



## Disclaimer:

Presentation slides from the Rare Disease Workshop Series are posted by the Kakkis EveryLife Foundation, for educational purposes only. They are for use by drug development professionals and statisticians, and are not to be used to guide the prescribing or use of any of the drugs mentioned in the slides. To obtain information on a particular drug, refer to the drug labeling. Do not reproduce or distribute the slides (full set or any portion of) without the permission of the author.



## RARE DISEASE WORKSHOP SERIES

Improving the Clinical Development Process

# Re-Analyses of Pivotal Trials for Approved Treatments in MPS

Mei Sheng Duh, ScD,  
James Signorovitch, PhD,  
Analysis Group Inc, Boston MA



# Objectives

- Apply a range of statistical tests to three trials of MPS treatments
- Compare estimated treatment effects and standard errors across tests
- Generate hypotheses – what kinds of tests have advantages for rare disease trials?



# Data

- Aldurazyme (N=45)
  - Visits at weeks 0, 4, 8, 12, 16, 20, 26
- Naglazyme (N=39)
  - Visits at weeks 0, 6, 12, 18, 24
- Elaprase (N=64)
  - Visits at weeks 0, 18, 38, 53



# Outcomes

- Change from baseline in 6-minute walk test (6MWT) in meters
- Change from baseline in forced vital capacity (FVC) in liters



# Methods:

## Tests Using Change to Last Visit

- T-test (Welch–Satterthwaite degrees of freedom)
- Wilcoxon rank sum test
- ANOVA (assuming equal variance)
- ANCOVA with adjustment for the baseline score for the outcome measure and
  - No further adjustment
  - Further adjustment for study center
  - Further adjustment for age and sex
  - Further adjustment for age, sex, height, weight, FVC, 6 and 12 minute walk test, apnea/hypopnea index (AHI), shoulder flexion



# Methods:

## Tests Using Change to Last Visit (cont'd)

- Augmentation (Tsiatis et al. 2008)
  - Considered all variables listed above as potential auxiliary variables
  - Selected variables for augmentation on each treatment arm by minimizing Mallows's  $C_p$
  - Two versions: limit to one selected variable per arm vs. an unlimited number of variables obtained via forward selection
  - Seeks to adjust for chance baseline imbalance and increase efficiency
  - Works by approximating the optimal semiparametric estimator
  - Provides a consistent estimate even if the augmentation is misspecified
  - Avoids post-hoc fishing for covariate adjustments and is unbiased



# Methods:

## Longitudinal Tests

Used generalized estimating equations and combined different specifications:

- Working correlation structure
  - Compound symmetry
  - Toeplitz
  - Unstructured
- Effects of time (includes main effects and interactions with treatment)
  - Categorical (separate effect for follow-up each visit)
  - Linear
  - Quadratic



# Methods:

## Longitudinal Tests (cont'd)

- Time of treatment effect assessment
  - Average across all visits
  - Assessed at endpoint
- Number of follow-up assessments included
  - All assessments
  - Last two assessments



# Methods:

## Longitudinal Tests (cont'd)

- Adjustment for baseline characteristics
  - None
  - Baseline score for the outcome measure
  - Baseline score for the outcome measure, age, sex
  - Baseline score for the outcome measure, age, sex, height, weight, FVC, 6 and 12 minute walk test, apnea/hypopnea index, shoulder flexion (as available)



# Methods:

## Aggregating Across Trials

- Mean treatment effects (signal) and standard errors (noise) were aggregated separately across trials
- The treatment effect mean and standard error within each trial were first standardized relative to the standard deviation of the outcome measure at baseline



# Methods: Other Details

- Signal and noise for the Wilcoxon test were computed from its Z-statistic and the median difference between treatment groups
- Missing outcomes were imputed by carrying the last observation forward
- Patients with missing baseline values were dropped as needed



# Results:

## Table 1. Baseline Variation

	6-Minute Walk Test (Meters)		
	Treatment Effect	Baseline Difference	Baseline SD
Trial 1	38	-48	124
Trial 2	53	-68	93
Trial 3	37	9	102



## Table 2. Tests Using Change to Last Visit

Model <sup>2</sup>	Treatment Effect Estimates <sup>1</sup>					
	6-Minute Walk Test			Forced Vital Capacity		
	Mean	SE	Z-Score	Mean	SE	Z-Score
1. T-Test	0.41	0.19	2.14	0.27	0.10	2.81
2. Wilcoxon	0.28	0.14	2.03	0.25	0.11	2.26
3. ANOVA	0.41	0.19	2.14	0.27	0.10	2.79
4. ANCOVA <sup>3</sup>	0.40	0.20	2.04	0.24	0.09	2.54
5. ANCOVA <sup>4</sup>	0.41	0.19	2.16	0.25	0.09	2.76
6. ANCOVA <sup>5</sup>	0.40	0.18	2.16	0.25	0.08	2.98
7. ANCOVA <sup>6</sup>	0.36	0.20	1.85	0.20	0.08	2.63
8. Augmentation (1) <sup>7</sup>	0.44	0.19	2.28	0.23	0.09	2.52
9. Augmentation (2) <sup>7</sup>	0.43	0.19	2.23	0.23	0.09	2.56

Notes:  
See handout for footnotes.



