Longitudinal data analysis in (pre-)clinical research on rare diseases

Georg Zimmermann[1][2]

[1]Department of Artificial Intelligence and Human Interfaces, Paris Lodron University, Salzburg, Austria

[2]Research Programme Biomedical Data Science, Paracelsus Medical University, Salzburg, Austria

WBS Colloquium, July 2 2025

Acknowledgments

Special thanks to...

- My cooperation partners with respect to this work: Johann Bauer, Martin Laimer, Verena Wally, Arne Bathke, Martin Geroldinger, Konstantin Thiel, Geert Molenberghs, and Johan Verbeeck (and perhaps even more).
- The federal state of Salzburg for providing funding to support the WISS 2025 project 'IDA-Lab Salzburg' (20204-WISS/225/197-2019 and 20102-F1901166-KZP), and 'EXDIGIT' (20204-WISS/263/6-6022)

My position

UNIKLINIKUM
SALZBURG

PARIS
LODRON
UNIVERSITÄT
SALZBURG

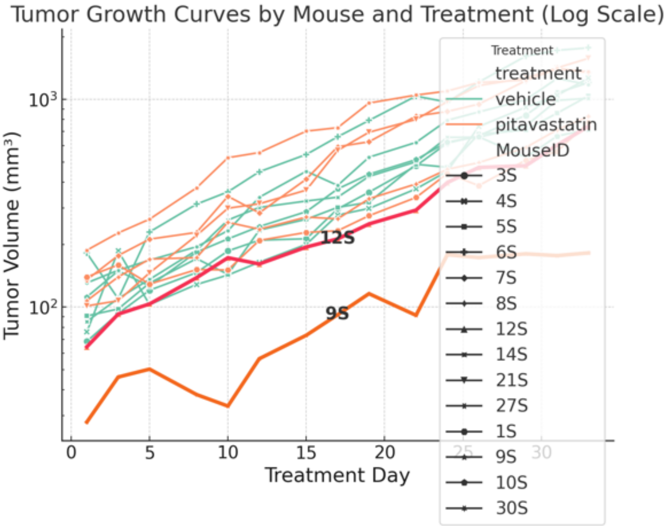
PARACELSUS
MEDIZINISCHE
UNIVERSITÄT

Outline

In my talk, I would like to . . .

- . . . provide a motivation why longitudinal data analyses are frequently encountered in (bio-)medical research, in particular in rare diseases
- . . . present some “points to consider” when deciding for the one or the other (nonparametric) approach
- . . . sketch some ideas for future research in this area

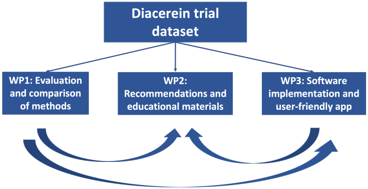
Motivation part 1: Tumor growth in preclinical research on rare diseases



Motivation part 1: Tumor growth in preclinical research on rare diseases

- (Maybe) a standard example – but:
- . . . (very) small samples
- . . . transformations of the data? Reliability of measurements?
- . . . missing data – missingness mechanism?
- . . . how to adjust for potential baseline differences in tumor volumes?
- . . . etc.
- Most of these challenges (and some more) also apply to clinical data

Motivation part 2: The EBStatMax project



Diacerein trial dataset

WP1: Evaluation and comparison of methods

WP2: Recommendations and educational materials

WP3: Software implementation and user-friendly app

The EBStatMax project's aims are to **reanalyze the data** using various state-of-the-art methodologies, **provide recommendations** for future trials, **devise computational tools** for practitioners in order to implement results in concrete trial analysis, and **design educational material**.

EUROPEAN JOINT PROGRAMME RARE DISEASES

IDA LAB INTELLIGENT DATA ANALYTICS SALZBURG

PARACELSUS MEDIZINISCHE PRIVATUNIVERSITÄT

PARIS LODRON UNIVERSITÄT SALZBURG

LAND SALZBURG

UNIKLINIKUM SALZBURG

eb Haus Austria.

UPPSALA UNIVERSITET

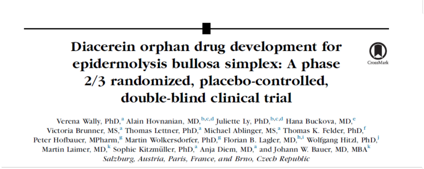
KU LEUVEN

I-BioStat

UHASSELT

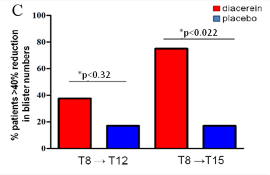
[This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825575]

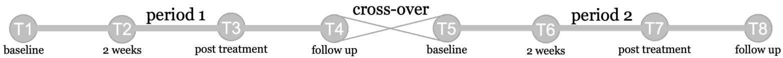
Motivation part 2: The EBStatMax project



