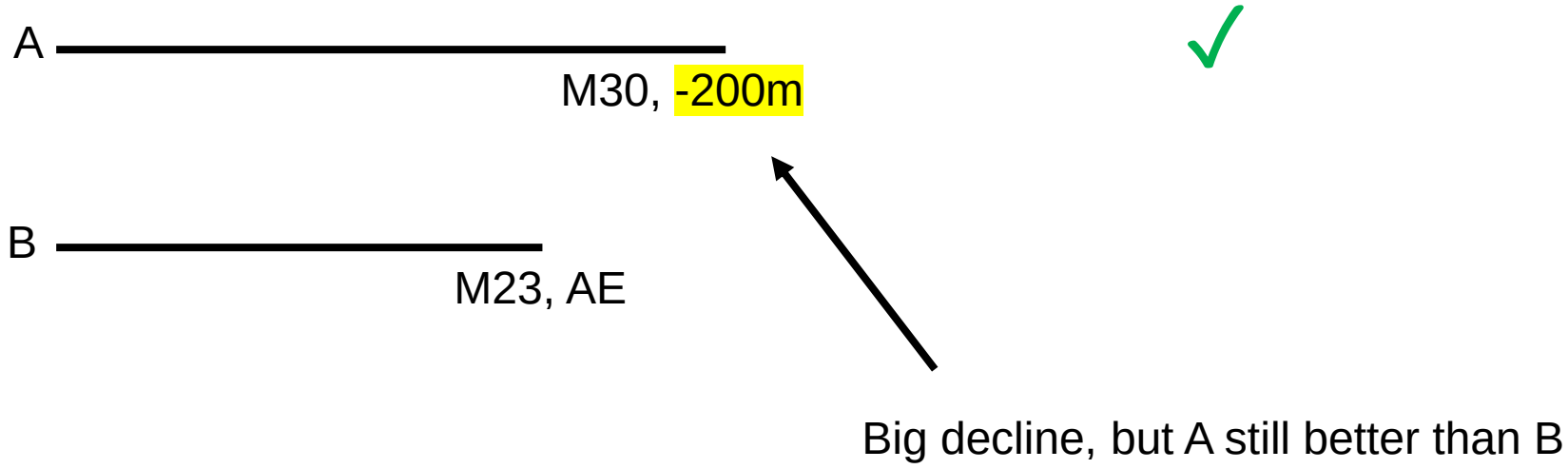


✓

Prioritization of Reasons

- Only one patient has month 30 6MWT data

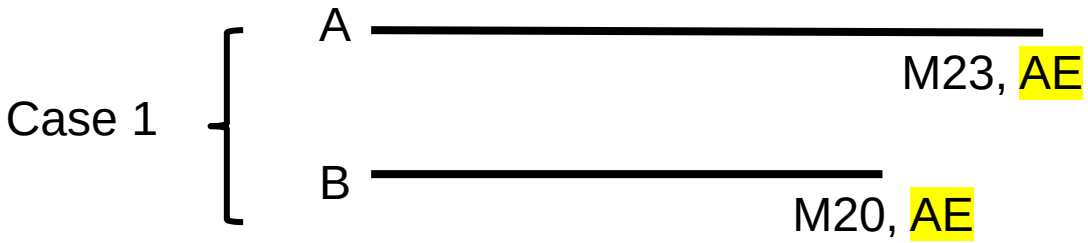
Who did better?