# Table 3. Longitudinal Models

## No Baseline Adjustment

Model	Specification				Treatment Effect Estimates <sup>5</sup>					
					6-Minute Walk Test			Forced Vital Capacity		
	Working Correlation <sup>1</sup>	Model for Effect of Time <sup>2</sup>	Time of Treatment Effect Assessment <sup>3</sup>	Included Visits <sup>4</sup>	Mean	SE	Z	Mean	SE	Z
10. Unstructured	Unstructured	None	Average	All	0.28	0.14	1.97	0.17	0.08	2.18
11. Toeplitz	Toeplitz	None	Average	All	0.28	0.15	1.91	0.18	0.07	2.45
12. CS	CS	None	Average	All	0.28	0.14	2.04	0.17	0.08	2.17
13. CS, Categorical	CS	Categorical	Last Visit	All	0.41	0.19	2.15	0.27	0.09	2.84
14. CS, Linear	CS	Linear	Last Visit	All	0.32	0.18	1.75	0.16	0.08	1.98
15. CS, Quadratic	CS	Quadratic	Last Visit	All	0.32	0.18	1.79	0.16	0.07	2.15

**Notes:**

See handout for footnotes.

CS = Compound Symmetry.

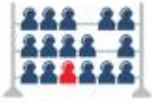


# Table 4. Longitudinal Models

*Adjusted for baseline score for the outcome measure*

Model	Specification				Treatment Effect Estimates <sup>5</sup>					
					6-Minute Walk Test			Forced Vital Capacity		
	Working Correlation <sup>1</sup>	Model for Effect of Time <sup>2</sup>	Time of Treatment Effect Assessment <sup>3</sup>	Included Visits <sup>4</sup>	Mean	SE	Z	Mean	SE	Z
16. Unstructured	Unstructured	None	Average	All	0.29	0.13	2.14	0.18	0.06	3.16
17. Toeplitz	Toeplitz	None	Average	All	0.29	0.14	2.10	0.17	0.06	2.99
18. CS	CS	None	Average	All	0.29	0.13	2.23	0.17	0.06	3.06
19. CS, Categorical	CS	Categorical	Last Visit	All	0.42	0.18	2.33	0.26	0.08	3.12
20. CS, Linear	CS	Linear	Last Visit	All	0.32	0.17	1.92	0.15	0.06	2.31
21. CS, Quadratic	CS	Quadratic	Last Visit	All	0.33	0.16	1.98	0.17	0.07	2.45
22. Unstr., Last Two	Unstructured	None	Average	Last Two	0.37	0.16	2.35	0.18	0.08	2.32
23. CS, Last Two	CS	None	Average	Last Two	0.37	0.16	2.29	0.18	0.08	2.34

Notes:  
See handout for footnotes.  
CS = Compound Symmetry.



# Table 5. Longitudinal Models

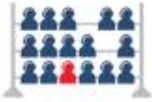
*Adjusted for baseline score for the outcome measure, age and gender*

Model	Specification				Treatment Effect Estimates <sup>5</sup>					
					6-Minute Walk Test			Forced Vital Capacity		
	Working Correlation <sup>1</sup>	Model for Effect of Time <sup>2</sup>	Time of Treatment Effect Assessment <sup>3</sup>	Included Visits <sup>4</sup>	Mean	SE	Z	Mean	SE	Z
24. Unstructured	Unstructured	None	Average	All	0.28	0.14	2.03	0.16	0.06	2.73
25. Toeplitz	Toeplitz	None	Average	All	0.28	0.14	1.98	0.17	0.06	2.93
26. CS	CS	None	Average	All	0.27	0.14	1.92	0.16	0.06	2.72
27. CS, Categorical	CS	Categorical	Last Visit	All	0.41	0.19	2.10	0.25	0.08	3.00
28. CS, Linear	CS	Linear	Last Visit	All	0.32	0.17	1.86	0.15	0.06	2.29
29. CS, Quadratic	CS	Quadratic	Last Visit	All	0.32	0.17	1.88	0.14	0.06	2.21

Notes:

See handout for footnotes.

CS = Compound Symmetry.



## Table 6. Longitudinal Models

*Adjusted for baseline score for the outcome measure, age, gender, weight, height, 6-minute walk test, forced vital capacity, apnea/hypopnea index and shoulder flexion*

Model	Specification				Treatment Effect Estimates <sup>5</sup>					
					6-Minute Walk Test			Forced Vital Capacity		
	Working Correlation <sup>1</sup>	Model for Effect of Time <sup>2</sup>	Time of Treatment Effect Assessment <sup>3</sup>	Included Visits <sup>4</sup>	Mean	SE	Z	Mean	SE	Z
30. Unstructured	Unstructured	None	Average	All	0.27	0.14	1.97	0.12	0.06	1.98
31. Toeplitz	Toeplitz	None	Average	All	0.27	0.14	1.86	0.13	0.06	2.15
32. CS	CS	None	Average	All	0.27	0.14	1.98	0.12	0.06	1.99
33. CS, Categorical	CS	Categorical	Last Visit	All	0.39	0.18	2.13	0.22	0.06	3.44
34. CS, Linear	CS	Linear	Last Visit	All	0.31	0.17	1.86	0.11	0.06	1.97
35. CS, Quadratic	CS	Quadratic	Last Visit	All	0.31	0.17	1.77	0.11	0.07	1.76

