Not all patients are created equal, but are there subgroups that are more homogenous?

PSI webinar – 29 November 2018

Alexander Schacht, PhD

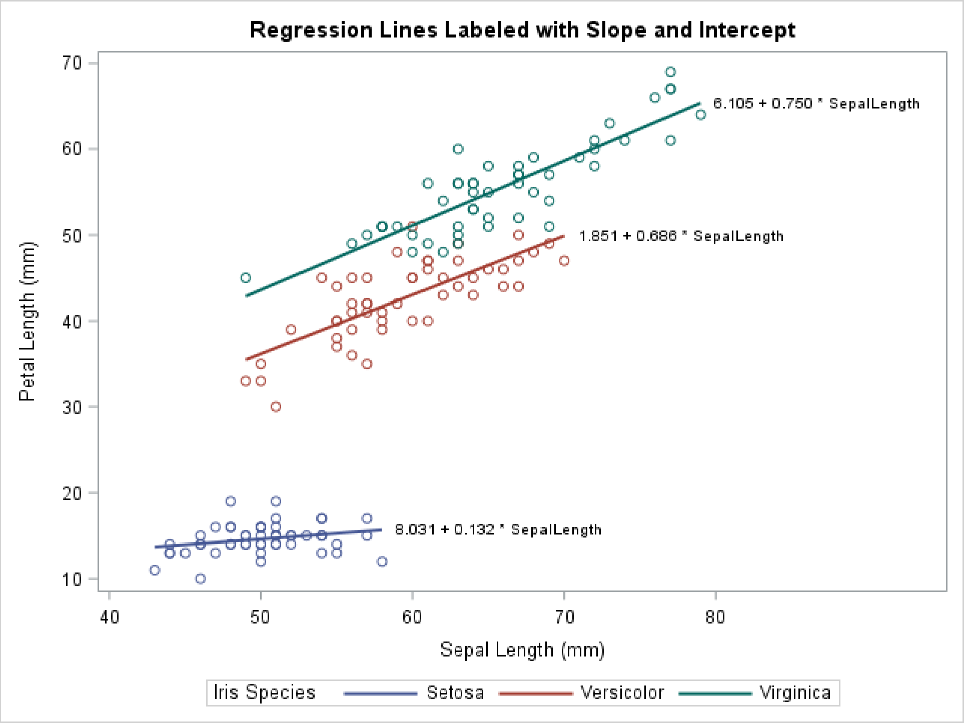
Do you wonder about

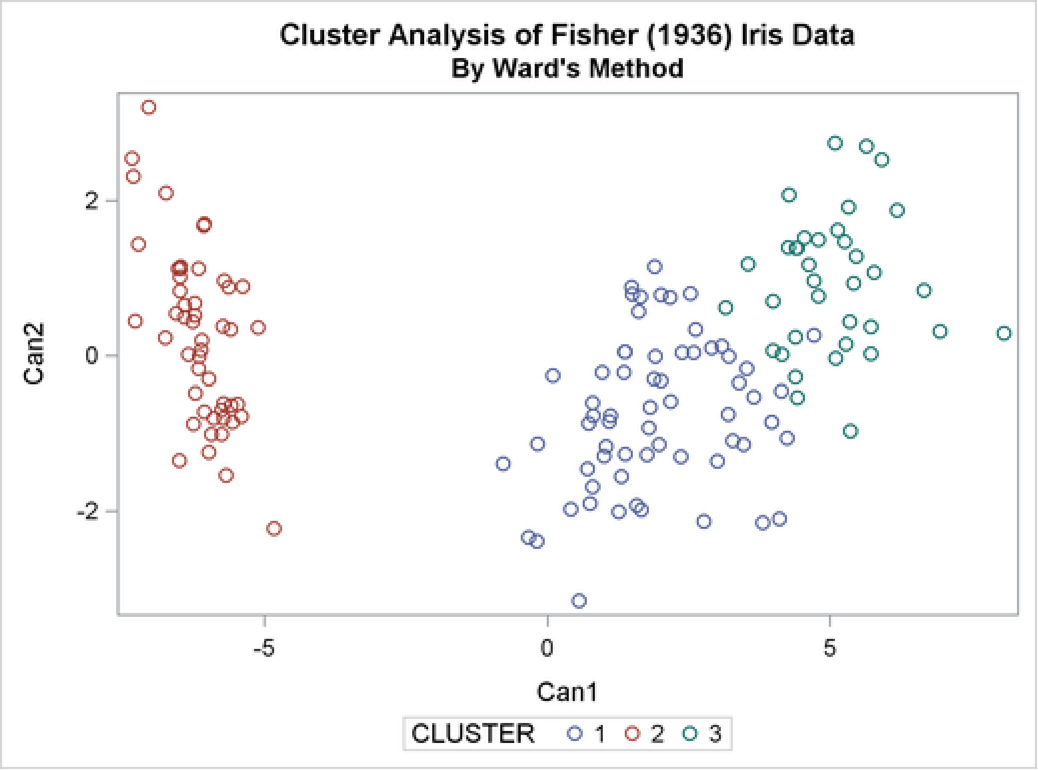
- segments of patients at baseline?
- patterns over time?
- concurrence of adverse events?

