Composite Endpoints

Dean Follmann

National Institute of Allergy and Infectious Diseases

Outline

- Simple Phase III Trial
- Composite endpoints
- Co-equal primary endpoints
- Co-equal surrogate endpoints
- Combining 1^0 & 2^0 endpoints
- Ranking Methods
- Conclusions

Simple Phase III trial

- Choose a single relevant endpoint
  - Death
  - Ejection fraction of the left ventricle
- Size trial for 90% power to detect a clinically important effect
  - 20% reduction in mortality
  - .05 difference in EF

Sample size formula

- Treatment Effect: more is better
  - Pick a responsive endpoint
- Variability: less is better
  - Get more measurements/stable endpoint
- Events: more are better
  - Include sicker patients
  - Lengthen follow-up

Characteristics of a clinical endpoint (Neaton et al 1994)

- Should be relevant and easy to interpret.
- Should be clinically apparent and easy to diagnose.
- Should be sensitive to treatment differences.

More complicated world

- Occasionally, a single primary endpoint undesirable. Why?
  - Clinically important events are rare.
  - Effect of treatment manifested on a variety of important endpoints.

Example: ACES trial

- ACES—trial to evaluate antibiotics versus placebo in patients at risk of CHD events.
- Primary endpoint is
  - Hospitalization for unstable angina
  - CHD death
  - Nonfatal MI
  - Revascularization

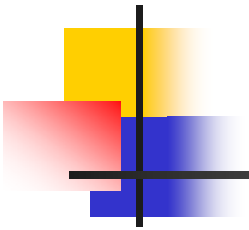
Composite Concerns

- With a composite endpoint, relative importance of various constituent endpoints determined by frequency.
- CHD death or revascularization
  - CHD death 1%
  - Revascularization 10%

Composite Concerns

- Only include constituent endpoints who are reasonably influenced by treatment.
  - Treatment: 50% on death, 20% on MI
  - Control rate Treatment rate
  - Death .01 .005
  - MI .01 .008
- Death alone vs Death or MI: same power

Bonferroni approach.



- Use p-values for the two endpoints.
- Reject if p_1 or p_2 less than .05/2
- Inference drawn for each endpoint
- Good if treatment has entire effect on one endpoint or the other, don't know which one.

Example: PEPI

- Postmenopausal Estrogen/Progestin Interventions Trial. HRT's effect on risk factors for heart disease.
- 875 women assigned to 5 combinations.
- Primary endpoints
  - HDL-C
  - SBP
  - Serum insulin
  - fibrinogen

O'Brien (1984) Rank-Sum method

- Rank each outcome and calculate an average rank for each patient
- See if average rank differs between groups.

Sub	X_1	R_1	X_2	R_2	Avg R
Fred	3.3	2	87	1	1.5
Joe	4.1	3	105	2	2.5
Sam	1.7	1	1000	3	2.0

O’Brien OLS method

- Standardize each endpoint. Compute the average endpoint for each person and perform a t-test on the averages.

Sub	X_1	$(X_1-❍)/♦$	X_2	$(X_2-❍)/♦$	avg
Fred	3.3	.22	87	-.72	-.25
Joe	4.1	.87	105	-.42	.23
Sam	1.7	-1.10	200	1.14	.03