Diacerein orphan drug development for epidermolysis bullosa simplex: A phase 2/3 randomized, placebo-controlled, double-blind clinical trial

Verena Wally, PhD,[a] Alain Hovnanian, MD,[b,c,d] Juliette Ly, PhD,[b,c,d] Hana Buckova, MD,[e]
Victoria Brunner, MS,[a] Thomas Lettner, PhD,[a] Michael Ablinger, MS,[a] Thomas K. Felder, PhD,[f]
Peter Hofbauer, MPharm,[g] Martin Wolkersdorfer, PhD,[g] Florian B. Lagler, MD,[h,i] Wolfgang Hitzl, PhD,[j]
Martin Laimer, MD,[k] Sophie Kitzmüller, PhD,[a] Anja Diem, MD,[a] and Johann W. Bauer, MD, MBA[k]
Salzburg, Austria, Paris, France, and Brno, Czech Republic





- Longitudinal cross-over design
  ⇒ Every subject k is observed repeatedly at t time points ($t = 4$ time points per period)

Motivation part 2: The EBStatMax project

- Ordinal outcomes: *visual analogue scales* or *quality of life* questionnaires
  ⇒ analyzed using nonparametric methods
- For complex longitudinal designs (e.g cross-over), appropriate methods for analyzing purely ordinal outcomes are scarce
- state-of-the-art nonparametric approaches:
  - **nparLD** – R package
  - **generalized pairwise comparisons (GPC)**

Methods: Introduction

- Two treatment groups (placebo vs. verum) within each period, t time points per period, n subjects.
- Furthermore, we assume $X_{iks}^{(j)} \overset{iid}{\sim} F_{is}^{(j)}$, that is, we denote the marginal distribution of group $i \in \{1, 2\}$ within period $j \in \{1, 2\}$ at time point $s \in \{1, \ldots, t\}$ by $F_{is}^{(j)}$.
- It should be noted that no specific parametric assumptions are made on $F_{is}^{(j)}$.

Methods: nparLD

- The `R` package `nparLD` provides user-friendly access to robust rank-based methods for the analysis of longitudinal data in factorial settings
- Notational system: each design depends on the number of factors
- Fx - LD - Fy, where x and y are the number of whole- and sub-plot factors, respectively.
- Our setting:
  - Number of levels of group (whole-plot factor): 2
  - Number of levels of time (sub-plot factor): 4
  - F1 - LD - F1 Model
- We are only interested in answering the question whether the longitudinal profiles of the VAS scores differ between verum and placebo – we are testing for a nonparametric interaction effect

Methods: nparLD

One may use the ANOVA-type statistic (ATS):

$$A_n(\mathbf{C}) = \frac{n}{tr(\mathbf{C}\hat{\mathbf{V}})} \hat{\boldsymbol{\theta}}^T \mathbf{C} \hat{\boldsymbol{\theta}}, \tag{1}$$

- where **C** is the hypothesis matrix,
- $\hat{\boldsymbol{\theta}}$ represents the vector of "estimated relative effects" $\hat{\theta}_{11}, \ldots, \hat{\theta}_{1t}, \hat{\theta}_{21}, \ldots, \hat{\theta}_{2t}$, and
- $\hat{\mathbf{V}}$ is the corresponding covariance matrix estimator.

The sampling distribution of $A_n(\mathbf{C})$ can be approximated by a $F_{(\hat{f},\infty)}$ distribution, where $\hat{f} = \frac{(tr(\mathbf{C}\hat{\mathbf{V}}))^2}{tr(\mathbf{C}\hat{\mathbf{V}}\mathbf{C}\hat{\mathbf{V}})}$

Relative effects

- For independent rv's $X \sim F$, $Y \sim G$,

$$\theta := P(X < Y) + \frac{1}{2}P(X = Y) = \int F \mathrm{d}G$$

- Pairwise relative effects: a independent samples, i.e., observations $Y_{i1}, \ldots, Y_{in_i} \overset{i.i.d.}{\sim} F_i$, $i \in \{1, \ldots, a\}$, all $Y_{11}, \ldots, Y_{an_a}$ independent,

$$\theta_{ij} := P(Y_{i1} < Y_{j1}) + \frac{1}{2}P(Y_{i1} = Y_{j1}),$$

- Drawback of pairwise relative effects – not transitive (e.g., Thangavelu and Brunner 2007)

Relative effects

- Comparison to a reference distribution:

$$\theta_i = P(W < Y_{i1}) + \frac{1}{2}P(W = Y_{i1}) \quad \text{or}$$
$$\psi_i = P(Z < Y_{i1}) + \frac{1}{2}P(Z = Y_{i1}),$$

  where $Y_{i1} \sim F_i$, $W \sim H$, and $Z \sim H^{\psi}$, $i \in \{1, 2, \ldots, a\}$.