➡ Find „similar“ patients!

Learning algorithms

Supervised learning	Unsupervised learning
Regression	Clustering
Classification	Dimension reduction

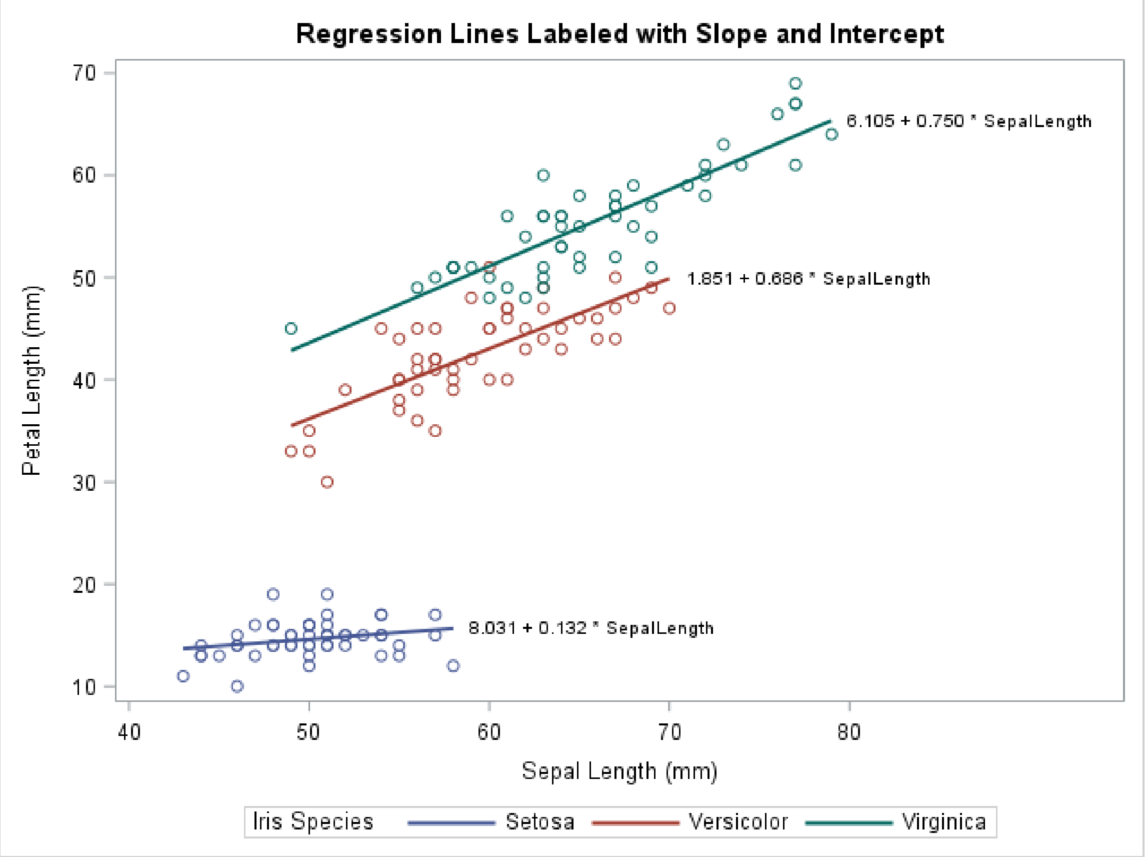




Picture source: SAS

Supervised learning questions

- What predicts response?
- Which patients drop out earlier?
- What leads to higher quality of life?