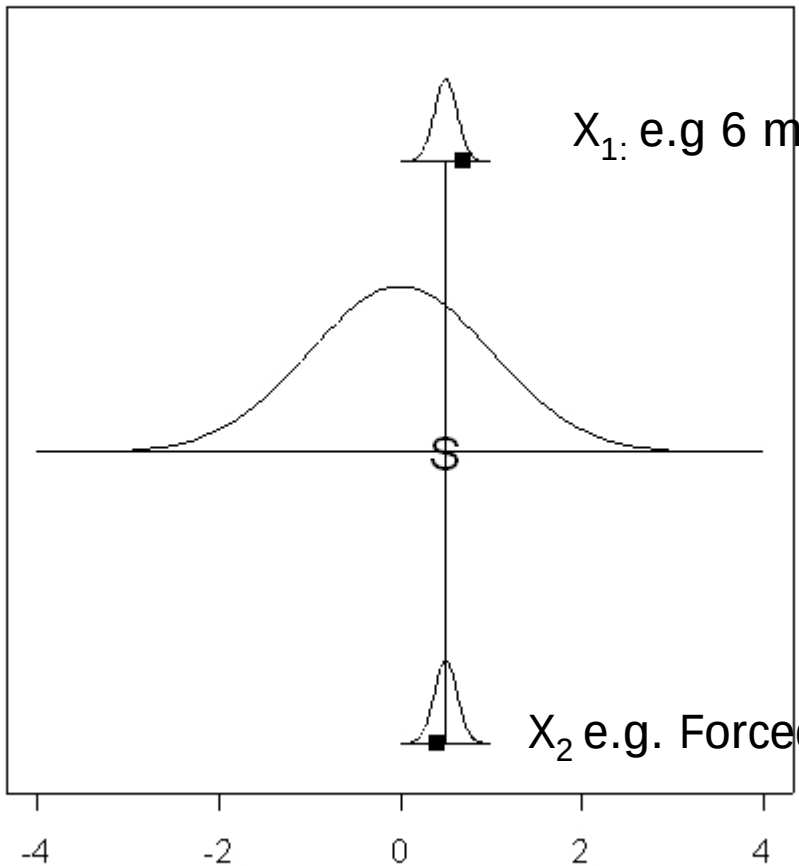
O’Brien GLS method

- Assume common treatment effect 
  - e.g. 1 standard deviation on both endpoints.
- Calculate a statistically optimal estimate of using a weighted average. (more correlated endpoints, less weight).
- Pocock Geller Tsiatis (1987) generalize to binary/survival etc endpoints.
- Many other methods conceptually similar: specify a model with the same  for many endpoints.

Latent Variable models

- Assume each person has an underlying severity, S, which influences several endpoints.
- E.g. MPS---Lysosomal enzyme deficency
  - **FVC**
  - **6 minute walk**
  - AHI
  - shoulder flexion
  - visual acuity
- Test whether the distribution of underlying severities is moved by treatment.

Conceptual framework for latent variable model



$X_{1:}$ e.g 6 minute walk distance

(latent) Severity dbn
in Treatment group

X_2 e.g. Forced vital capacity

👎 = Mean(S) in control - Mean(S) in treatment

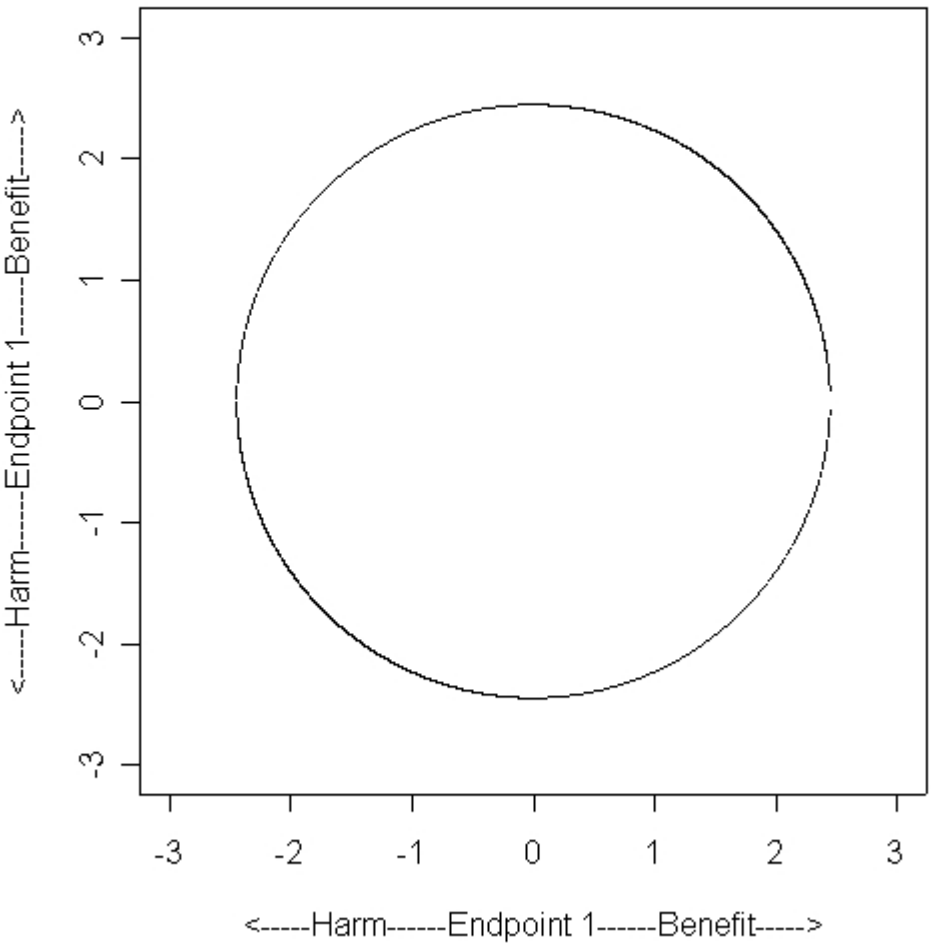
A model

- Simple Model
  - Y_{i1} = B_01 + D Z_i + b_i + e_{i1}
  - Y_{i2} = B_02 + D Z_i + b_i + e_{i2}
    - e_ij ~ N(0, Vej)
    - S_i ~ N(0, Vs)