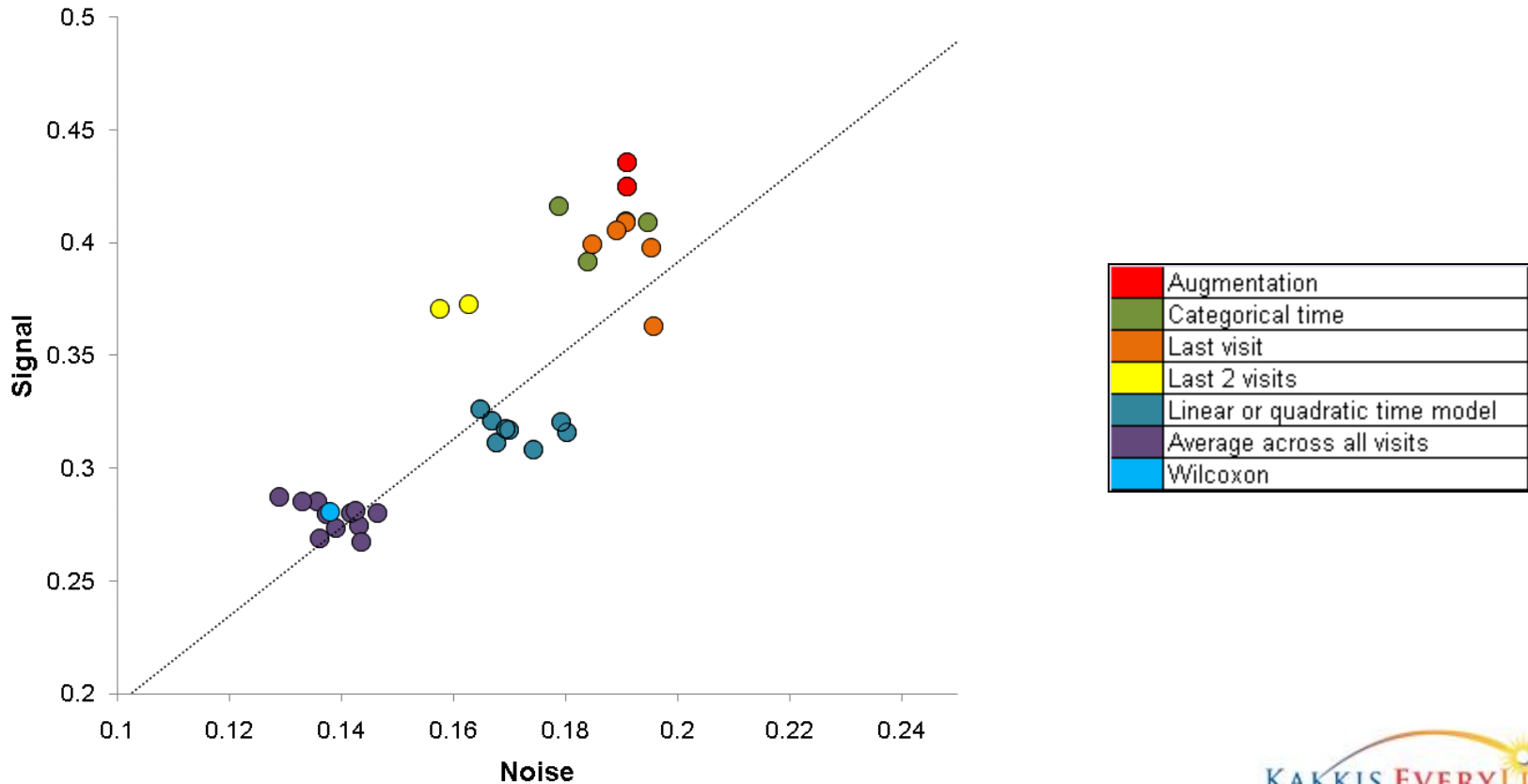
**Notes:**

See handout for footnotes.

CS = Compound Symmetry.