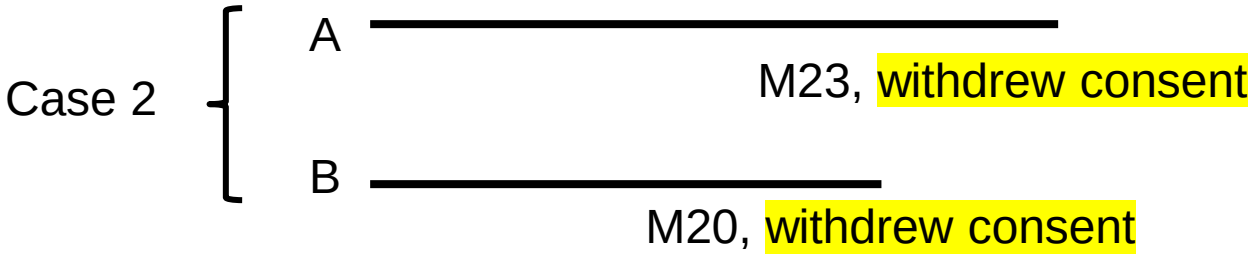
Prioritization of Reasons

- Both patients have missing data for the same reason

Who did better?



✓

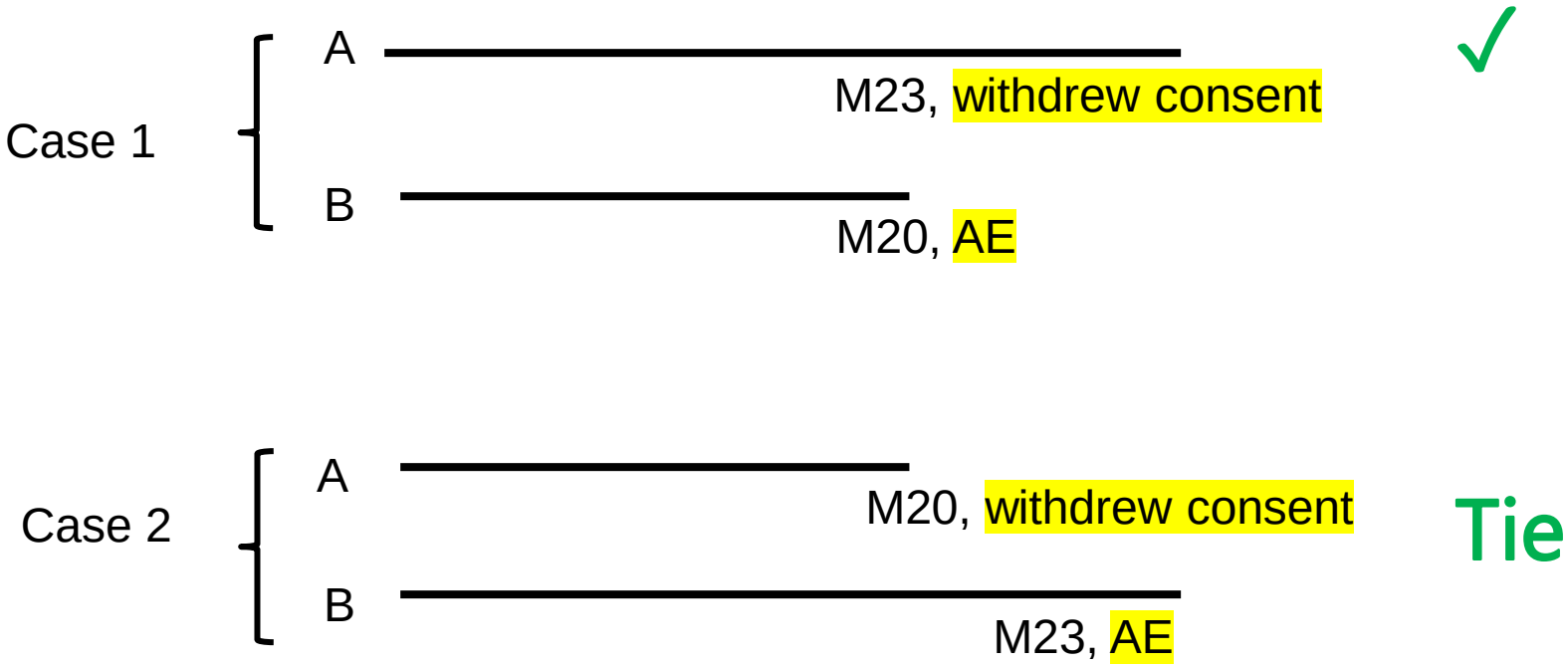


Tie

Prioritization of Reasons

- Both patients have missing data for different reasons

Who did better?



