Received: 27 August 2022 | Revised: 27 December 2022 | Accepted: 30 January 2023

DOI: 10.1002/bimj.202200236

Biometrical Journal

RESEARCH ARTICLE

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

Martin Geroldinger[1,2] | Johan Verbeeck[3] | Konstantin E. Thiel[1,2] | Geert Molenberghs[3,4] | Arne C. Bathke[5] | Martin Laimer[6] | Georg Zimmermann[1,2]

[1]Team Biostatistics and Big Medical Data, IDA Lab Salzburg, Paracelsus Medical University, Salzburg, Austria

[2]Department of Research and Innovation, Paracelsus Medical University, Salzburg, Austria

[3]Data Science Institute (DSI), Interuniversity Institute for Biostatistics and Statistical Bioinformatics (I-BioStat), Hasselt University, Hasselt, Belgium

[4]Interuniversity Institute for Biostatistics and Statistical Bioinformatics (I-BioStat), KULeuven, Leuven, Belgium

[5]Intelligent Data Analytics (IDA) Lab Salzburg, Department of Artificial Intelligence and Human Interfaces, Faculty of Digital and Analytical Sciences, Paris Lodron University of Salzburg, Salzburg, Austria

[6]Department of Dermatology and Allergology, Paracelsus Medical University, Salzburg, Austria

Correspondence
Martin Geroldinger, Team Biostatistics and Big Medical Data, IDA Lab Salzburg, Paracelsus Medical University
Strubergasse 21, A-5020 Salzburg, Austria.
Email: martin.geroldinger@pmu.ac.at

Funding information
H2020 Societal Challenges, Grant/Award Number: 825575

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

Abstract

Ordinal data in a repeated measures design of a crossover study for rare diseases usually do not allow for the use of standard parametric methods, and hence, nonparametric methods should be considered instead. However, only limited simulation studies in settings with small sample sizes exist. Therefore, starting from an Epidermolysis Bullosa simplex trial with the above-mentioned design, a rank-based approach using the R package nparLD and different generalized pairwise comparisons (GPC) methods were compared impartially in a simulation study. The results revealed that there was not one single best method for this particular design, because a trade-off exists between achieving high power, accounting for period effects, and for missing data. Specifically, nparLD as well as the unmatched GPC approaches do not address crossover aspects, and the univariate GPC variants partly ignore the longitudinal information. The matched GPC approaches, on the other hand, take the crossover effect into account in the sense of incorporating the within-subject association. Overall, the prioritized unmatched GPC method achieved the highest power in the simulation scenarios, although this may be due to the specified prioritization. The rank-based approach yielded good power even at a sample size of $N = 6$, whereas the matched GPC method could not control the type I error.

KEYWORDS
Epidermolysis Bullosa simplex, generalized pairwise comparison (GPC), neutral comparison, nonparametric marginal model (nparLD), repeated measures

1 | INTRODUCTION

Ordinal outcomes are frequently used in clinical practice and biomedical research, in order to make decisions and outcome assessments more patient-centered. Examples include the visual analog scale (VAS; e.g., Kristensen et al., 2021; Lange et al., 2021) and quality of life (QOL) questionnaires (e.g., SF-36, Böhm et al., 2021), just to name a few. From a statistical perspective, these outcomes should not be analyzed using classical parametric methods, because they are purely ordinal, which renders location-shift effect measures inappropriate. Consequently, some well-established rank-based approaches (e.g., Wilcoxon–Mann–Whitney test [Mann & Whitney, 1947; Wilcoxon, 1945] Kruskal–Wallis test [Kruskal & Wallis, 1952]) are preferable in such a setting.

However, data from subjects with rare diseases pose several challenges: The low prevalence results in small patient numbers that are eligible for inclusion in clinical trials (e.g., Zimmermann et al., 2019). Hence, statistical methods with a good finite-sample performance are needed. Moreover, to compensate for the power loss due to small sample sizes, crossover or, more generally, repeated measures designs are often employed. Yet, for these more complex longitudinal designs, which may include between- as well as within-subject factors, appropriate statistical methods for analyzing purely ordinal outcomes are scarce. Especially the ANOVA-type test that is implemented in the `R` package `nparLD` might be a promising approach (Noguchi et al., 2012). The underlying idea is an extension of the nonparametric rank-based methodology that has been developed for analyzing ordinal outcomes in general factorial designs (with between-subjects factors only), which is, in turn, an extension of the Wilcoxon–Mann–Whitney test (for an overview, see Brunner et al., 2019). The corresponding effect measure is not a location shift, but the so-called *relative effect* (or *probabilistic index*), which—informally speaking—quantifies the tendency toward larger values of the outcome for each of the groups (e.g., treatment groups, time points). Consequently, this type of effect measure and the corresponding tests are indeed appropriate for analyzing purely ordinal outcomes. Alternatively, one may consider using the framework of *generalized pairwise comparisons* (GPC). The nonparametric GPC method is an extension of the Gehan test (Gehan, 1965) to the case of multivariate outcomes (for a comparison of different approaches, see Verbeeck et al., 2019). Interestingly, this approach is related to the above-mentioned rank-based methods: Assuming a univariate outcome and no missing data, the GPC/Gehan test is a linear transformation of the Wilcoxon–Mann–Whitney test (Verbeeck et al., 2021).