# Figure 1. Signal vs. Noise Across Models *6-Minute Walk Test*



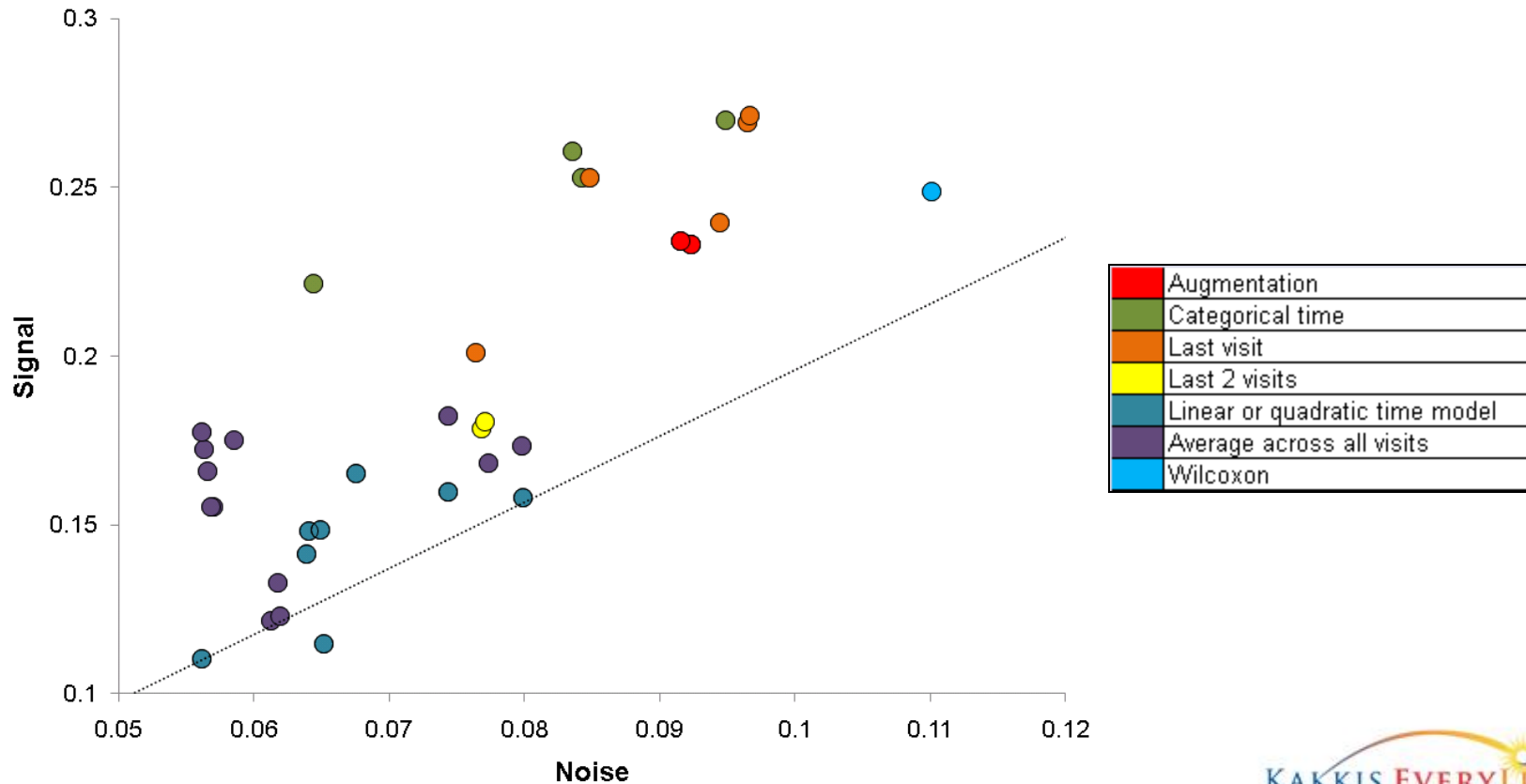


## Observations for 6MWT

- Tests based only on change to last visit performed well for most baseline adjustments; compared to these tests:
  - Augmentation helped increase signal vs. noise
  - Longitudinal models with categorical effects of time preserved signal vs. noise across baseline adjustments
  - Looking only at the last 2 visits increased signal vs. noise
  - Linear and quadratic effects of time reduced signal vs. noise
  - Averaging treatment effects across all visits reduced noise and signal
  - The Wilcoxon test detected low signal and low noise relative to others



## Figure 2. Signal vs. Noise Across Models *Forced Vital Capacity*





## Observations for FVC

- Tests based only on change to last visit performed well. Compared to these tests:
  - Augmentation was not helpful or harmful
  - Longitudinal models with categorical effects of time often increased signal vs. noise
  - Looking at the last 2 visits decreased signal vs. noise
  - Linear and quadratic effects of time decreased signal vs. noise
  - Averaging treatment effects across all visits reduced noise and signal
  - Among tests using only the last visit, the Wilcoxon had the least signal vs. noise



# Table 7. Longitudinal Models

*Unstructured working correlation, various baseline adjustments*

Model	Specification				Treatment Effect Estimates <sup>5</sup>					
					6-Minute Walk Test			Forced Vital Capacity		
	Working Correlation <sup>1</sup>	Model for Effect of Time <sup>2</sup>	Time of Treatment Effect Assessment <sup>3</sup>	Included Visits <sup>4</sup>	Mean	SE	Z	Mean	SE	Z
10. No adjustment	Unstructured	None	Average	All	0.28	0.14	1.97	0.17	0.08	2.18
16. Adjustment 1	Unstructured	None	Average	All	0.29	0.13	2.14	0.18	0.06	3.16
24. Adjustment 2	Unstructured	None	Average	All	0.28	0.14	2.03	0.16	0.06	2.73
30. Adjustment 3	Unstructured	None	Average	All	0.27	0.14	1.97	0.12	0.06	1.98