- H and H^{ψ} denote the weighted and unweighted averages, respectively,

$$H(x) := \frac{1}{N}\sum_{i=1}^{a} n_i F_i(x),$$
$$H^{\psi}(x) := \frac{1}{a}\sum_{i=1}^{a} F_i(x).$$

- Extensions to multi-factorial designs (including repeated measures) by splitting up the index i

Estimation

- Applying the plug-in principle (i.e., replacing the population CDFs by their empirical counterparts) yields

$$\hat{\theta}_i := \frac{1}{N}\left(\bar{R}_{i.} - \frac{1}{2}\right),$$
$$\hat{\psi}_i := \frac{1}{N}\left(\bar{R}^{\psi}_{i.} - \frac{1}{2}\right).$$

- Here, $\bar{R}_{i.}$ and $\bar{R}^{\psi}_{i.}$ denote the group-specific averages of the classical ranks $R_{i\ell}$ and the so-called pseudo-ranks $R^{\psi}_{i\ell}$, which are defined as follows:

$$R_{i\ell} := \frac{1}{2} + N\hat{H}(Y_{i\ell}),$$
$$R^{\psi}_{i\ell} := \frac{1}{2} + N\hat{H}^{\psi}(Y_{i\ell}),$$

for $i \in \{1, 2, \ldots, a\}$ and $\ell \in \{1, 2, \ldots, n_i\}$.

Methods: GPC

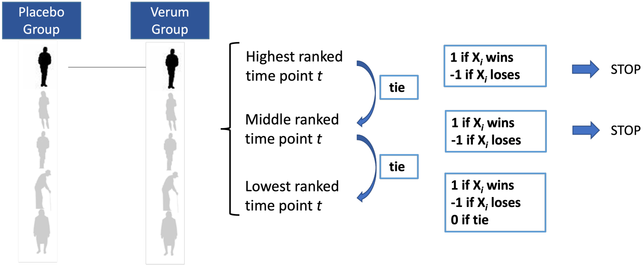
- With a single outcome and no missing data, the GPC test is a linear transformation of the Mann-Whitney test
- The GPC method evaluates $\mathbf{X}_{ik}^{(j)}$ (i.e., the vector of period-specific longitudinal measurements of subject k in group i) by constructing all possible pairs (one from each treatment arm), and subsequently assigning a score to each pair.
- GPC variants:
  - Univariate GPC
  - Prioritized GPC
  - Non-prioritized GPC
  - Matched GPC
  - Unmatched GPC

Methods: GPC

- A summary measure per period can be constructed, which is compared per pair (= univariate GPC) or the longitudinal VAS scores can be compared in a multivariate way by comparing the VAS scores per timepoint between pairs (= multivariate GPC).
- Matched GPC compares treatment arms only within the same subject, while the unmatched approach compares each subject from the placebo group with each subject of the treatment group.
- Per pair, a score $U_{k\ell}$ corresponding to the uni- or multivariate comparison of the VAS scores, denoted by V_{1k} for patient k under verum and $V_{2\ell}$ for patient ℓ under placebo, is assigned as follows (with $k, \ell \in \{1, \ldots, n\}$ for the unmatched GPC and $k = \ell$ for the matched GPC) :

$$U_{k\ell} = \begin{cases} 1, & \text{if } V_{1k} > V_{2\ell} \\ -1, & \text{if } V_{1k} < V_{2\ell} \\ 0, & \text{if } V_{1k} = V_{2\ell}, \end{cases} \tag{2}$$

Methods: matched vs. unmatched prioritized GPC



Placebo Group

Verum Group

Highest ranked time point *t*

tie

Middle ranked time point *t*

tie

Lowest ranked time point *t*

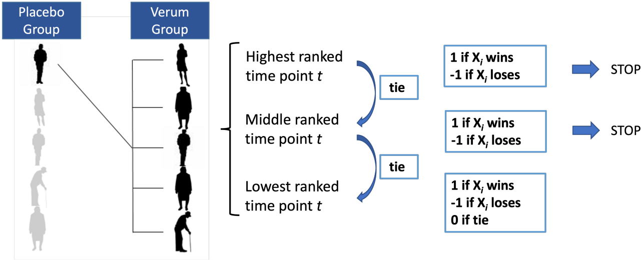
1 if X_i wins
-1 if X_i loses

STOP

1 if X_i wins
-1 if X_i loses

STOP

1 if X_i wins
-1 if X_i loses
0 if tie



Placebo Group

Verum Group

Highest ranked time point *t*

tie

Middle ranked time point *t*

tie

Lowest ranked time point *t*

1 if X_i wins
-1 if X_i loses

STOP

1 if X_i wins
-1 if X_i loses

STOP

1 if X_i wins
-1 if X_i loses
0 if tie

Methods: GPC

- In order to construct a GPC test statistic, the scores $U_{k\ell}$ are averaged and divided by an appropriate estimator of the standard error.
- Effect measure: average of the scores = “net benefit”
- Finally, “classical” approaches (e.g., sign test) can be used for calculating p-values, etc.
- Details are provided, e.g., in Buyse (2010).