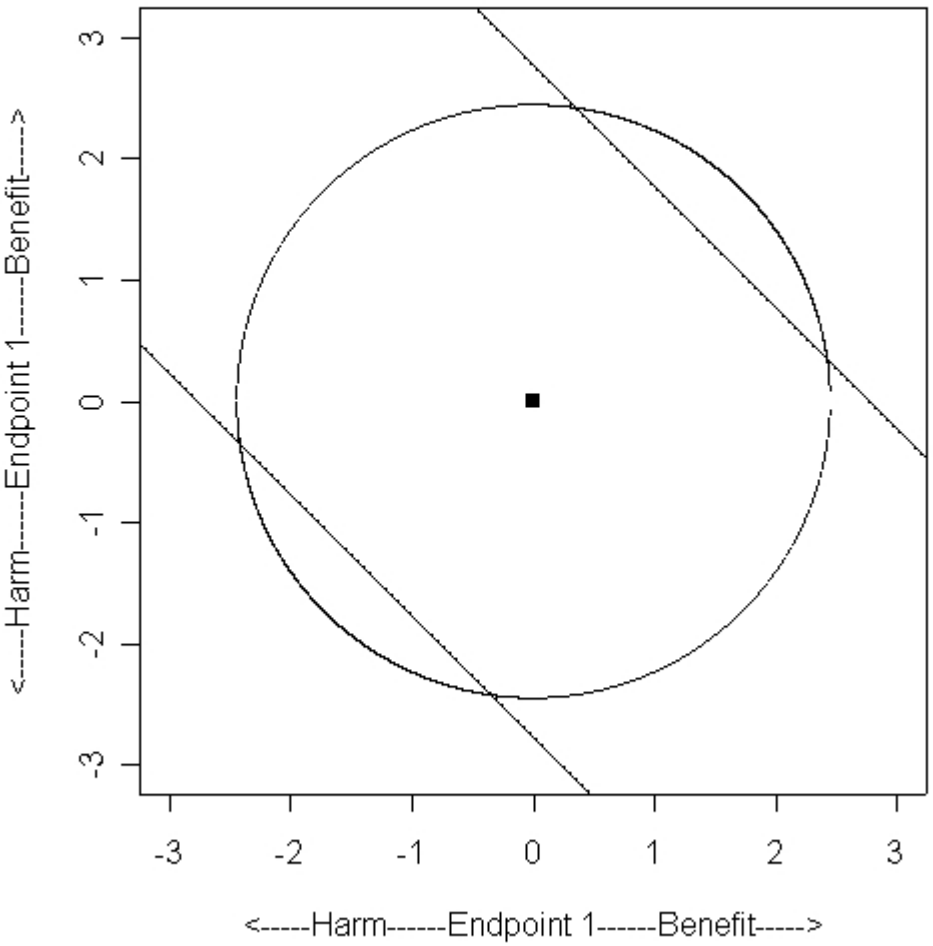
Global Tests

- Hotelling T^2---multivariate t-test
- Good for *any* treatment effect, so less good for uniformly beneficial treatment effects.

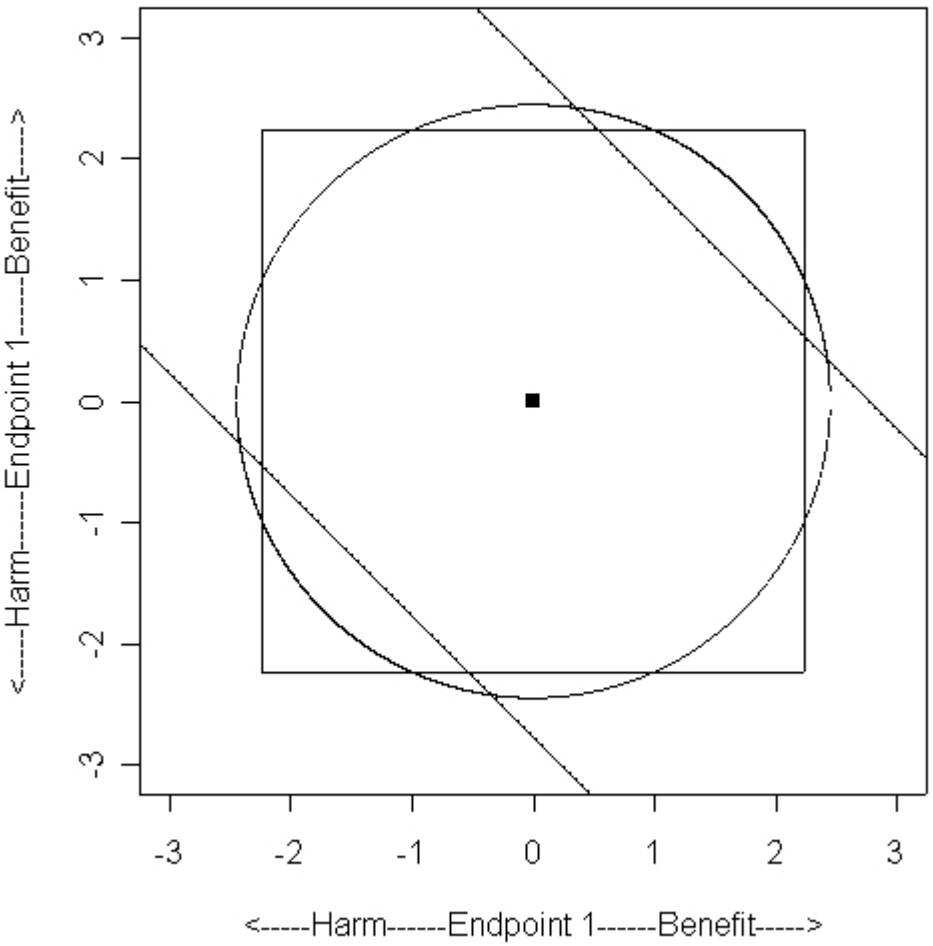
Rejection Region for Hotelling's T^2 Test