**Notes:**

*See handout for footnotes.*

*Adjustment 1 = baseline outcome score.*

*Adjustment 2 = baseline outcome score, age and sex.*

*Adjustment 3 = baseline outcome score, age, sex, weight, height, 6-minute walk test, forced vital capacity, apnea/hypopnea index and shoulder flexion (as available)*



## Table 8. Longitudinal Models

*Compound symmetry working correlation, various baseline adjustments*

Model	Specification				Treatment Effect Estimates <sup>5</sup>					
	Working Correlation <sup>1</sup>	Model for Effect of Time <sup>2</sup>	Time of Treatment Effect Assessment <sup>3</sup>	Included Visits <sup>4</sup>	6-Minute Walk Test			Forced Vital Capacity		
					Mean	SE	Z	Mean	SE	Z
12. No Adjustment	CS	None	Average	All	0.28	0.14	2.04	0.17	0.08	2.17
18. Adjustment 1	CS	None	Average	All	0.29	0.13	2.23	0.17	0.06	3.06
26. Adjustment 2	CS	None	Average	All	0.27	0.14	1.92	0.16	0.06	2.72
32. Adjustment 3	CS	None	Average	All	0.27	0.14	1.98	0.12	0.06	1.99

**Notes:**

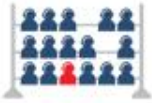
See *handout for footnotes.*

CS = *Compound Symmetry.*

Adjustment 1 = *baseline outcome score.*

Adjustment 2 = *baseline outcome score, age and sex.*

Adjustment 3 = *baseline outcome score, age, sex, weight, height, 6-minute walk test, forced vital capacity, apnea/hypopnea index and shoulder flexion (as available)*



## Table 9. Longitudinal Models

*Categorical effect of time, various baseline adjustments*

Model	Specification				Treatment Effect Estimates <sup>5</sup>					
	Working Correlation <sup>1</sup>	Model for Effect of Time <sup>2</sup>	Time of Treatment Effect Assessment <sup>3</sup>	Included Visits <sup>4</sup>	6-Minute Walk Test			Forced Vital Capacity		
					Mean	SE	Z	Mean	SE	Z
12. No Adjustment	CS	Categorical	Last Visit	All	0.41	0.19	2.15	0.27	0.09	2.84
18. Adjustment 1	CS	Categorical	Last Visit	All	0.42	0.18	2.33	0.26	0.08	3.12
26. Adjustment 2	CS	Categorical	Last Visit	All	0.41	0.19	2.10	0.25	0.08	3.00
32. Adjustment 3	CS	Categorical	Last Visit	All	0.39	0.18	2.13	0.22	0.06	3.44

**Notes:**

See handout for footnotes.

CS = Compound Symmetry.

Adjustment 1 = baseline outcome score.

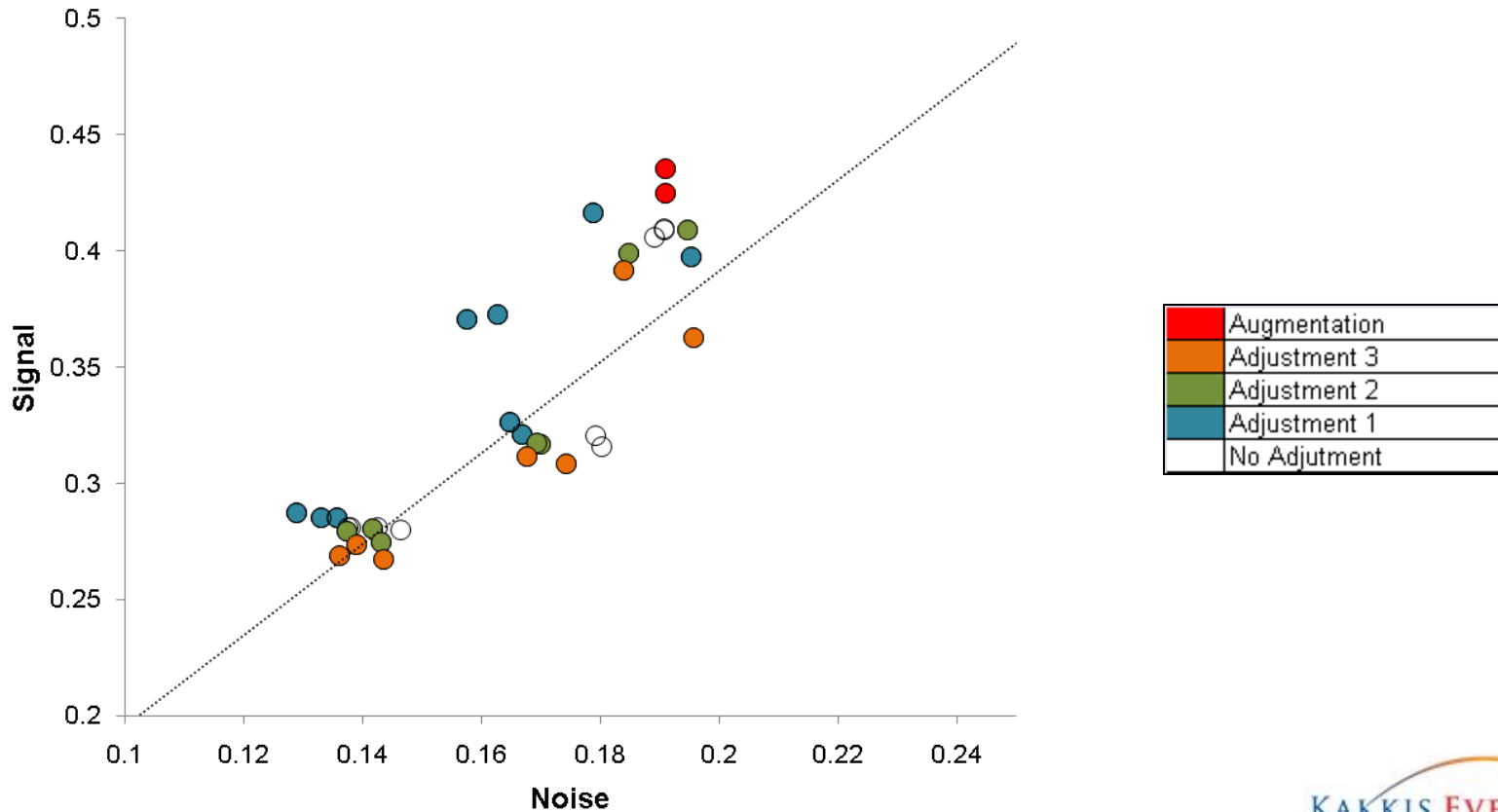
Adjustment 2 = baseline outcome score, age and sex.