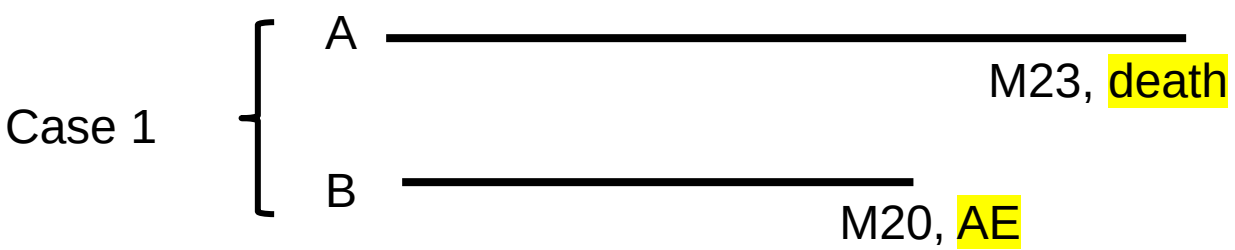
✓

Tie

Prioritization of Reasons

- Both patients have missing data for different reasons

Who did better?



✓

✓

Analysis

As before

$$\widehat{Prob}(\text{A} > \text{B}) = \frac{1}{nm}\sum\sum I(\text{Group1}_j > \text{Group2}_i)$$

$$\hat{\theta} = \frac{\widehat{Prob}(\text{A} > \text{B})}{\widehat{Prob}(\text{A} < \text{B})}$$

95% C.I. for $\hat{\theta}$: $\exp\left(\ln\hat{\theta} \mp 1.96\sqrt{var}\right)$

Missing Data and Prioritization of Reasons

- Endpoint is now a combination endpoint
  - Rules for handling missing data transparent and easily understood
- Loss of power possible compared with model-based methods
- Idea extends to other kinds of endpoints

Conclusions (Benefit-Risk, Missing Data)

- **Benefit-Risk**
  - Pairwise comparisons of hierarchically arranged endpoints provides a framework for evaluation of whether benefit > risk, and if so, by how much
  - For drugs treating the same condition, can compare the benefit-risk
- **Missing data**
  - Prioritizes reasons (including timing) for missingness. Non-parametric solution to the missing data problem

Universal (non-parametric) Sample Size Formula for Parallel Group Trials

Mann-Whitney *U* Test

- Nonparametric test of H_0: Pr(A>B)=Pr(A<B), where A and B are randomly selected values from two populations

$$U = \sum_{i=1}^{n}\sum_{j=1}^{m} S(A_i, B_j)\,,$$

with

$$S(A,B) = \begin{cases} 1, & \text{if } B < A, \\ 0, & \text{if } B = A, \\ -1, & \text{if } B > A, \end{cases}$$

Under H_0 (in the absence of tied ranks),

$$E(U) = 0$$

$$Var(U) = \frac{nm(n+m+1)}{3}$$

$$Z = \frac{U - E(U)}{\sqrt{Var(U)}}$$

Reject H_0 if $|Z| \geq Z_{1-\frac{\alpha}{2}}$

General Sample Size Formula for Mann-Whitney *U* Test Is Not Available

- Only available for special cases, e.g.
  - Continuous endpoint (Noether 1987)
  - Ordinal endpoint (Whitehead 1993; Zhao et al. 2008)
  - Recurrent event endpoint under equal duration of follow up

