Statistical properties of hypothesis tests using Goal Attainment Scaling

Susanne Urach
Section for Medical Statistics, Medical University of Vienna

Joint Conference on Biometrics & Biopharmaceutical Statistics (CEN ISBS), Aug 28 - Sep 1, 2017, Vienna, Austria

joint work with Gaasterland C.M.W., Rosenkranz G., Jilma B., Roes K., Van der Lee J.H., Posch M., Ristl R.

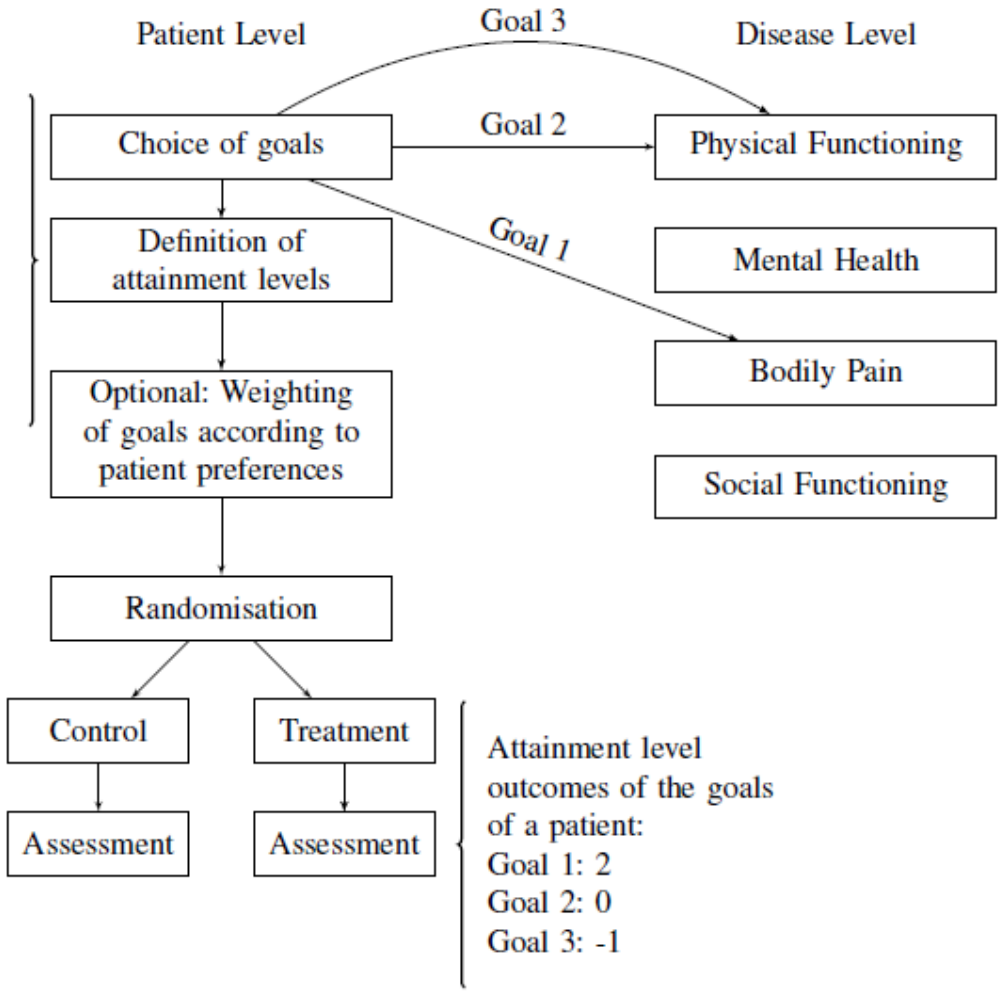
This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement number FP HEALTH 2013-603160. ASTERIX Project - http://www.asterix-fp7.eu/

Motivation: Duchenne muscular dystrophy (DMD)

- Duchenne results from defects in the X gene for dystrophin, a structural protein required to maintain muscle integrity.
- It affects 1/3300 males and is considered an orphan disease.
- Disease with very heterogeneous courses or stages:
  1. first signs: abnormal ambulation due to proximal muscle weakness
  2. 8 years: falling, standing up from supine or climbing stairs
  3. 10-14 years: restricted to a wheelchair
- Symptoms differ substantially between patients:
  - walking abnormalities
  - elbow/knee flexion/extension, shoulder abduction
  - endurance
  - cardiorespiratory status
- No standardized outcome measure applicable to all available: e.g. 6-min Walk Test restricted to patients without a wheelchair