Adjustment 3 = baseline outcome score, age, sex, weight, height, 6-minute walk test, forced vital capacity, apnea/hypopnea index and shoulder flexion (as available)



# Figure 3. Signal vs. Noise Across Models

## 6-Minute Walk Test





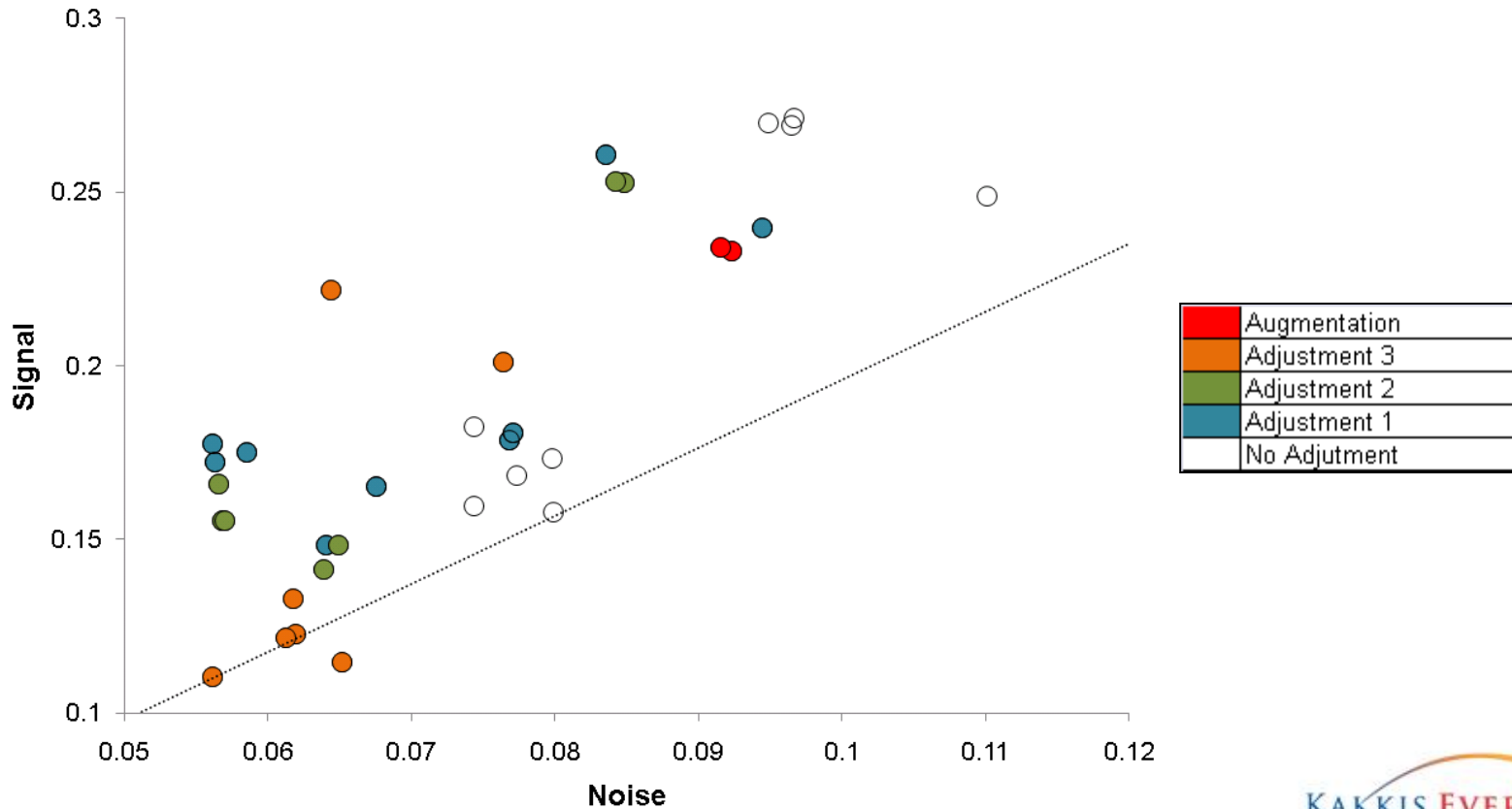
## Observations for 6MWT

- In models using only change to last visit, adjustment for baseline characteristics had mixed impacts on signal vs. noise
- In longitudinal models, adjusting for baseline scores was often better than no adjustment, but further adjustment tended to decrease signal and increase noise