Inputs to Universal Sample Size Formula for Parallel Group Trials

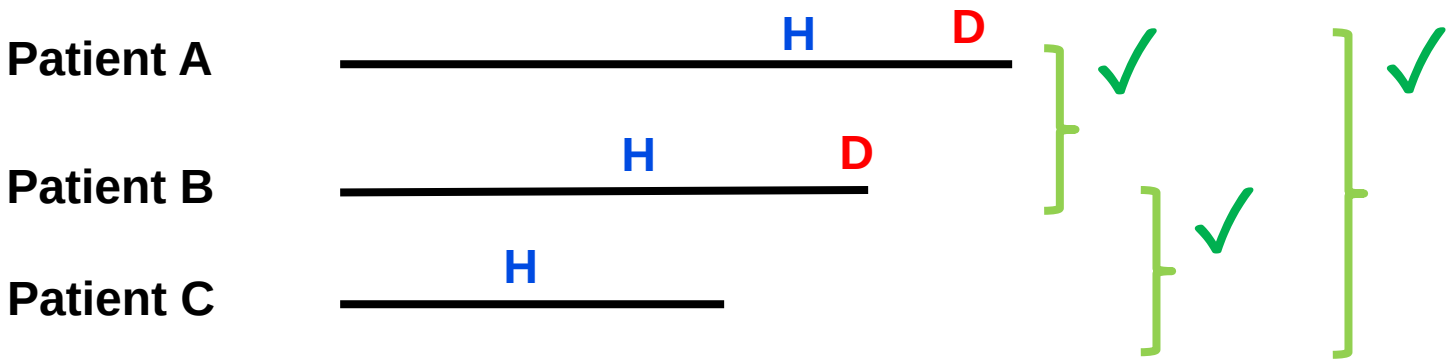
- $H_0: \theta = 1$

  $H_1: \theta > 1$

$$\theta = \frac{\Pr(A > B)}{\Pr(A < B)}$$

- Inputs to universal sample size formula are:
  - Allocation ratio, θ under H_1, α, β
  - Probabilities of transitive relationships (under H_0)
  - Probabilities of intransitive relationships (under H_0)