Rejection Regions for Hotelling's T^2 Test & O'Brien test



Rejection Regions for Hotelling's T^2, O'Brien, & Bonferroni Tests



Tilley et al (1996) for Stroke trial

- Trial of t-PA versus placebo in patients with acute ischemic stroke.
- Dichotomized 4 stroke scales.
- Discussed use of Bonferroni, Hotelling's Test & O'Brien's GLS test.
- Reject if
  - Mean(Z) > 1.96 * [(1+ 3 ❒  
  - > 1.96 if ❒ 
  -  1.96/2 if ❒ 

Combining co-equal but surrogate endpoints

- Suppose both endpoints are surrogates.
- Ideally form a risk score.
  - R = w_1 DBP + w_2 SBP + w_3 serum insulin...
  - R = w_1 Hepatitis + w_2 sex for drugs + ...
- Do a t-test using R.

Combining 1^0 and auxiliary...

- 1^0 endpoint alone: use Wilcoxon Rank sum approach.
- Compare each pair of treatment/control patients
- Y_{ij} = 1 if "i" (in T) lives past "j" (in P)
- = 1/2 if both live
- = 0 if "i" (in T) dies before "j" (in P)
- Form mean(Y_{ij}) = Pr(live longer on T than P)
- Equivalent to ranking by death time.

Combining 1^0 and auxiliary...

- If both live, replace ½ with
  - p_{ij} = Pr(i lives longer than j | CD4s)
- May be useful if
  - CD4/death relationship in past = future
  - Treatment effects CD4 counts & they differ at end
- Similar approach taken by Faucett Schenker Taylor (2002) who imputed death times.

“Utility” Ranking Methods

- May be hard to say MI is half as bad as death. But clearly death is worse.
  - Death is worst
    - Rank by death time
  - 2 Strokes worse than 1
    - Rank by time of first stroke
  - 1 Stroke worse than nothing
    - Rank by time of stroke.
- Compare the ranks between groups

Table III. The cards ordered by consensus ranks and the mean and standard deviation of the 19 ranks

Consensus rank	Time of event (years) 0 1 2 3 4	Card number	Rank Mean	Rank Std	Time to death	Number of events
1	✱	6	1·2	0·4	0·4	1
2	@	8	1·9	0·5	0·4	1
3	@	27	3·0	0·3	0·8	1
4	D	29	4·0	0·3	1·1	1
5	S ✱	35	5·3	0·9	2·0	1
6	S D	16	6·9	1·5	2·1	2
7	S S ✱	28	7·5	1·4	2·9	3
8	H H D	1	8·6	1·8	3·0	3
9	H ✱	40	8·8	1·6	2·8	2
10	H D	13	11·5	1·8	3·0	2
11	S @	39	11·7	1·8	3·3	2
12	✱	24	12·2	2·7	2·8	1
13	S H@	21	12·2	3·1	3·7	3
14	H @	2	12·5	1·0	3·5	2
15	H H@	38	15·2	2·3	3·4	3
16	D	10	16·7	2·1	3·5	1
17	@	36	17·6	2·2	3·6	1
18	✱	34	18·5	1·6	3·7	1
19	S S S	32	21·0	8·1		3
20	S H	15	21·5	1·9		2
21	H S	42	21·7	1·9		2
22	H H	3	23·5	4·7		2
23	H H	19	23·8	4·6		2
24	S S	4	24·2	4·0		2
25	H H H	14	24·7	4·4		3
26	S S	20	25·2	4·6		2
27	H S	23	27·7	3·7		2
28	S	18	28·4	5·3		1
29	S S	17	28·6	4·4		2
30	H H	26	30·0	4·4		2
31	H S	30	30·9	3·5		2
32	H S	41	31·9	3·7		2
33	S H	9	32·2	3·6		2
34	H	12	32·4	4·7		1
35	S S	37	34·0	3·5		2
36	H	33	35·1	3·4		1
37	H H	31	35·7	2·8		2
38	S	7	35·8	3·5		1
39	S	22	38·5	2·7		1
40	H	43	39·5	2·9		1
41	S	5	39·8	2·0		1
42	H	11	41·4	2·5		1
43		25	43·0	0·0		0