Process of goal setting and measurement

Attainment level definitions
for goal "walking":
-2 unable to walk
-1 can walk for 3 steps
0 can walk for 5 minutes
+1 can walk for 15 minutes
+2 can walk for a longer period



Advantages and disadvantages of GAS

Goal attainment scaling (GAS) is a patient centered outcome measure capturing the treatment effect across manifestations:

- Advantages:
  - Increase in the relevance of the endpoint to the patient
  - Higher possible sample sizes for the clinical trials: patients with very heterogenous symptoms can be included because the endpoint is individualized

- Disadvantages:
  - Process of goal setting time consuming
  - Not a validated measurement instrument
  - Clinicians are unsure about the concept GAS measures
  - Only inferences on a global effect are possible, but not on the individual endpoints

Research questions

The flexibility in the choice and number of goals provides several statistical challenges in the analysis and interpretation of results:

- Analyzing trials:
  - How to test for a treatment effect in an optimal way?
  - Interpretation of significant hypothesis test?
  - What kind of weights should be applied to the individual goals?

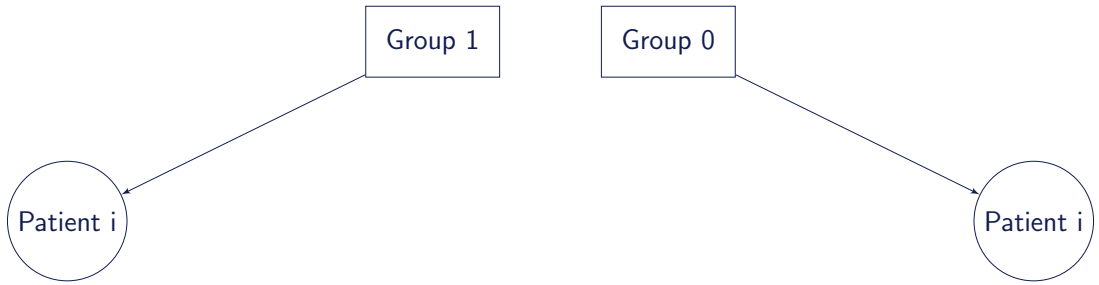
- Designing trials:
  How is a hypothesis test using a GAS endpoint affected by
  - Maximum number of goals
  - Correlation between the goals
  - Proportion of goals affected by the treatment
  - Number of attainment levels

Multilevel hierarchical model

- Randomized parallel group comparison between two arms

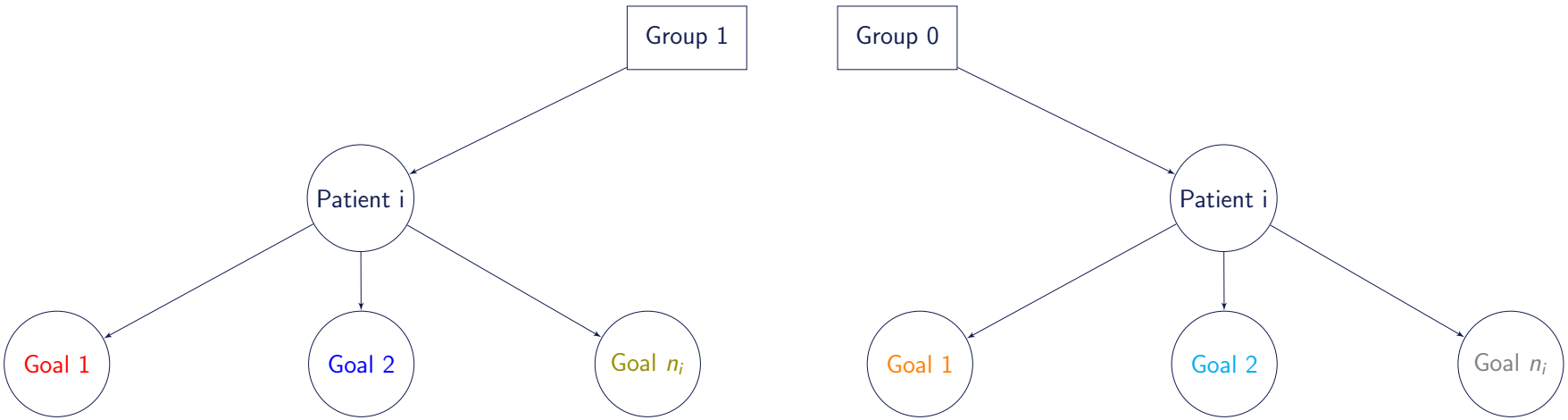
Multilevel hierarchical model

- Randomized parallel group comparison between two arms
- The patients are clustered within the groups.