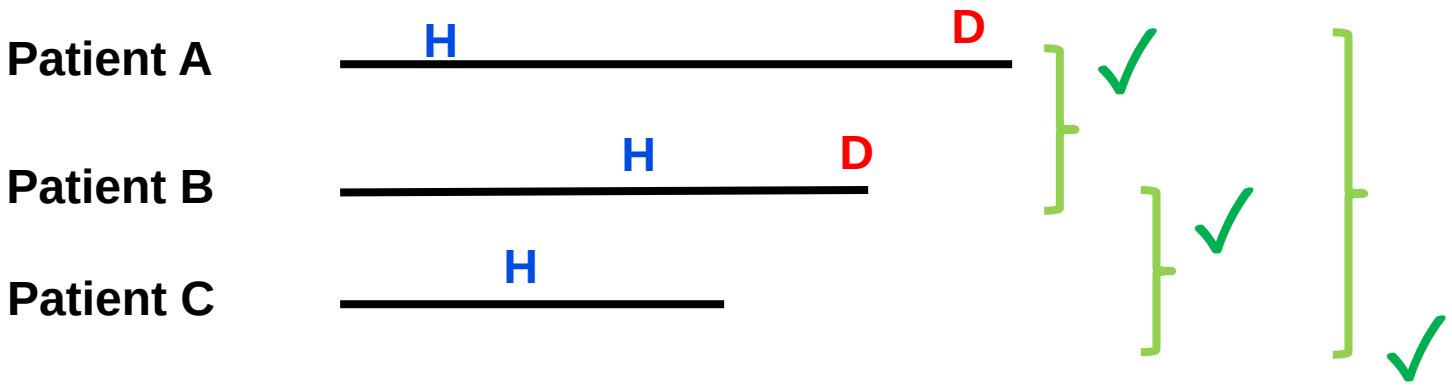
Transitive Relationships



EP1 - Time to **D**eath EP2 - Time to first **H**ospitalization

A > B, B > C, A > C

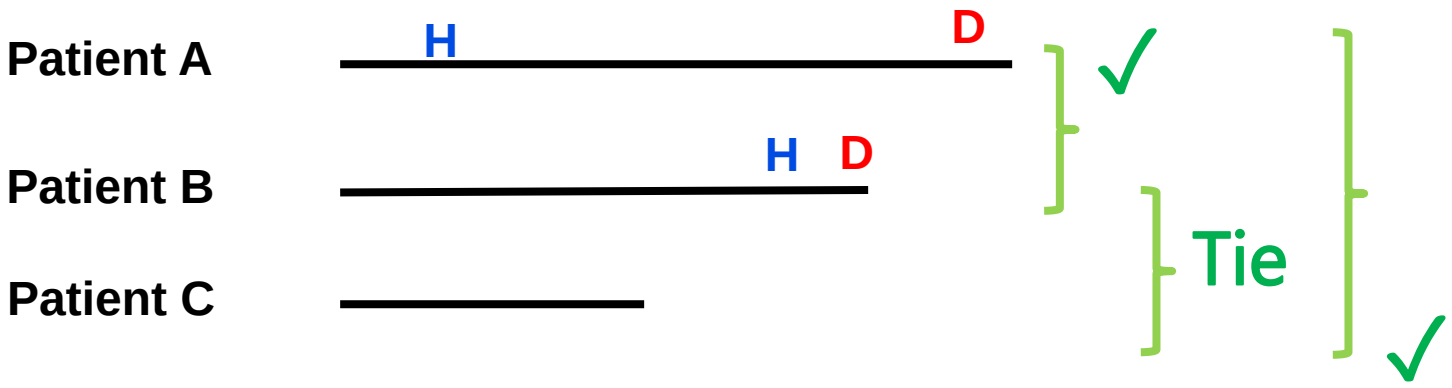
Intransitivity – Severe



EP1 - Time to **D**eath EP2 - Time to first **H**ospitalization

A > B, B > C, but C > A!

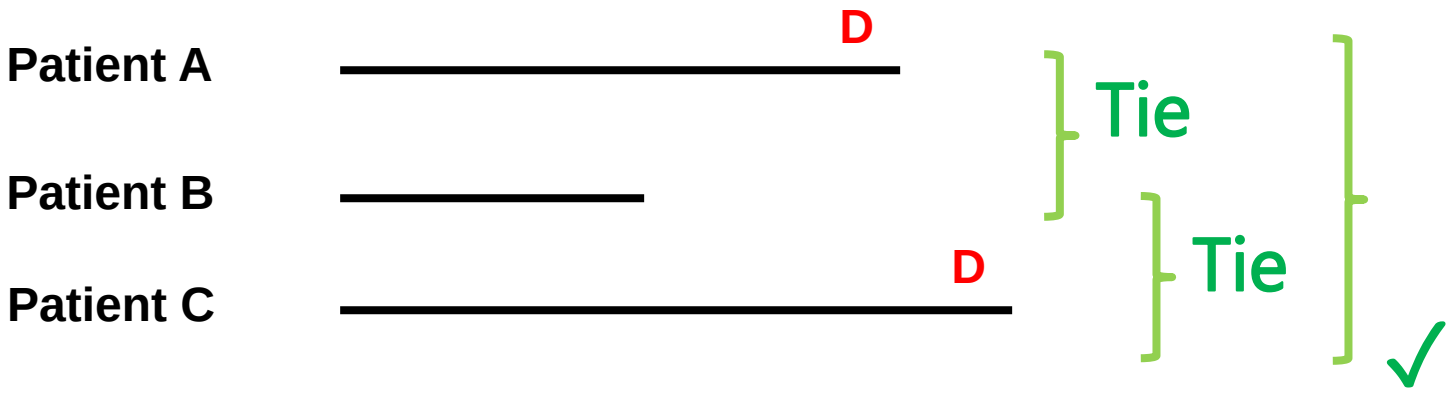
Intransitivity – Moderate



EP1 - Time to **D**eath EP2 - Time to first **H**ospitalization

A > B, B = C, but C > A!

Intransitivity – Mild



EP1 - Time to **D**eath

A = B, B = C, but C > A!