0 1 2 3 4
Time of event (years)

The mean and std (standard deviation) of the rank are based on the sample of 19 rankers.
H non-fatal heart attack
@ fatal heart attack
✱ fatal stroke
S non-fatal stroke
D death from neither heart attack nor stroke

HIV vaccine trials

- Want HIV vaccine to reduce acquistion and also post-infection viral load for those infected. How to combine?
  - Those who are uninfected get best rank
  - Those who are infected are ranked by viral load “setpoint” lower setpoints get higher ranks.

Weighting

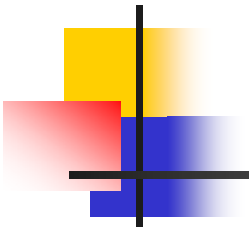
- You may not be interested in weighting, but weighting is interested in you.
- Approaches we discussed.
  - Equal weight for all endpoints (e.g. OLS)
  - More weight for frequent events (e.g. composite)
  - Less correlated outcomes more weight (e.g. GLS)

- Clinically interpretable weights?

Conclusions

- Common approaches are to pick a composite endpoint or adopt a Bonferroni correction.
- Clinical relevance / interpretability paramount.
- Appropriate approach depends heavily on the application.
- Novel endpoints/analysis approaches should be thoroughly investigated.

- Bandeen-Roche K, Miglioretti DL, Zeger SL, et al. (1997) “Latent variable regression for multiple discrete outcomes” Journal of the American Statistical Association, 1375-1386.
- Bjorling L, Hodges J, (1997) “Rule-Based Ranking schemes for antiretroviral trials” Statistics in Medicine,1175-1191.
- Faucett C, Schenker N, Taylor J, (2002) “Survival Analysis using auxiliary variables via multiple imputation, with application to AIDS clinical trial data” Biometrics, 37-47.
- Follmann D, (1995) “Multivariate Tests for Multiple Endpoints in Clinical Trials” Statistics in Medicine, 1163-1176.
- Follmann D, (1996) “A Simple Multivariate Test for One Sided Alternatives” Journal of the American Statistical Association, 854-861.



- Follmann D, Wittes J, Cutler J, (1992) “A Clinical Trial Endpoint Based on Subjective Rankings” Statistics in Medicine, 427-438.
- Lefkopoulou M, Ryan L, (1993) “Global tests for multiple binary outcomes” Biometrics, 975-988.
- Legler JM, Ryan LM, (1997) “Latent variable models for teratogenesis using multiple binary outcomes” Journal of the American Statistical Association, 13-20.
- Miller VT, Larosa J, Barnabei V, et al. (1995) “Effects of Estrogen or Estrogen/Progestin Regimens on Heart-Disease Risk Factors in Postmenopausal Women” The Postmenopausal Estrogen/Progestin Interventions (PEPI) Trial, 199-208.
- Neaton J, Wentworth D, Rhame, et al. (1994) “Considerations in choice of a clinical endpoint for AIDS clinical trials”Statistics in Medicine, 2107-2125.

- O'Brien PC, (1984) “Procedures for comparing samples with multiple endpoints” Biometrics, 1079-1087.
- O'Brien PC, Geller NL, (1997) “Interpreting tests for efficacy in clinical trials with multiple endpoints” Controlled Clinical Trials, 222-227.
- Pocock S, Geller N, Tsiatis A, (1987) “The analysis of multiple endpoints in clinical trials” Biometrics, 487-498.
- Sammel M, Lin X, Ryan L, (1999) “Multivariate linear mixed models for multiple outcomes” Statistics in Medicine, 2479-2492.
- Sammel MD, Ryan LM, Louise M, Legler JM, (1997) “Latent variable models for mixed discrete and continuous outcomes” Journal of the Royal Statistical Society, Series B, Methodological, 667-678.