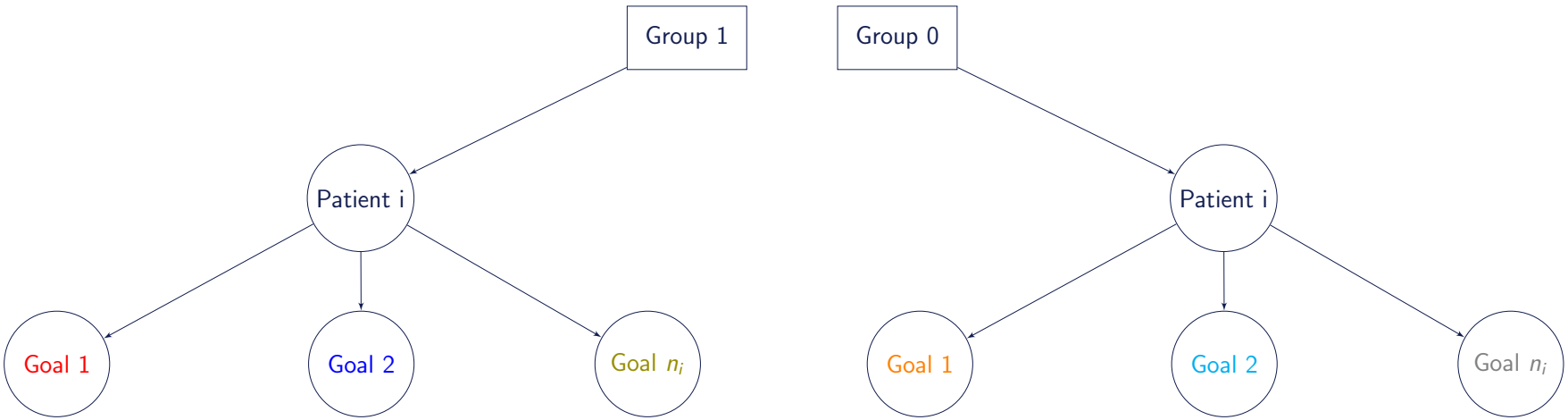
Multilevel hierarchical model

- Randomized parallel group comparison between two arms
- The patients are clustered within the groups.
- Goal outcomes are clustered and equi-correlated within patients.
- Patients individually choose number and kind of goals
- Latent continuous goal attainment score for goal k of subject i: Y_{ik}