Different Types of Intransitivity

Severe

A>B, B>C, but C>A

Moderate

A>B, B=C, but C>A

Mild

A=B, B=C, but C>A

Conditions Giving Rise to Intransitivity

- Variable lengths of follow up

and

- Endpoint
  - Survival endpoint (mild intransitivity)
  - Recurrent event endpoint (any kind of intransitivity)
  - Hierarchical combination of endpoints (any kind of intransitivity)
  - Etc.

Derivation of Null Variance

- Null variance involves comparisons of a triplet of patients
- Each pairwise comparison has three possible outcomes: win, loss, or tie
- As a result, there are $3^3 = 27$ possible scenarios
- 13 are transitive scenarios; 14 are intransitive scenarios

Seven Distinct Probabilities

- Transitive relationships
  - A>B>C
  - A>B=C
  - A=B>C
  - A=B=C

- Intransitive relationships
  - A>B, B>C, C>A (severe)
  - A>B, B=C, C>A (moderate)
  - A=B, B=C, C>A (mild)

Null Variance Formula

- $$Var\left(\widehat{\Pr}(A > B) - \widehat{\Pr}(A < B)\right) \approx \frac{n+m}{nm}\left(\frac{p_1+p_2+p_3}{3} - p_5 - \frac{p_6}{3}\right)$$

$$= \frac{n+m}{3nm}\left[1 - p_4 - (4p_5 + 2p_6 + p_7)\right],$$

where

$$p_1 = \Pr(A > B > C)$$
$$p_2 = \Pr(A > B = C)$$
$$p_3 = \Pr(A = B > C)$$
$$p_4 = \Pr(A = B = C) = \Pr(3\text{−way tie})$$
$$p_5 = \Pr(A > B, B > C, C > A) = \Pr(\text{severe intransitivity})$$
$$p_6 = \Pr(A > B, B = C, C > A) = \Pr(\text{moderate intransitivity})$$
$$p_7 = \Pr(A = B, B = C, C > A) = \Pr(\text{mild intransitivity})$$

Special Cases

$$\frac{n+m}{3nm}\left[1-p_4-(4p_5+2p_6+p_7)\right]$$

- No tie, no intransitivity

$$Var\left(\widehat{\Pr}(A>B)-\widehat{\Pr}(A<B)\right)\approx\frac{n+m}{3nm}$$

- No intransitivity (ties okay)

$$Var\left(\widehat{\Pr}(A>B)-\widehat{\Pr}(A<B)\right)\approx\frac{n+m}{3nm}(1-p_4)$$

- Survival endpoint

$$Var\left(\widehat{\Pr}(A>B)-\widehat{\Pr}(A<B)\right)\approx\frac{n+m}{3nm}(1-p_4-p_7)$$

Sample Size Formula for Parallel Group Trials

$$Var\left(\log\left(\hat{\theta}\right)\right) \approx \frac{4}{\left(1 - \Pr(2\text{−way tie})\right)^2} Var\left(\widehat{\Pr}(A > B) - \widehat{\Pr}(A < B)\right)$$

$$= \frac{\sigma^2}{N}. \qquad (N = n + m)$$

$$\boldsymbol{N = \frac{\sigma^2\left(Z_{1-\alpha} + Z_{1-\beta}\right)^2}{\log^2(\theta)}}$$

Conclusions

- Mann-Whitney *U* test has been around for a long time, but until now a general sample size formula has not been available
- Hierarchically combined endpoint is getting popular, sample sizes are calculated via simulations which are time-consuming and not straightforward
- To use the proposed formula, we need to build experience with the component probabilities:
  - Probability of a 3-way tie
  - Probabilities of intransitive outcomes

References

- Noether, G. E. (1987). Sample size determination for some common nonparametric tests. *Journal of the American Statistical Association*, *82*(398), 645-647.
- Whitehead, J. (1993). Sample size calculations for ordered categorical data. *Statistics in medicine*, *12*(24), 2257-2271.
- Zhao, Y. D., Rahardja, D., & Qu, Y. (2008). Sample size calculation for the Wilcoxon–Mann–Whitney test adjusting for ties. *Statistics in medicine*, *27*(3), 462-468.