- Tang DI, Geller NL, Pocock SJ, (1993) “On the design and analysis of randomized clinical trials with multiple endpoints” Biometrics, 23-30.
- Tilley BC, Marler J, Geller NL, et al. (1996) “Use of a global test for multiple outcomes in stroke trials with application to the national institute of neurological disorders and stroke t-PA stroke trial” Stroke, 2136-2142.
- Wassmer G, Reitmeir P, Kieser M, Lehmacher W, (1999) “Procedures for testing multiple endpoints in clinical trials: An overview” Journal of Statistical Planning and Inference, 69-81.
- Zhang J, Quan H, Ng J, et al. (1997) “Some statistical methods for multiple endpoints in clinical trials” Controlled Clinical Trials, 204-221.
- Zhong A, Song C, Reiss TF, (2004) “An endpoint for worsening asthma: Development of a sensitive measure and its properties” Drug Information Journal, 5-13.

Mariamman: Goddess of pox

- Afflicted individuals provide offerings
- Follow them home
- Successful ring vaccination
- Smallpox eradicated

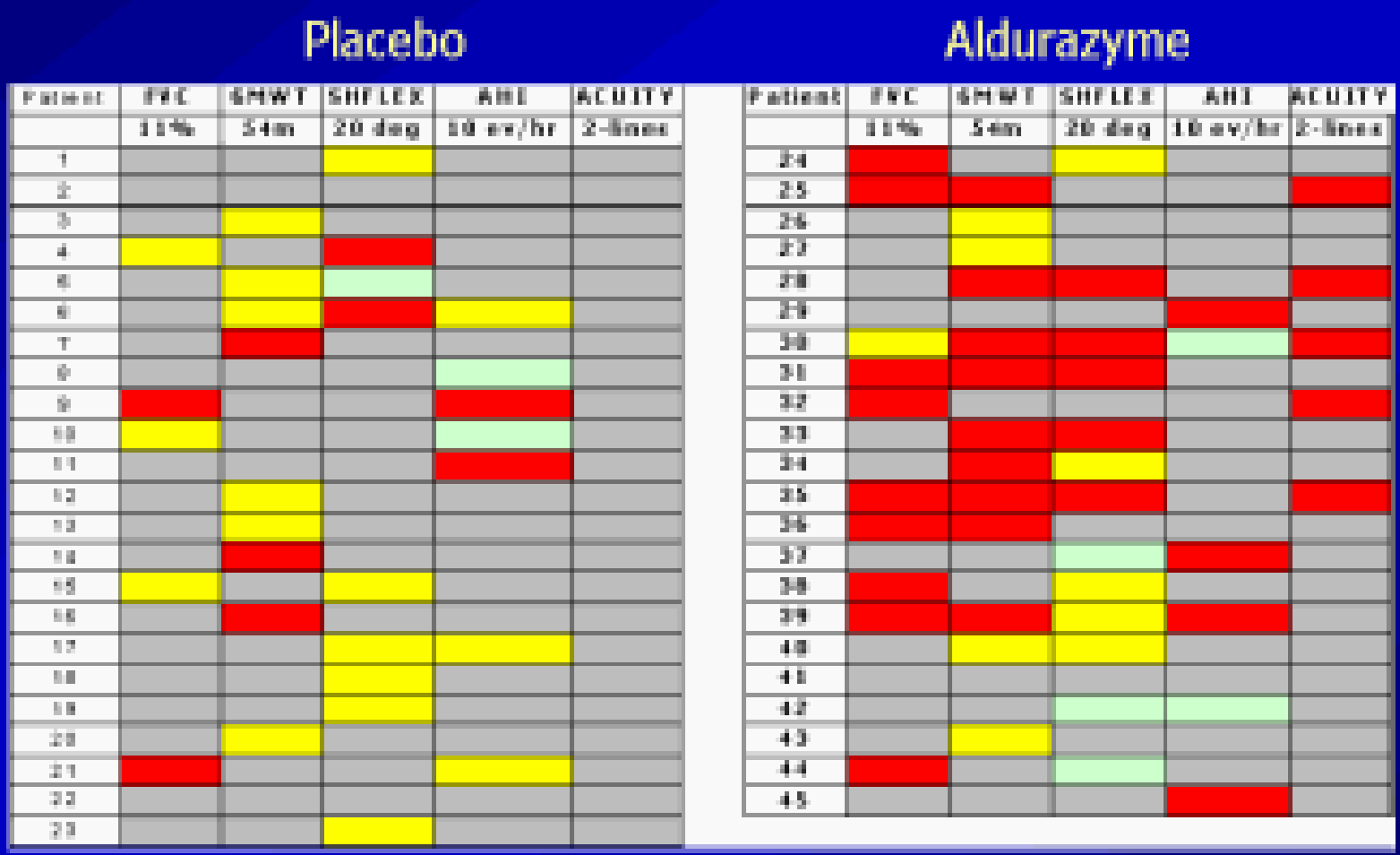
Novel Design Issues

- Would a crossover trial make sense?
  - Area under EDSS curve over time.
- Enroll patients during a remission?
- For a phase II study, could a placebo be ethically used for a short while?
- Could all patients receive drug at end of study?
- Can we cross-over at time of failure?