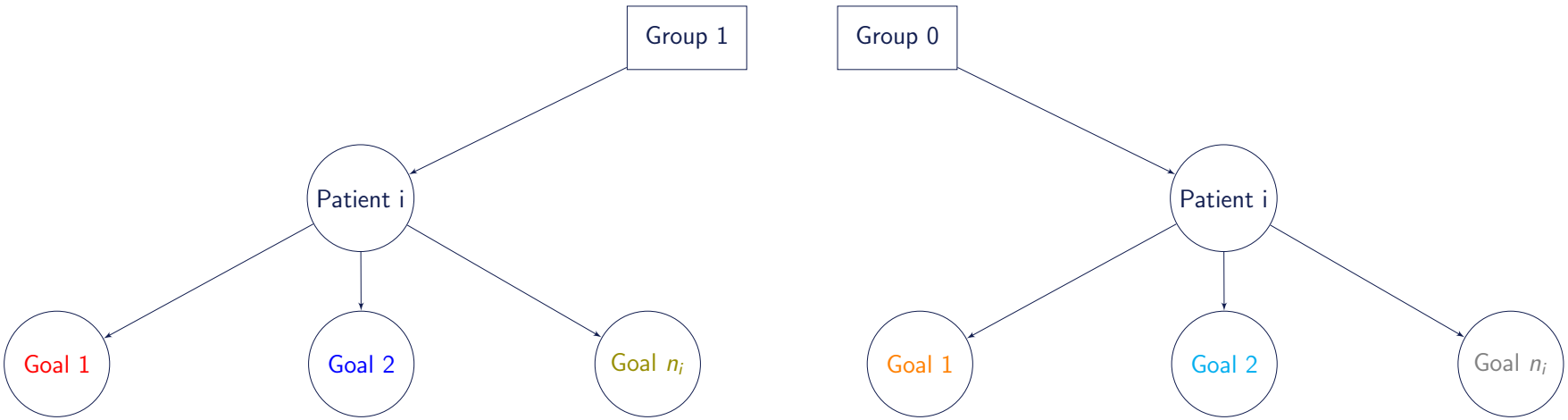
Multilevel hierarchical model

- Randomized parallel group comparison between two arms
- The patients are clustered within the groups.
- Goal outcomes are clustered and equi-correlated within patients.
- Patients individually choose number and kind of goals
- Latent continuous goal attainment score for goal k of subject i: Y_{ik}



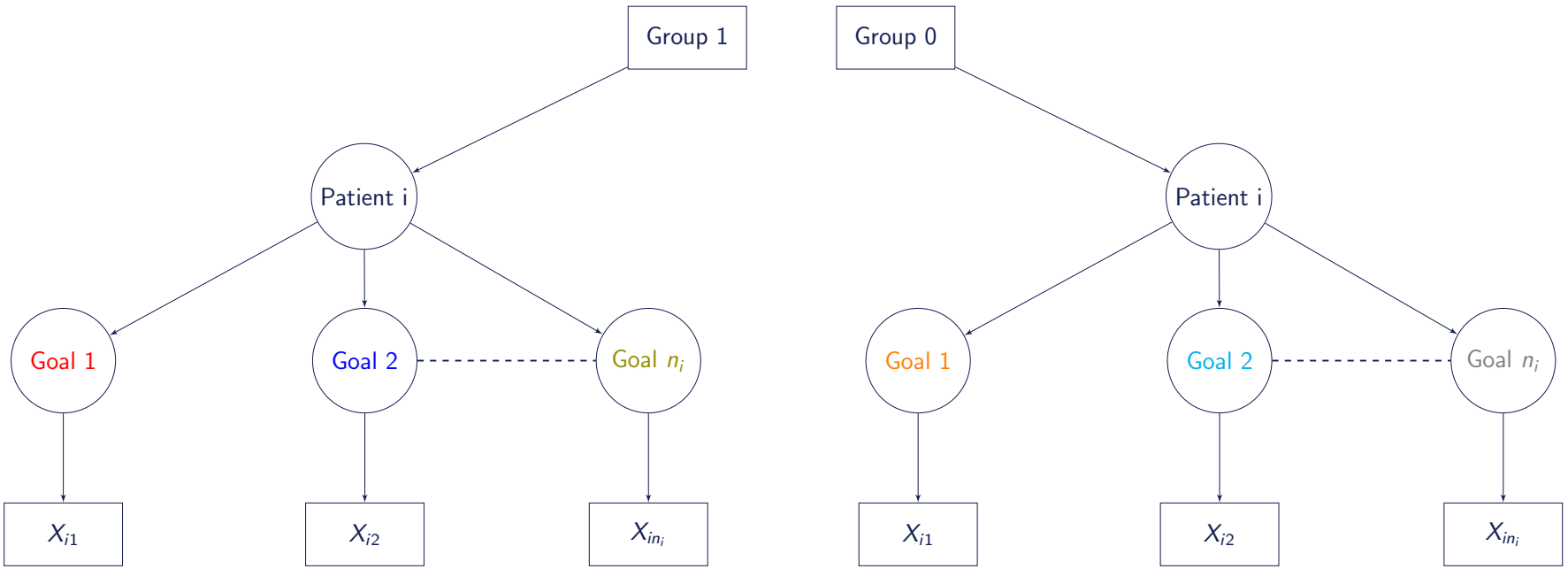
Multilevel hierarchical model

- Randomized parallel group comparison between two arms
- The patients are clustered within the groups.
- Goal outcomes are clustered and equi-correlated within patients.
- Patients individually choose number and kind of goals
- Latent continuous goal attainment score for goal k of subject i: Y_{ik}