Simulation design

- **Main aim:** Ensure that the simulation setting closely resembles the real-life data, while at the same time being as “neutral” as possible w.r.t. comparing the different methodological approaches!
- We have n subjects observed repeatedly at $t = 4$ time points per period in a crossover trial
- For each subject $k \in \{1, 2, \ldots, n\}$, we have a pair $(\mathbf{X}_{1k}, \mathbf{X}_{2k})$ of vectors with 4 components each (corresponding to the 4 time points per period)
- In each simulation run, the blocks $\mathbf{X}_{ik}$ were randomly permuted across all subjects.

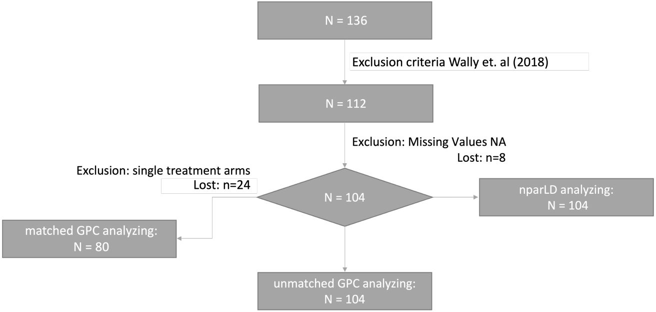
Simulation Design

For the power simulations, the following steps were carried out:

1. Random variables $Z_k \overset{iid}{\sim} \mathcal{D}$, $k \in \{1, 2, \ldots, n\}$, were generated, where $\mathcal{D}$ was either a normal distribution $\mathcal{N}(\mu_{\text{norm}}, 1)$ or a lognormal distribution $LN(\mu_{\text{log}}, 1)$, with $\mu_{\text{norm}} \in \{2, 3, 4\}$ and $\mu_{\text{log}} \in \{0.2, 0.6, 0.9\}$.
2. These random variables $(Z_k)_{k=1}^n$ were subsequently added to the observations from the placebo group. Two different scenarios were considered:
    - Scenario 1: The random variables were added to the VAS scores under placebo at the third time point (*i.e.*, the post-treatment visit) only.
    - Scenario 2: The random variables were added to the VAS scores under placebo at the third time point and additionally, $(Z_k/2)_{k=1}^n$ were added at the fourth time point.
3. The corresponding “new” observations resulting from Step 2 were appropriately cut off and rounded, if required, in order to adequately represent VAS scores.

Simulation Design

- This setup is closely aligned with clinical expertise (w.r.t. distributions & parameters)
- $R = 5000$ simulation runs were performed. The resulting empirical power values are based on using the two-sided level $\alpha = 0.05$.
- Following in- and exclusion criteria were used:



Results

	Type I Error	
	Pruritus	Pain
nparLD two-sided Period 1	0.0586	0.056
nparLD two-sided Period 2	0.0618	0.0646
univariate matched GPC one-sided	0.0592	0.0666
univariate matched GPC two-sided	0.024	0.0344
univariate unmatched GPC one-sided	0.0444	0.051
univariate unmatched GPC two-sided	0.0468	0.0492
prioritized matched GPC one-sided	0.0538	0.0646
prioritized matched GPC two-sided	0.0214	0.0252
prioritized unmatched GPC one-sided	0.0446	0.048
prioritized unmatched GPC two-sided	0.0472	0.049
non prioritized unmatched GPC one-sided	0.0484	0.054
non prioritized unmatched GPC two-sided	0.0496	0.0508

Results

Power

	Pain		**Pruritus**	
	Secenario 1	Secenario 2	Secenario 1	Secenario 2
	nparLD			
$\mu_{\text{log}} = 0.2$	0.2402	0.2616	0.2846	0.3128
$\mu_{\text{log}} = 0.6$	0.3476	0.3566	0.3642	0.3808
$\mu_{\text{log}} = 0.9$	0.4522	0.4552	0.4418	0.4438
$\mu_{\text{norm}} = 2$	0.28	0.2888	0.3	0.3252
$\mu_{\text{norm}} = 3$	0.5112	0.4872	0.4532	0.4542
$\mu_{\text{norm}} = 4$	0.7322	0.6846	0.604	0.5694
	prioritized unmatched GPC			
$\mu_{\text{log}} = 0.2$	0.6404	0.6432	0.8808	0.888
$\mu_{\text{log}} = 0.6$	0.786	0.7888	0.9334	0.9402
$\mu_{\text{log}} = 0.9$	0.8826	0.8844	0.9642	0.9694
$\mu_{\text{norm}} = 2$	0.6702	0.6758	0.889	0.891
$\mu_{\text{norm}} = 3$	0.8834	0.889	0.95	0.95
$\mu_{\text{norm}} = 4$	0.9778	0.9788	0.9528	0.9546
	univariate unmatched GPC			
$\mu_{\text{log}} = 0.2$	0.1106	0.1812	0.1398	0.223
$\mu_{\text{log}} = 0.6$	0.1768	0.3068	0.2024	0.3486
$\mu_{\text{log}} = 0.9$	0.2638	0.4626	0.2902	0.5
$\mu_{\text{norm}} = 2$	0.1236	0.222	0.1616	0.264
$\mu_{\text{norm}} = 3$	0.2284	0.4364	0.2626	0.4638
$\mu_{\text{norm}} = 4$	0.38	0.7014	0.4086	0.6732