Cluster analysis

Goal: group similar patients

Similarity?

- Vector of variables
- Distance

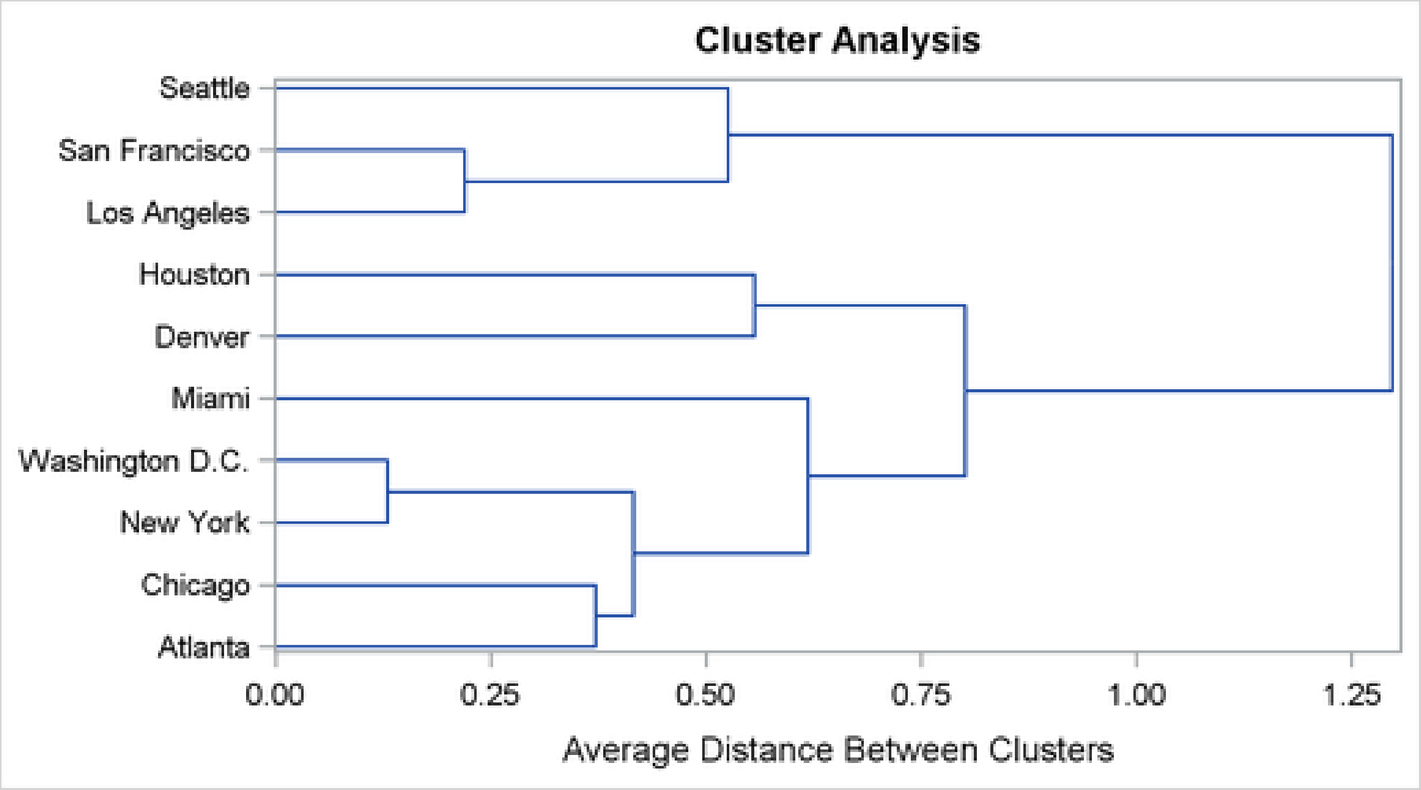
Hierarchically clustering

Step 1: Each patient is one cluster

Step 2: Find the closest 2 clusters

Step 3: Combine these 2 clusters

Repeat steps 2 and 3 until only one cluster exists



Picture source: SAS

Distance – a selection (SAS notation)