Multilevel hierarchical model

- Randomized parallel group comparison between two arms
- The patients are clustered within the groups.
- Goal outcomes are clustered and equi-correlated within patients.
- Patients individually choose number and kind of goals
- Latent continuous goal attainment score for goal k of subject i: Y_{ik}
- Discretization of the latent continuous normal variables Y_{ik} via same set of thresholds $\rightarrow$ observed ordinal goal attainment level X_{ik}



Generating clustered ordinal outcomes

Random effect model for latent continuous goal outcome

$i = 1, \ldots, m$, with per group sample size m
$k = 1, \ldots, n_i$, with number of goals $n_i \sim G$ of patient i

$$Y_{ik} = u_i + g_i b_{ik} + \epsilon_{ik}$$

Y_{ik} ... latent continuous outcome for goal k of patient i
u_i ... random patient effect, $u_i \sim N(0, \sigma_u^2)$
g_i ... treatment group indicator, $g_i = 0, 1$
b_{ik} ... random treatment effect on goal k of patient i, $b_{ik} \sim F$
with $E(b_{ik}) = \delta$ and $Var(b_{ik}) = \sigma_b^2$
ϵ_{ik} ... random error term $\epsilon_{ik} \sim N(0, 1)$

The expected goal attainment in the treatment group $E(Y_{ik}) = \delta$ can be interpreted as the average treatment effect on the different goals.

Analysis of GAS data

Null hypothesis $H_0 : E(X_{ik}^1) \leq E(X_{ik}^0)$

The average goal attainment level of the experimental group is less or equal to the average goal attainment level in the control group.

Challenges:

- Clustered observations:
  Since goal attainment levels from within patients tend to be more alike than observations from different patients, those observations provide less information about a group.

- Different number of goals per patient:
  Less correlated or more goals of a patient provide more information about the overall treatment effect.

Wald tests based on estimates of $E(X_{ik}^g)$

Weighting of the contribution of each patient to the overall test statistic:

- Two sample t-test on per-subject means $\bar{X}_i^g = \frac{1}{n_i} \sum_{k=1}^{n_i} X_{ik}^g$
  - underestimates the standard deviation of X_{ik}^g
  - $\bar{X}_i^g$ only approximately normally distributed
  - t test very robust against deviations from the normal distribution
  - alternatively one could also apply a Mann-Whitney-U Test
- Two sample t-test on Kiresuk and Sherman T scores:
  - ordinary least square (OLS) estimator of $E(X_{ik}^g)$= sum of standardised mean goal attainment levels
  - assumed average correlation
- Generalised estimation equation (GEE) approach:
  - calculates generalized least square (GLS) estimator of $E(X_{ik}^g)$= minimum variance unbiased estimator = attainment levels are weighted by the inverse of the covariance matrix
  - estimates unknown covariance matrix using working correlation structure

Wald tests based on estimates of $E(X_{ik}^g)$

Weighting of the contribution of each patient to the overall test statistic:

- Two sample t-test on per-subject means $\bar{X}_i^g = \frac{1}{n_i} \sum_{k=1}^{n_i} X_{ik}^g$
  - underestimates the standard deviation of X_{ik}^g
  - $\bar{X}_i^g$ only approximately normally distributed
  - t test very robust against deviations from the normal distribution
  - alternatively one could also apply a Mann-Whitney-U Test
- Two sample t-test on Kiresuk and Sherman T scores:
  - ordinary least square (OLS) estimator of $E(X_{ik}^g)$= sum of standardised mean goal attainment levels
  - assumed average correlation
- Generalised estimation equation (GEE) approach:
  - calculates generalized least square (GLS) estimator of $E(X_{ik}^g)$= minimum variance unbiased estimator = attainment levels are weighted by the inverse of the covariance matrix
  - estimates unknown covariance matrix using working correlation structure

Wald tests based on estimates of $E(X_{ik}^g)$

Weighting of the contribution of each patient to the overall test statistic:

- Two sample t-test on per-subject means $\bar{X}_i^g = \frac{1}{n_i} \sum_{k=1}^{n_i} X_{ik}^g$
  - underestimates the standard deviation of X_{ik}^g
  - $\bar{X}_i^g$ only approximately normally distributed
  - t test very robust against deviations from the normal distribution
  - alternatively one could also apply a Mann-Whitney-U Test
- Two sample t-test on Kiresuk and Sherman T scores:
  - ordinary least square (OLS) estimator of $E(X_{ik}^g)=$ sum of standardised mean goal attainment levels
  - assumed average correlation
- Generalised estimation equation (GEE) approach:
  - calculates generalized least square (GLS) estimator of $E(X_{ik}^g)=$ minimum variance unbiased estimator = attainment levels are weighted by the inverse of the covariance matrix
  - estimates unknown covariance matrix using working correlation structure

Comparison of GEE and Kiresuk method

If we assume equal correlations ρ for all pairs $(X_{ik}, X_{ik'})$, $k \neq k'$:

Kiresuk method

$$\frac{\bar{T} - 50}{10} = \frac{1}{m} \sum_{i=1}^{m} \sqrt{\frac{n_i}{1 + (n_i - 1)\rho}} \bar{X}_i$$

Sum of the standardised mean goal attainment levels.

GEE method

$J = (1, ..., 1)$, $J' = (1, ..., 1)^T$

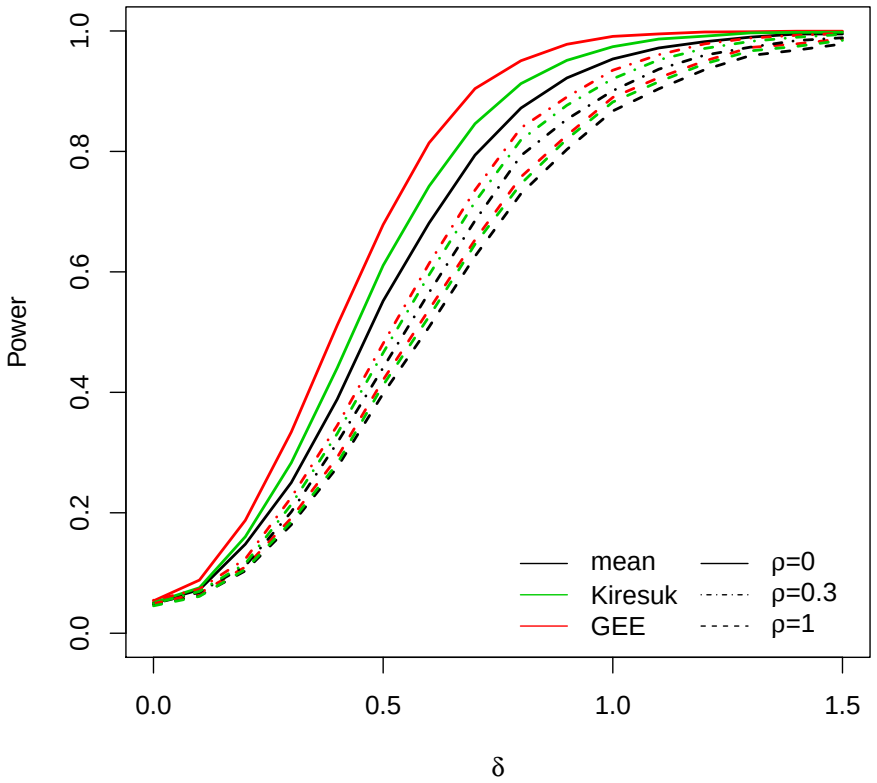
$$\frac{J'\Sigma^{-1}X}{J'\Sigma J} = \frac{\sum_{i=1}^{m} \frac{n_i}{1+(n_i-1)\rho} \bar{X}_i}{\sum_{i=1}^{m} \frac{n_i}{1+(n_i-1)\rho}}$$

Weighting the goal attainment levels with the inverse of the covariance matrix Σ^{-1}.

- GEE approach calculates grand mean for $\rho = 0$ and both Kiresuk and GEE reduce to the mean of per-subject means for $\rho = 1$.
- In both cases the means are weighted accounting for the different number of goals and the correlation between them.