Discussion

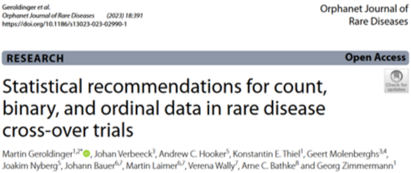
Still, comparing the methods “neutrally” is somewhat challenging:

- nparLD: analyses could only be conducted for each period separately
  $\Rightarrow$ cross-over aspect partially lost
- univariate GPC: based on summary measurement
  $\Rightarrow$ longitudinal information partially lost
- matched GPC: based on a pairwise comparison between both periods
  $\Rightarrow$ several subjects had to be excluded due to missing data
- missing data: problem for nparLD and univariate GPC approaches

Discussion

- Matched GPC was rather conservative
- nparLD liberal only in a few scenarios
- nparLD: high power despite a smaller sample size ($n = 6$, $n = 7$; as a result of period-specific analyses) → good performance with (very) small sample sizes
- prioritized unmatched GPC achieved highest power
  $\Rightarrow$ prioritization of the time points has a big impact on power (prioritized based on clinical reasoning)
  $\Rightarrow$ different prioritization might lead to a deterioration

EBStatMax – project output



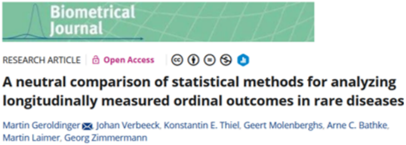
Geroldinger *et al.*
Orphanet Journal of Rare Diseases (2023) 18:391
https://doi.org/10.1186/s13023-023-02990-1

Orphanet Journal of
Rare Diseases

RESEARCH **Open Access**

Statistical recommendations for count, binary, and ordinal data in rare disease cross-over trials

Martin Geroldinger[1,2*], Johan Verbeeck[3], Andrew C. Hooker[5], Konstantin E. Thiel[1], Geert Molenberghs[3,4], Joakim Nyberg[5], Johann Bauer[6,7], Martin Laimer[6,7], Verena Wally[7], Arne C. Bathke[8] and Georg Zimmermann[1]

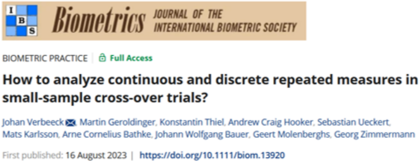


Biometrical Journal

RESEARCH ARTICLE | Open Access

A neutral comparison of statistical methods for analyzing longitudinally measured ordinal outcomes in rare diseases

Martin Geroldinger, Johan Verbeeck, Konstantin E. Thiel, Geert Molenberghs, Arne C. Bathke, Martin Laimer, Georg Zimmermann



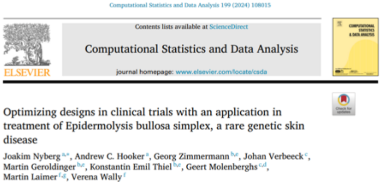
Biometrics JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY

BIOMETRIC PRACTICE | Full Access

How to analyze continuous and discrete repeated measures in small-sample cross-over trials?

Johan Verbeeck, Martin Geroldinger, Konstantin Thiel, Andrew Craig Hooker, Sebastian Ueckert, Mats Karlsson, Arne Cornelius Bathke, Johann Wolfgang Bauer, Geert Molenberghs, Georg Zimmermann

First published: 16 August 2023 | https://doi.org/10.1111/biom.13920



Computational Statistics and Data Analysis 199 (2024) 108015

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

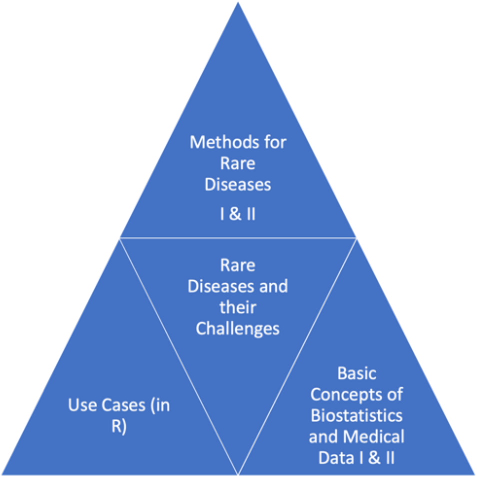
journal homepage: www.elsevier.com/locate/csda