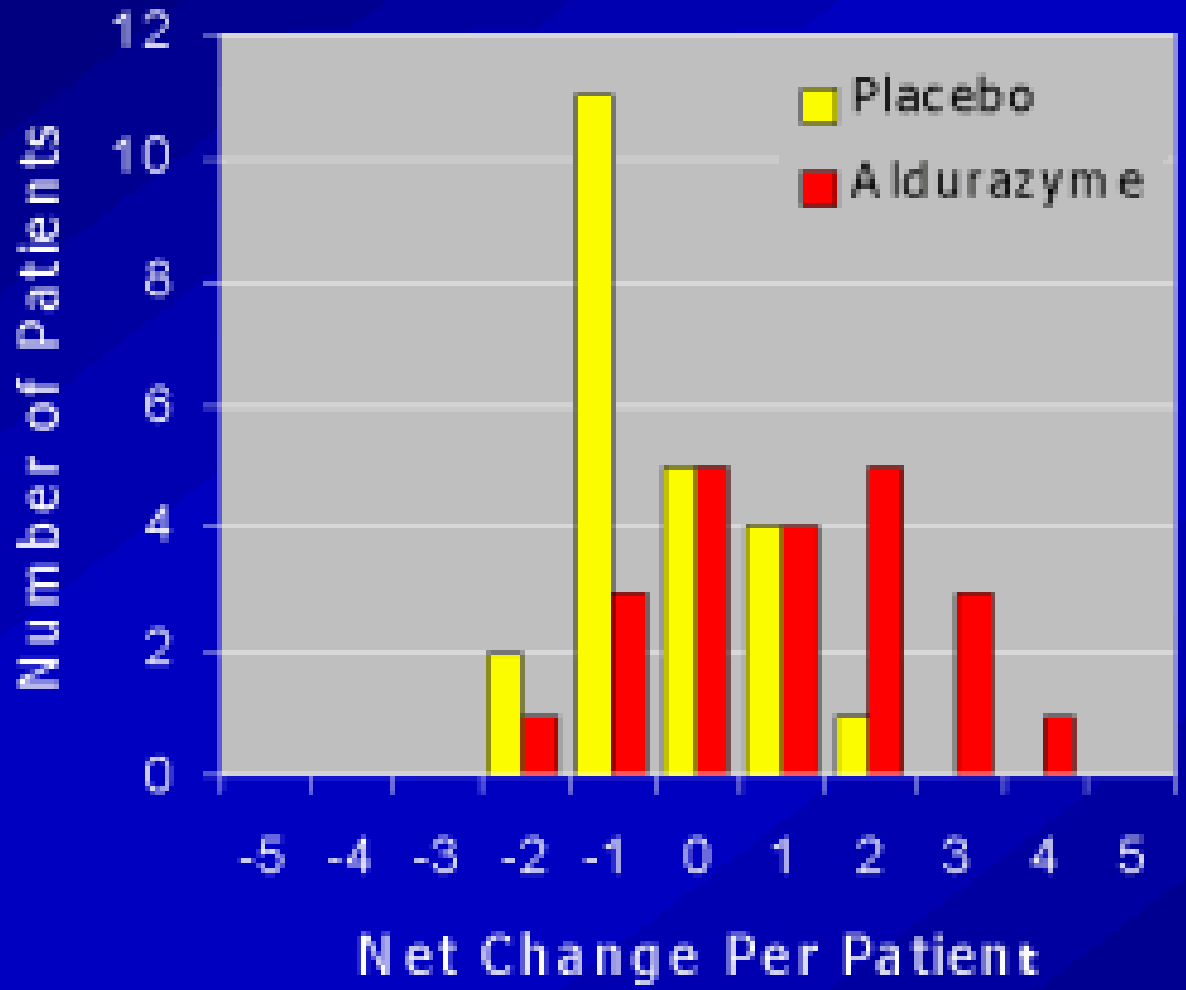
Aldurazyme trial in MPS

- MPS: lysosomal enzyme deficency, leads to GAG accumulation with multisystemic effects.
- Inclusion criteria:
  - Stand 6 minutes, walk > 5 meters
- weekly IV infusion for ½ year.
- N=45
- Endpoints: **FVC, 6 minute walk**, AHI, shoulder flexion, visual acuity.