Power of the hypothesis test: GEE vs Kiresuk

The GEE approach has better power for testing $E(X_{ik}^1) \leq E(X_{ik}^0)$:



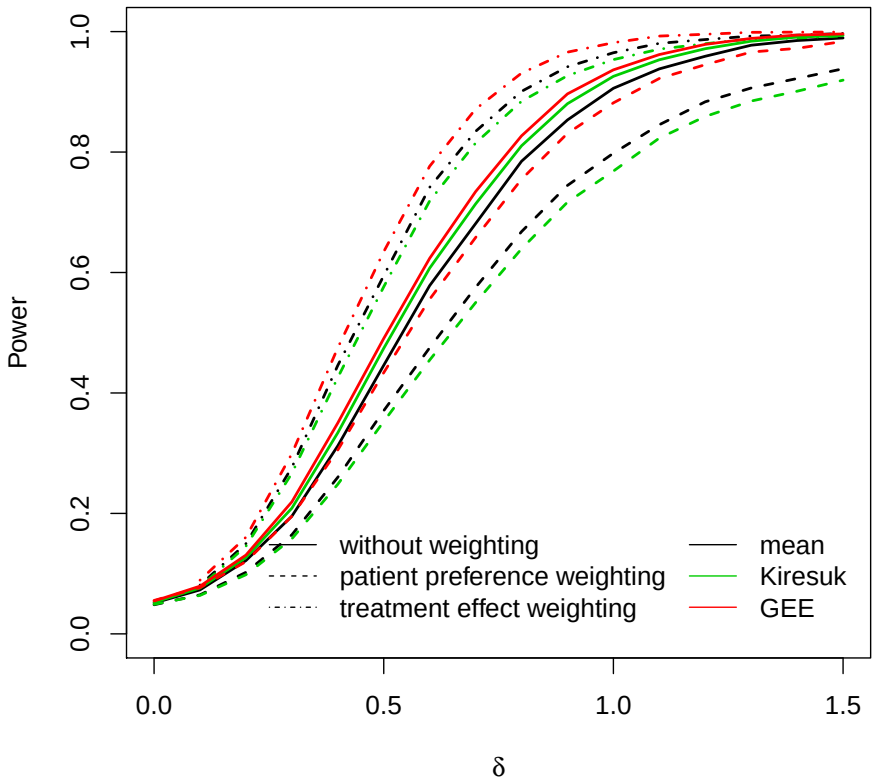
Power, $\delta = 0.5$, $\rho = 0$

GEE: 68%
Kiresuk: 61%
mean: 55.4%

m=20, $n_{max} = 5$, thresholds $c_j = \Phi^{-1}(p_j)$, $p = (0.2, 0.4, 0.6, 0.8)$
$n_{gi} \sim U\{1, \ldots, n_{max}\}$, $b_{ik} \sim U(0, 2\delta)$

Weighting of goal attainment outcomes

If the weights are not correlated with the treatment effect on the goals, weighting leads to a substantial loss in power.



Power, $\delta = 0.5$, $\rho = 0$

GEE

without weighting: 68%

preference weighting: 57%

effect weighting: 79%

Kiresuk

without weighting: 61%

preference weighting: 51%

effect weighting: 75%

mean

without weighting: 55%

preference weighting: 53%

effect weighting: 76%

Impact of design aspects on power

- The power increases with the number of goals affected by the treatment, but the increase levels off: For weak correlation between goals, there can be substantial power increase up to about 5 goals.
- If goals chosen by a patient are very similar, the gain in power by adding goals is small.
- Including goals that are not affected by the treatment can lead to a substantial loss in power.
- A scale with 5 levels appears to be sufficient. Further increasing the number of level has little influence on the power.

Conclusions

- The optimal way to test for a change in average goal attainment levels between groups would be to use the GEE approach.
- Using weights for the goal attainment levels which are not correlated with the treatment effect reduces power.
- The statistical implications of design choices (as, e.g., the maximum number of goals) should be considered.
- When presenting the results, the individual goals chosen should be investigated as well, maybe for certain domain clusters.

References

Agresti, A. and M. Kateri (2011).
Categorical data analysis.
Springer.

Hedeker, D. and R. D. Gibbons (1994).
A random-effects ordinal regression model for multilevel analysis.
Biometrics, 933–944.

Kiresuk, T. J. and M. R. E. Sherman (1968).
Goal attainment scaling: A general method for evaluating comprehensive community mental health programs.
Community mental health journal 4(6), 443–453.

Liang, K.-Y. and S. L. Zeger (1986).
Longitudinal data analysis using generalized linear models.
Biometrika, 13–22.