ELSEVIER

Optimizing designs in clinical trials with an application in treatment of Epidermolysis bullosa simplex, a rare genetic skin disease

Joakim Nyberg[a,*], Andrew C. Hooker[a], Georg Zimmermann[b,e], Johan Verbeeck[c], Martin Geroldinger[b,e], Konstantin Emil Thiel[b,e], Geert Molenberghs[c,d], Martin Laimer[f,g], Verena Wally[f]

EBStatMax – project output



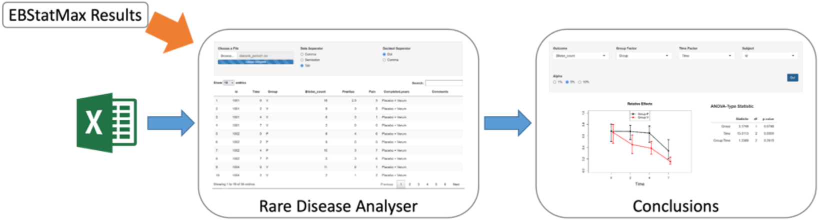
Methods for Rare Diseases I & II

Rare Diseases and their Challenges

Use Cases (in R)

Basic Concepts of Biostatistics and Medical Data I & II

EBStatMax – project output



EBStatMax Results

Choose a File

Browse... diacerin_period1.tsv

Upload complete

Data Separator

○ Comma
○ Semicolon
● Tab

Decimal Separator

● Dot
○ Comma

Show 10 entries Search:

	Id	Time	Group	Blister_count	Pruritus	Pain	Completed.years	Comments
1	1001	0	V	18	2.5	5	Placebo + Verum	
2	1001	2	V	9	3	5	Placebo + Verum	
3	1001	4	V	6	3	1	Placebo + Verum	
4	1001	7	V	2	0	0	Placebo + Verum	
5	1002	0	P	8	4	6	Placebo + Verum	
6	1002	2	P	9	0	0	Placebo + Verum	
7	1002	4	P	10	3	7	Placebo + Verum	
8	1002	7	P	5	3	6	Placebo + Verum	
9	1004	0	V	11	6	1	Placebo + Verum	
10	1004	2	V	2	1	2	Placebo + Verum	

Showing 1 to 10 of 56 entries

Previous 1 2 3 4 5 6 Next

Rare Disease Analyser

Outcome: Blister_count

Group Factor: Group

Time Factor: Time

Subject: Id

Alpha

○ 1% ● 5% ○ 10%

Go!

ANOVA-Type Statistic

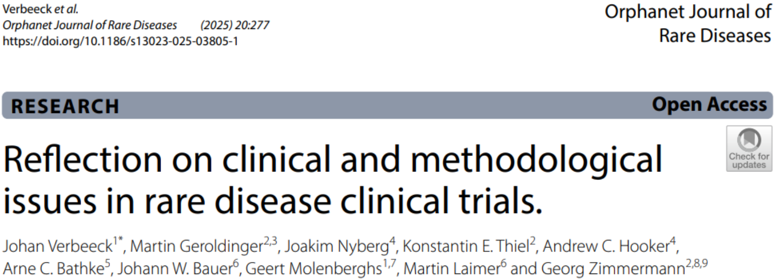
	Statistic	df	p.value
Group	3.1749	1	0.0748
Time	15.3113	2	0.0000
Group:Time	1.3389	2	0.2615

Conclusions

`https://ebstatmax.ejprarediseases.org/`

`https://imt.erdera.org/collection/ebstatmax/` (more generally on EBStatMax and the key project outcomes)

EBStatMax – project output



Verbeeck *et al.*
Orphanet Journal of Rare Diseases (2025) 20:277
https://doi.org/10.1186/s13023-025-03805-1

Orphanet Journal of
Rare Diseases

RESEARCH **Open Access**

Reflection on clinical and methodological issues in rare disease clinical trials.