- Single
- Complete
- Average
- Centroid
- k-th nearest neighbor
- Ward

Single

$$D_{KL} = \min_{i \in C_K} \min_{j \in C_L} d(x_i, x_j)$$

- minimum distance between an observation in one cluster and an observation in the other cluster
- No constraints on shape of clusters
- Good in elongated and irregular clusters
- Chaining tendency
- Could combine with trimming

Complete

$$D_{KL} = \max_{i \in C_K} \max_{j \in C_L} d(x_i, x_j)$$

- maximum distance between an observation in one cluster and an observation in the other cluster
- Biased towards equal diameter clusters
- Sensitive to outliers

Average

$$D_{KL} = \frac{1}{N_K N_L} \sum_{i \in C_K} \sum_{j \in C_L} d(x_i, x_j)$$

- average distance between pairs of observations, one in each cluster
- Tends to join clusters with small variances
- Biased towards similar variance clusters

Centroid $D_{KL} = \| \bar{\mathbf{x}}_K - \bar{\mathbf{x}}_L \|^2$

- Euclidean distance between their centroids or means
- Robust to outliers

k-th nearest neighbor

$$d^*(x_i, x_j) = \begin{cases} \frac{1}{2}\left(\frac{1}{f(x_i)} + \frac{1}{f(x_j)}\right) & \text{if } d(x_i, x_j) \leq \max\ (r_k(x_i), r_k(x_j)) \\ \infty & \text{otherwise} \end{cases}$$

- $r_k(x)$ = distance to k-th nearest neighbor of x
- f(x) = density of observations in sphere with radius $r_k(x)$
- Good for high density clusters
- Similar approach with uniform kernel possible

Ward

$$D_{KL} = B_{KL} = \frac{\|\bar{\mathbf{x}}_K - \bar{\mathbf{x}}_L\|^2}{\frac{1}{N_K} + \frac{1}{N_L}}$$

- ANOVA sum of squares between the two clusters added up over all the variables
- Minimizes within-cluster sum of squares at each step
- Joins small clusters
- Roughly equal sized clusters
- Sensitive to outliers

The curse of choice

- Apply various approaches
- Check robustness
  - Outliers
  - Trimming
  - Number of clusters
- Check interpretability
  - Characteristics of clusters (means across variables)
- Check for unreasonable large or small clusters

How many clusters?

- Variance – bias trade-off
  - Use dendrogram
- Interpretability
  - Look at characteristics of clusters
- Sample size (more patients – more clusters)
- „More art than science“

Practical topics

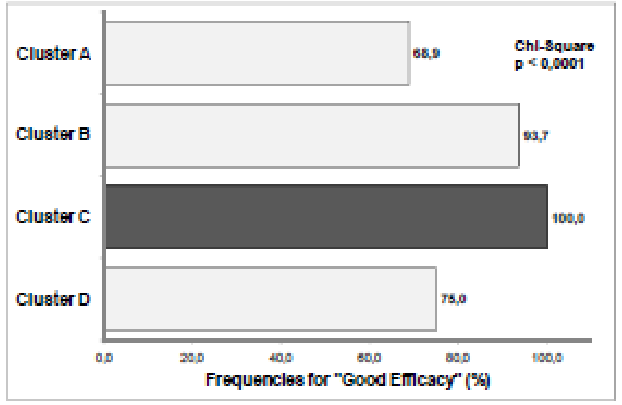
- Missing data
- Standardization
- Outliers
- Correlated data