- Augmentation performed better than or similar to ANCOVA adjusted only for the baseline score



# Figure 4. Signal vs. Noise Across Models

## *Forced Vital Capacity*



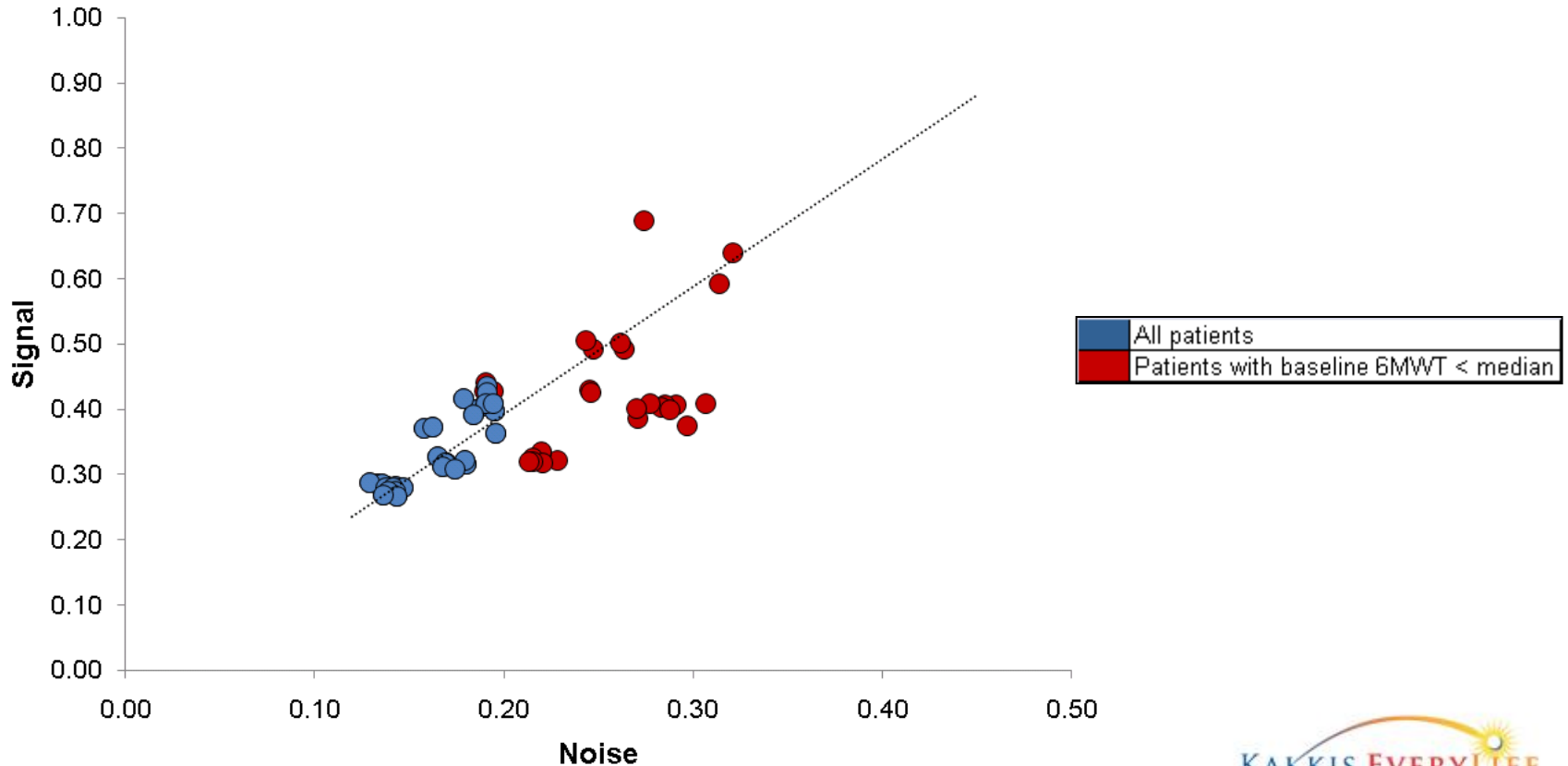


## Observations for FVC

- In models using only change to last visit, adjustment for baseline characteristics had mixed impacts on signal vs. noise
- In longitudinal models, adjusting for baseline scores was often better than no adjustment, but further adjustment tended to decrease signal and increase noise
- Augmentation performed similarly to ANCOVA adjustment for the baseline score