Johan Verbeeck[1*], Martin Geroldinger[2,3], Joakim Nyberg[4], Konstantin E. Thiel[2], Andrew C. Hooker[4], Arne C. Bathke[5], Johann W. Bauer[6], Geert Molenberghs[1,7], Martin Laimer[6] and Georg Zimmermann[2,8,9]

Ongoing and future research

- Consider a simplified, yet still sensible version of the EB example
- Primary endpoint: VAS score at post-treatment visit
- Adjustment for the baseline VAS score (see “EMA guideline on adjustment for baseline covariates in clinical trials”, EMA/CHMP/295050/2013)
- Semiparametric mean-based setting: (M)ANCOVA with minimal assumptions (e.g., Zimmermann et al., JMVA 2020).
- Nonparametric (rank-based) uni- and multivariate analysis of covariance?
- From a project-level perspective, this research is embedded within *servEB* (federal state of Salzburg; grant no. 20102/F2300645-FPR) and a *WEAVE project* (FWO – FWF; grant no. 10.55776/PIN9834224)

Ongoing and future research

- Let $X_{i1k} \sim F_{i1}$ denote an iid sample of the outcome variable and $X_{irk} \sim F_{ir}$, $r = 2, \ldots, h$ denote samples of the $h - 1$ covariates, $i \in \{1, 2, \ldots, a\}$.
- The corresponding relative effects are denoted by q_{i1} and $q_{i2}, \ldots, q_{ih}$, respectively.
- The estimated covariate-adjusted relative effects $\hat{q}_1^*, \ldots, \hat{q}_a^*$ are defined as follows (Bathke and Brunner 2003):

$$\hat{q}_i^* = \hat{q}_{i1} - \sum_{r=2}^{h} \hat{\gamma}_r \hat{q}_{ir} \tag{3}$$

- Thereby, the procedure underlying the estimation of the coefficients $\hat{\gamma}_2, \ldots, \hat{\gamma}_h$ is based on the idea of minimizing the variance.

Ongoing and future research

- $H_0 : \mathbf{TF} = \mathbf{0}$, where $\mathbf{T}$ is an appropriate contrast matrix, and $\mathbf{F}$ denotes the vector $(F_{11}, \ldots, F_{a1})'$, i.e., the group-specific CDFs of the outcome.
- Using $\hat{\mathbf{q}}^* := (\hat{q}_1^*, \ldots, \hat{q}_a^*)'$, the ANOVA-type statistic is defined as follows:

$$A_N = \frac{Nf \cdot (\hat{\mathbf{q}}^*)'\mathbf{T}\hat{\mathbf{q}}^*}{\mathrm{tr}(\mathbf{T}\hat{\boldsymbol{\Sigma}}_N^*)} \tag{4}$$

- The distribution of the ATS under H_0 can be approximated by a $\chi^2_{\hat{f}}$ distribution, where

$$\hat{f} = \frac{\mathrm{tr}(\mathbf{T}\hat{\boldsymbol{\Sigma}}_N^*)^2}{\mathrm{tr}(\mathbf{T}\hat{\boldsymbol{\Sigma}}_N^*\mathbf{T}\hat{\boldsymbol{\Sigma}}_N^*)} \tag{5}$$

- The estimator of the covariance matrix $\hat{\boldsymbol{\Sigma}}_N^*$ is quite complicated (see Bathke and Brunner 2003).

Ongoing and future research

- As an alternative to the approximation, we consider a classical nonparametric as well as a wild bootstrap approach
- Bootstrapping is performed at the level of the so-called “rank transforms” (i.e., the estimated average CDF evaluated at the original observations)
- The bootstrap version of the ATS is then essentially the ATS (4), which is calculated based on the bootstrapped rank transforms instead of the original rank transforms.
- Under mild standard assumptions in an asymptotic framework, this approach yields an asymptotic level α test.
- Formal details and proofs are provided in the preprint Thiel et al. (2025).

Simulation results (example)

Table: Empirical type-I error on discrete ordinal data with $\alpha = 5\%$. Values exceeding a 95% Wald interval are highlighted. Legend: (FA1) $\mathcal{F}$ approximation unadjusted; (CA) χ^2 approximation NANCOVA; (FA2) $\mathcal{F}$ approximation NANCOVA; (EB) Efron bootstrap NANCOVA.

n_1:n_2	FA1	CA	FA2	EB
10:10	5.14	8.34	6.76	3.82
8:12	4.70	8.16	5.98	3.40
5:15	6.60	10.46	7.36	4.92
20:20	4.68	6.14	6.42	4.64
16:24	5.44	6.32	5.18	4.26
10:30	5.58	7.66	5.36	4.92

Simulation results (example)

Table: Empirical power on discrete ordinal data with $\alpha = 5\%$. Configurations where the empirical type-I error substantially exceeds α are greyed out. Legend: (FA1) $\mathcal{F}$ approximation unadjusted; (CA) χ^2 approximation NANCOVA; (FA2) $\mathcal{F}$ approximation NANCOVA; (EB) Efron bootstrap NANCOVA.

n_1:n_2	FA1	CA	FA2	EB
10:10	49.38	73.30	68.60	59.38
8:12	46.68	72.22	64.68	56.26
5:15	40.98	62.36	51.64	40.82
20:20	79.58	94.84	93.20	93.32
16:24	77.64	94.48	92.62	92.00
10:30	63.66	87.00	82.48	78.12