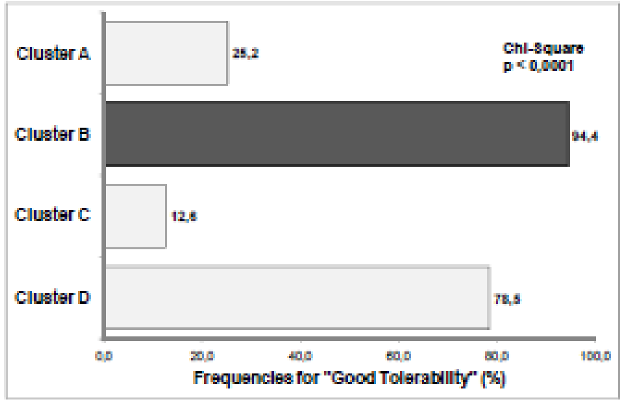
Software

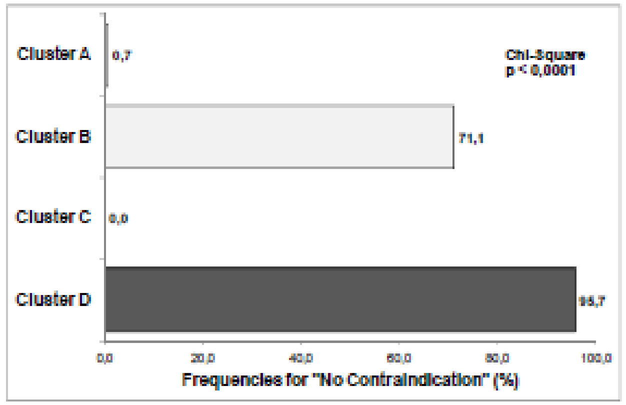
- SAS PROC CLUSTER
  https://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#cluster_toc.htm
- R package „cluster“
  https://cran.r-project.org/web/packages/cluster/cluster.pdf

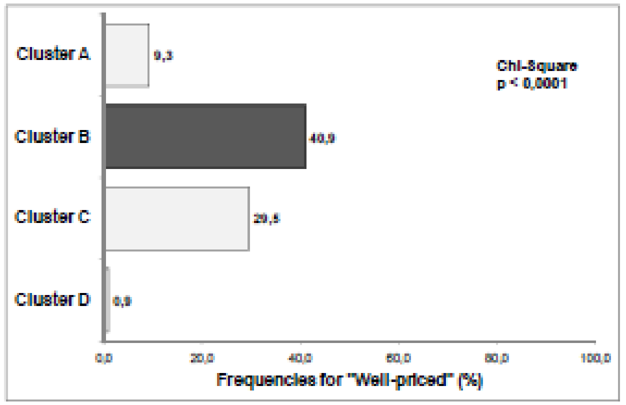
Choice of drugs

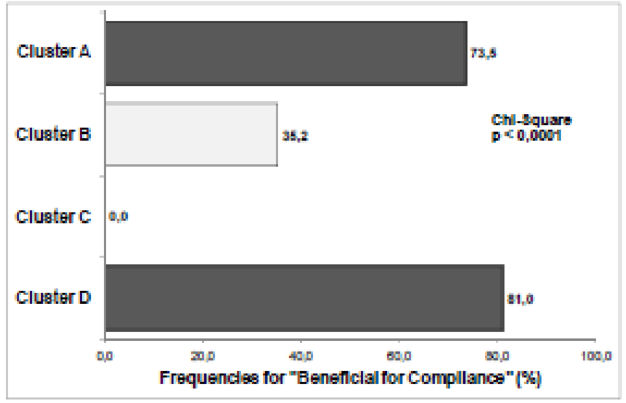
- 7 binary questions
  - Good efficacy
  - Good tolerability
  - No contraindication
  - Patient/parent decision
  - Beneficial for compliance
  - Well-priced
  - Duration of action
- 504 patients