## Figure 5. Signal vs. Noise Across Models *6-Minute Walk Test*





**Table 10. Tests Using Change to Last Visit for 6MWT**  
*All patients vs. the subgroup with baseline 6MWT < median*

Model <sup>2</sup>	6MWT Treatment Effect Estimates <sup>1</sup>					
	All Patients			Subgroup		
	Mean	SE	Z-Score	Mean	SE	Z-Score
1. T-Test	0.41	0.19	2.14	0.41	0.31	1.33
2. Wilcoxon	0.28	0.14	2.03	0.32	0.23	1.40
3. ANOVA	0.41	0.19	2.14	0.41	0.29	1.39
4. ANCOVA <sup>3</sup>	0.40	0.20	2.04	0.37	0.30	1.26
5. ANCOVA <sup>4</sup>	0.41	0.19	2.16	0.48	0.36	1.34
6. ANCOVA <sup>5</sup>	0.40	0.18	2.16	0.49	0.26	1.87
7. ANCOVA <sup>6</sup>	0.36	0.20	1.85	0.69	0.27	2.51
8. Augmentation (1) <sup>7</sup>	0.44	0.19	2.28	0.59	0.31	1.89
9. Augmentation (2) <sup>7</sup>	0.43	0.19	2.23	0.64	0.32	1.99

Notes:  
See handout for footnotes.



## Table 11. Longitudinal Models for 6MWT

*All patients vs. the subgroup with baseline 6MWT < median  
Adjusted for baseline score for the outcome measure*

Model	Specification				6MWT Treatment Effect Estimates <sup>5</sup>					
					All Patients			Subgroup		
	Working Correlation <sup>1</sup>	Model for Effect of Time <sup>2</sup>	Time of Treatment Effect Assessment <sup>3</sup>	Included Visits <sup>4</sup>	Mean	SE	Z	Mean	SE	Z
16. Unstructured	Unstructured	None	Average	All	0.29	0.13	2.14	0.43	0.19	2.25
17. Toeplitz	Toeplitz	None	Average	All	0.29	0.14	2.10	0.44	0.19	2.32
18. CS	CS	None	Average	All	0.29	0.13	2.23	0.43	0.19	2.20
19. CS, Categorical	CS	Categorical	Last Visit	All	0.42	0.18	2.33	0.50	0.26	1.91
20. CS, Linear	CS	Linear	Last Visit	All	0.32	0.17	1.92	0.49	0.25	1.99
21. CS, Quadratic	CS	Quadratic	Last Visit	All	0.33	0.16	1.98	0.51	0.24	2.08
22. Unstr., Last Two	Unstructured	None	Average	Last Two	0.37	0.16	2.35	0.43	0.25	1.75
23. CS, Last Two	CS	None	Average	Last Two	0.37	0.16	2.29	0.43	0.25	1.73

Notes:

See handout for footnotes.

CS = Compound Symmetry.



# Observations for 6MWT Subgroup Analyses

- Restriction to the most impaired subgroup at baseline did not substantially affect the unadjusted treatment effect in analyses of change to last visit
- Baseline adjustment tended to increase the treatment effect in the subgroup
- Restriction to the most impaired subgroup at baseline did increase the estimated treatment effect in longitudinal analyses
- Estimating the average effect across all visits, or using a linear or quadratic effect of time, tended to increase signal vs. noise compared to looking at the last two visits or allowing a more flexible model for time



# Summary

- Baseline adjustment can have a mixed impact on signal vs. noise
- Augmentation was 'safe' compared to ANCOVA – performance was similar to or better than ANCOVA adjusted only for baseline score; ANCOVAs with further adjustment were both better and worse
- Averaging the treatment effect across all study visits tended to decrease both signal and noise
- Using all study visits and including a categorical effect of time performed well for both endpoints; linear and quadratic effects of time were not helpful on average
- Increased treatment effects in the most impaired subgroup were more apparent with baseline adjustment or longitudinal models



# Limitations

- Purely statistical
- Exploratory
- May not generalize beyond the studied data



# These trials provide a resource for studying methods for rare disease trials

- Can systematically try out methods side by side
- Can go beyond purely statistical investigations since trials have similar designs and outcome measures
- Can generate hypotheses that can be tested in simulations or new trials



# Topics for Discussion

- How to optimize the following in settings with small samples and extreme baseline heterogeneity?
  - Baseline adjustment
  - Incorporation of longitudinal data
  - Analysis population / responder definition for different outcomes
- Opportunities to use blinded data while preserving type I error?