Back to preclinical research: Tumor growth

- The applied researchers asked many questions
- Structured approach: Systematically collecting the questions from a "core group" of researchers . . .
- . . . and a subsequent rating process.
- Prioritization of 2-3 topics.
- Then: Asking the collaboration partners for data examples $\rightarrow$ basis for simulation scenarios
- Current status: Preparing the datasets and simulation scenarios, selection of methodological approaches / literature search.
- Final goal: Answering the questions by simulations and/or theoretical considerations (or existing literature)

Wrap-up and take-home messages

- Research at the interface between statistics and applications in rare diseases means: Whenever you are not quite sure which method to use, there is a good reason for doing methodological research.
- There are many different approaches for longitudinal data analysis available, which use (slightly) different effect measures (e.g., importantly, interaction effects based on relative effects vs. GPC / net benefit)
- Therefore, systematic comparisons of these different approaches as well as detailed investigations regarding various subtle issues are much needed
- So, on the one hand, there is a huge number of potentially useful methods in some situations...
- ... on the other hand, however, there is still room for methodological improvements and even for developing novel methods in some highly relevant settings (e.g., covariate adjustment)

References

- Bathke, A. & Brunner, E. (2003), 'A nonparametric alternative to analysis of covariance', in Akritas, M. & Politis, D. (eds), 'Recent Advantages and Trends in Nonparametric Statistics', Elsevier, Amsterdam, 109–120.
- Boulesteix, A.-L., et al. (2018), 'On the necessity and design of studies comparing statistical methods', Biometrical Journal 60(1), 216–218.
- Boulesteix, A.-L., Lauer, S. & Eugster, M. J. A. (2013), 'A plea for neutral comparison studies in computational sciences', PLOS ONE 8(4), 1–11.
- Brunner, E., Domhof, S. & Langer, F. (2002). Nonparametric analysis of longitudinal data in factorial experiments. John Wiley & Sons.
- Brunner, E., Bathke, A. C. & Konietschke, F. (2019). Rank and pseudo-rank procedures for independent observations in factorial designs, using R and SAS. Springer.
- Brunner, E., Konietschke, F., Bathke, A.C., and Pauly, M. (2020). 'Ranks and Pseudo-ranks–Surprising Results of Certain Rank Tests in Unbalanced Designs.' International Statistical Review 89 (2): 349–366.

References

- Buyse, M. (2010), ‘Generalized pairwise comparisons of prioritized outcomes in the two-sample problem’, Statistics in Medicine 29, 3245–3257.
- Geroldinger, M., et al. (2023). ’A neutral comparison of statistical methods for analyzing longitudinally measured ordinal outcomes in rare diseases.’ Biom J, 66, 2200236, doi: 10.1002/bimj.202200236.
- Noguchi, K. et al. (2012). nparLD: An R software package for the nonparametric analysis of longitudinal data in factorial experiments. Journal of Statistical Software, 50(12), 1–23. https://doi.org/10.18637/jss.v050.i12
- Nyberg, J., Hooker, A.C., Zimmermann, G., Verbeeck, J., Geroldinger, M., Thiel, K.E., Molenberghs, G., Laimer, M., Wally, V. (2024). ’Optimizing designs in clinical trials with an application in treatment of Epidermolysis bullosa simplex, a rare genetic skin disease’. Computational Statistics & Data Analysis 199, 108015. doi: 10.1016/j.csda.2024.108015.

References

- Thangavelu, K. and Brunner, E. (2007). 'Wilcoxon-Mann-Whitney test for stratified samples and Efron's paradox dice.' Journal of Statistical Planning and Inference 137: 730737.
- Thiel, K.E., Sattler, P., Bathke, A.C., Zimmermann, G. (2025). 'Resampling NANCOVA: Nonparametric Analysis of Covariance in Small Samples'. Computational Statistics and Data Analysis, under review, DOI: 10.48550/arXiv.2412.17513.
- Verbeeck, J., Ozenne, B. & Anderson, W. (2020), 'Evaluation of inferential methods for the net benefit and win ratio statistics', Journal of Biopharmaceutical Statistics 30(5), 765–782.
- Verbeeck, J., Geroldinger, M., Nyberg, J., Thiel, K. E., Hooker, A. C., Bathke, A. C., Bauer, J. W., Molenberghs, G., Laimer, M., Zimmermann, G. (2025). 'Reflection on clinical and methodological issues in rare disease clinical trials.' Orphanet Journal of Rare Diseases 20(1), 277. doi: 10.1186/s13023-025-03805-1
- Wally, V. et al. (2018). Diacerein orphan drug development for epidermolysis bullosa simplex: A phase 2/3 randomized, placebo-controlled, double-blind clinical trial. J Am Acad Dermatol, 78(5), 892–901.

Thank you for your attention!