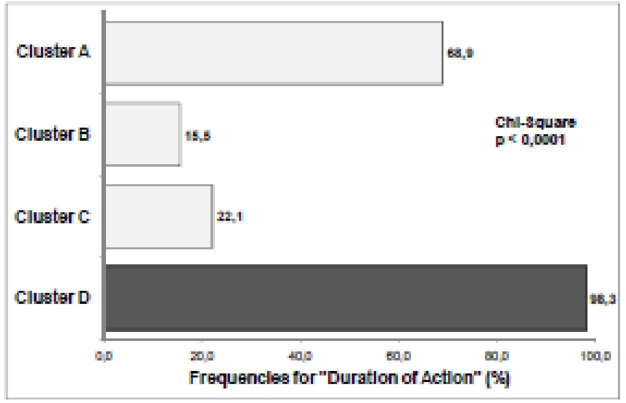


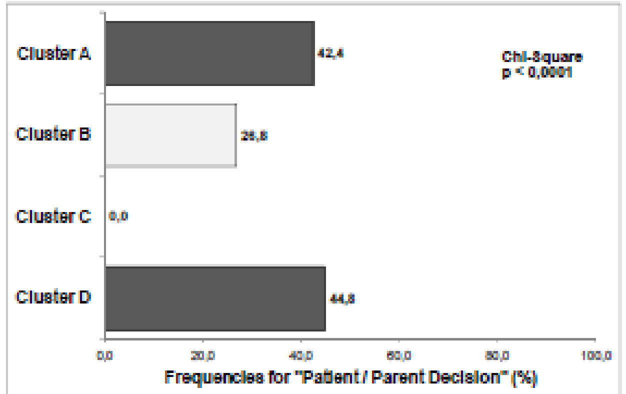






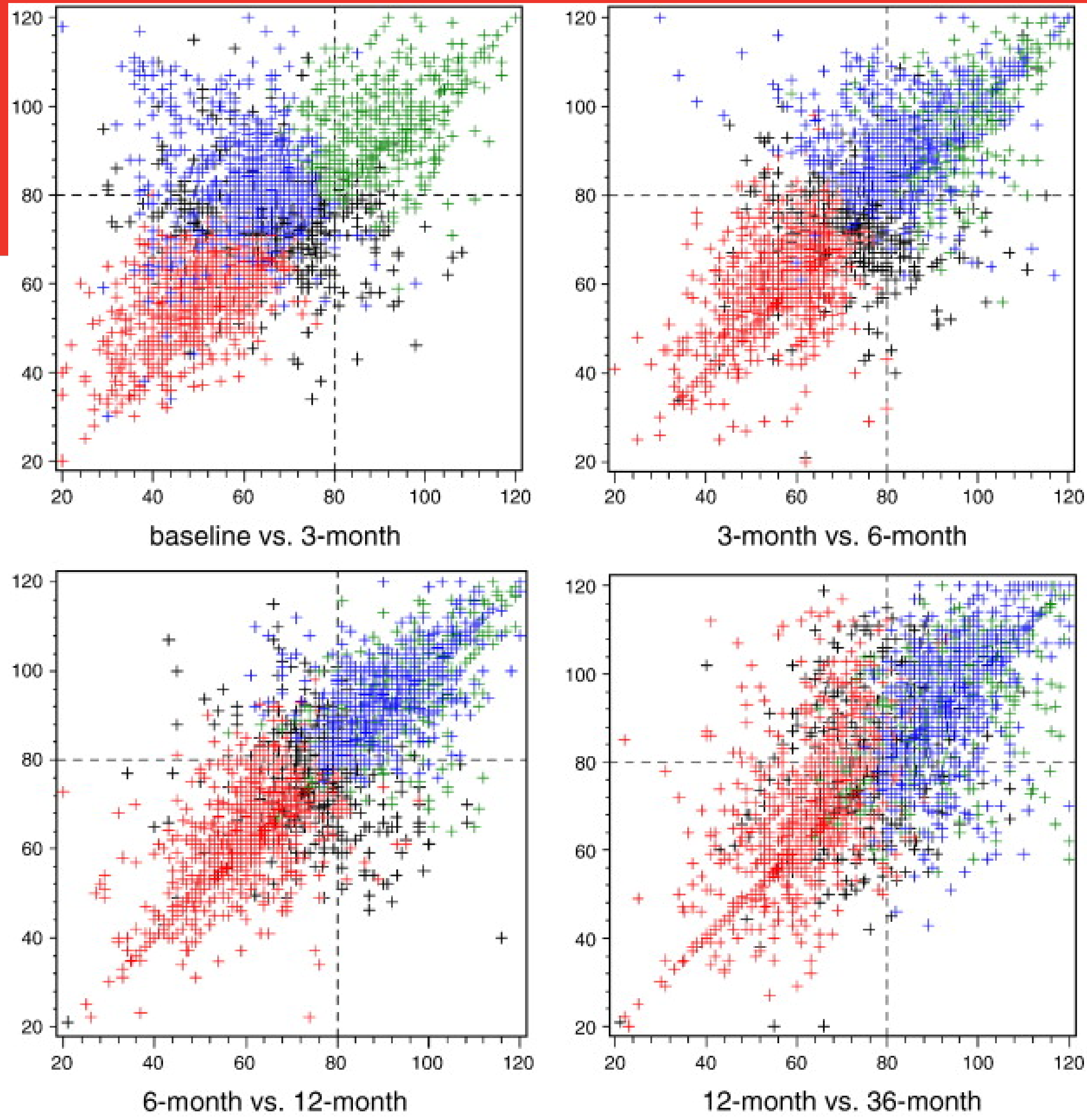


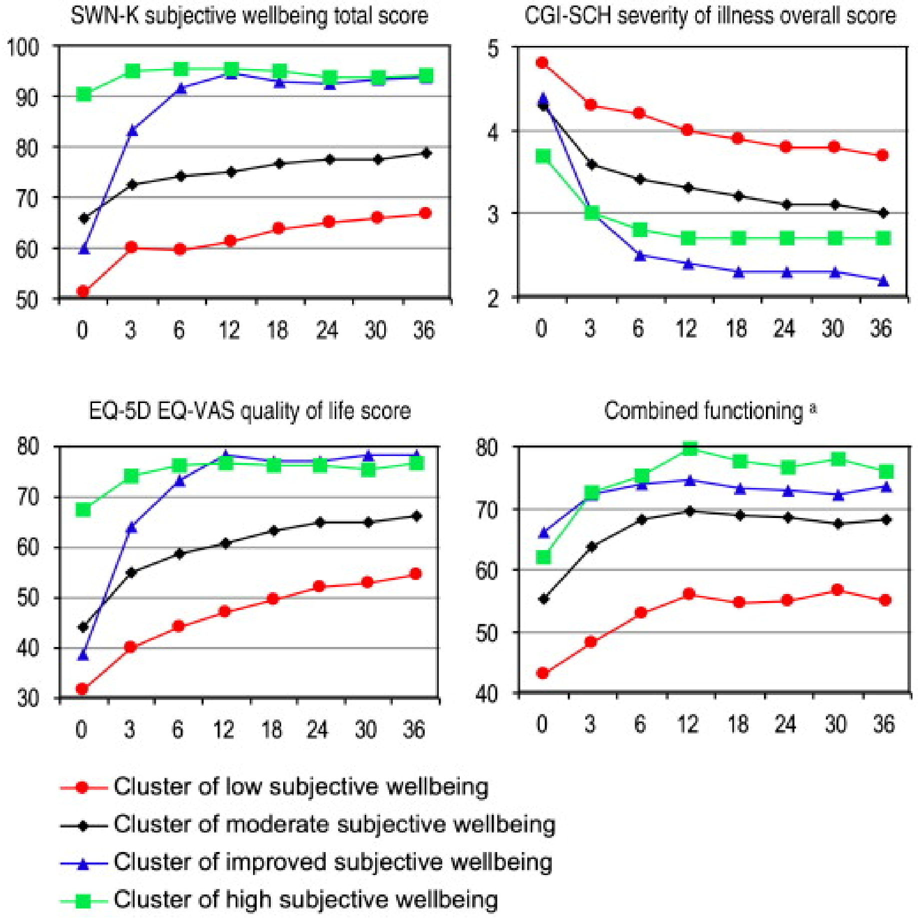




Efficacy over time

- QoL over time
- N=2842





References

- Wehmeier et al. Reasons for Physicians' Choice of Medication in Medication-Naïve Patients with ADHD: Baseline Data from the COMPLY Observational Study. Current Drug Therapy, 2010, 5, 139-150.
- Lambert et al. Long-term patterns of subjective wellbeing in schizophrenia: cluster, predictors of cluster affiliation, and their relation to recovery criteria in 2842 patients followed over 3 years. Schizophr Res. 2009 Feb;107(2-3):165-72.