Phase 3 Study: Post-Hoc Analysis Composite Endpoint



Phase 3 Study: Post-Hoc Analysis Aldurazyme Leads to Net Improvement



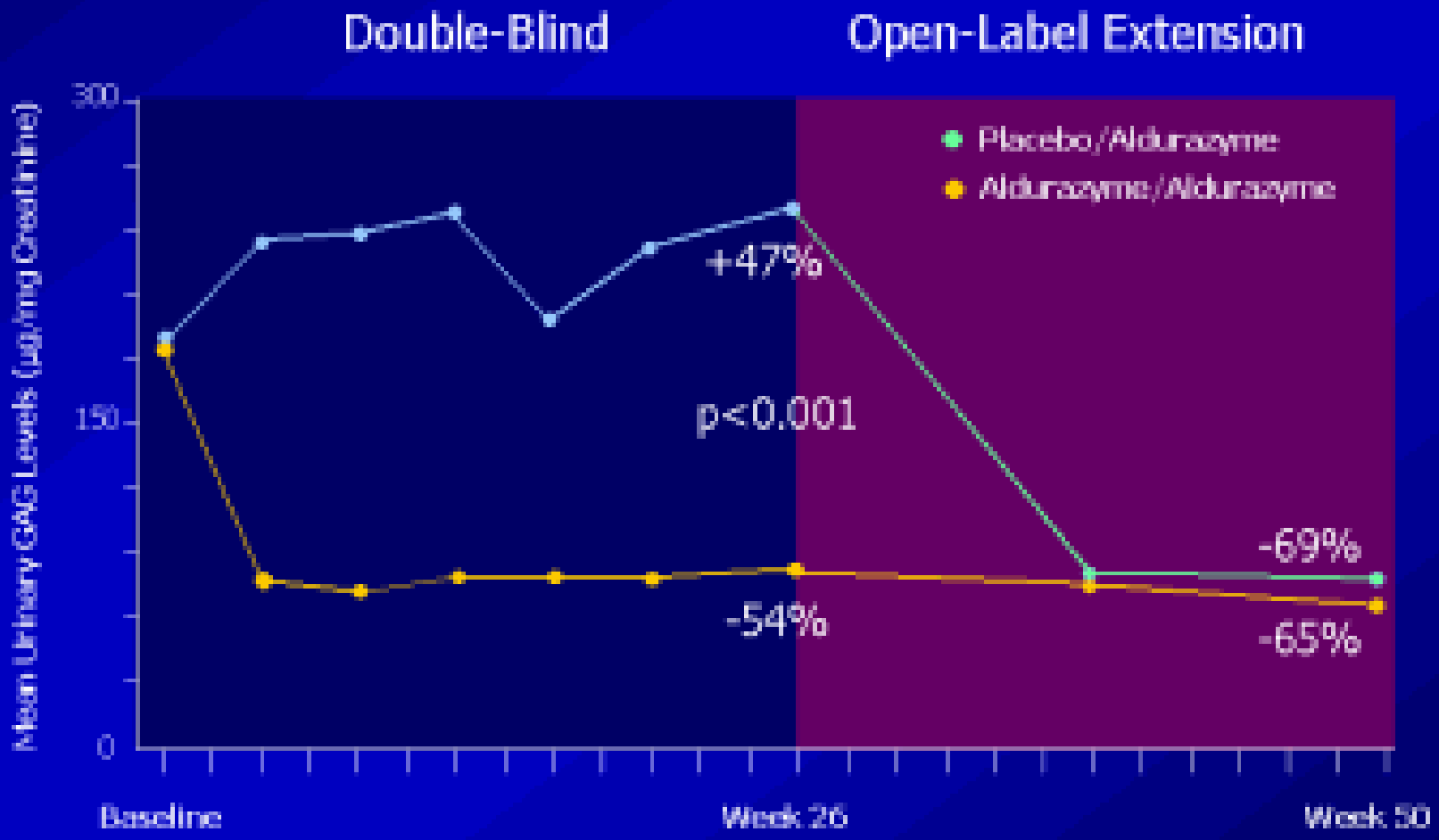
Responders

59% Aldurazyme
22% Placebo
(p=0.016)

Mean Net Change

1.0 Aldurazyme
-0.4 Placebo

Phase 3 Study: Aldurazyme Reduces Urinary GAG Levels



Example: asthma score

- Asthma: manifold symptoms, periodic worsening.
- Zhang, Song, Reiss (2004) proposed
  - PEF decrease >20%
  - 2+ puffs/day of beta-agnoist
  - Increase in symptom score > 50%
  - 3+ nighttime awakenings
  - PEF < 180 L/min
  - Hospital visit
- Showed good correlation with other global evaluations.

Two endpoints—setup

- Let X_1 and X_2 be two endpoints.
  - Two stroke scales, DBP & SBP, time to AIDS/Death & CD4.
- Let Z_1 and Z_2 be the associated standardized test statistics.
  - E.g. two tests of proportions, two t-tests, log-rank & t-test.
- Let p_1 and p_2 be the two p-values.
- Let's assume X_1 and X_2 are independent

Ranking generalization

- Compare each pair of treatment/control patients
- = 1 if “i” (in T) does better “j” (in P)
- Y_{ij} = 1/2 if same
- = 0 if “i” (in T) does worse “j” (in P)
- Can compare “i” & “j” over common followup.