However, the empirical evidence on the performance of these methods in small-sample size settings is limited, and the GPC method has not been applied to longitudinal analyses of ordinal outcomes so far. On top of that, a simulation-based comparison of these approaches has not been conducted yet. At this point, it should be noted that there might be other promising methods, such as the well-known Friedman test (Friedman, 1937) or approaches based on ordinal random-effects regressions (e.g., Hedeker & Gibbons, 1994; Hedeker & Mermelstein, 2000). However, the main aim of the present manuscript is to compare the nparLD and GPC methodologies in a neutral way, due to the similarities of the underlying concepts. By contrast, including the above-mentioned methodologies in the neutral comparison would have been challenging if not impossible at all: it would have been very difficult to set up a common framework that does not favor the one or the other method a priori, due to different underlying assumptions and methodological approaches.

Therefore, in the present manuscript, we performed a systematic empirical comparison of the above-mentioned methods that is informed by clinical considerations as well as the methodological expertise of a group of statisticians with different yet complementary research interests. The authors of this manuscript are statistical and clinical experts who are part of the EBStatMax project consortium (funded by the European Joint Programme on Rare Diseases, EU Horizon 2020 grant no. 825575), aimed at developing guidance regarding appropriate statistical methods for analyzing longitudinally collected outcomes based on data from patients with Epidermolysis Bullosa as a motivating case study. Analogously to comparing different interventions to each other in randomized clinical trials, statisticians should not only focus on developing "new" methods and generating the corresponding affirmative simulation evidence, but also conduct systematic comparisons of existing approaches for analyzing data from particular study designs (Boulesteix et al., 2017). We follow these "neutral comparison" principles that have been proposed in, for example, Boulesteix et al. (2013, 2018) as closely as possible. This is reflected in the simulation setup (see Sections 3 and 4) as well as in the interdisciplinary composition of the EBStatMax consortium that comprises statisticians whose respective research interests complement each other well.

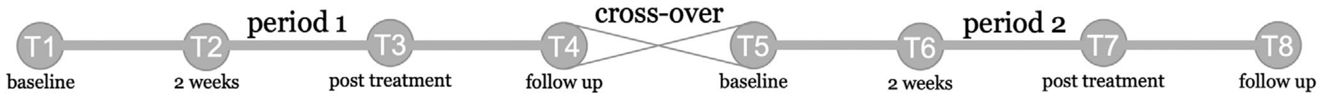

FIGURE 1 Illustration of the crossover study design of the EB trial.

In Section 2, we give a brief overview of a motivating example from Epidermolysis Bullosa, which forms the basis of this manuscript as well as of the EBStatMax project as a whole. This part is followed by an outline of the key concepts underlying the rank-based ANOVA-type approach and the GPC method (Section 3). The simulation settings and the corresponding results are reported in Section 4, and the conduct of a real-life data analysis using the statistical methods under consideration is demonstrated in Section 5. Finally, Section 6 contains a discussion of the key results and summarizes the main conclusions. Together with simulation evidence from settings with other designs and outcomes, the conclusions of the present manuscript constitute the basis for a guidance document and corresponding dissemination activities (e.g., workshop materials) on analyzing longitudinal data in rare diseases. These will be developed as a next step in the near future.

2 | MOTIVATION

When investigating rare diseases, one frequently encounters small sample sizes due to the rarity of the disease. This is in particular the case with Epidermolysis Bullosa simplex (=EB), which is a rare genetic skin disease characterized by fragility of epithelial-lined tissues and surfaces with recurrent mucocutaneous blistering for which there is no cure. Treatments that address the pathophysiology of EB as well as accompanying symptoms such as burdensome pain and itch are needed (Wally et al., 2018). A particular data set from this research area forms the basis of the EBStatMax project (see Section 1), and hence, also serves as a motivating example of the present manuscript: In 2013, a randomized, placebo-controlled, two-period crossover phase II/III trial was conducted at EB House Austria, Salzburg, Austria, which is a designated national center of expertise for genodermatoses with a focus on EB and member of the European Reference Network for Rare Skin Diseases (ERN Skin). The main aim of that study was to assess the impact of 1% Diacerein cream versus placebo in reducing the number of blisters in EB. Diacerein is a rhein prodrug and anthraquinone that was shown to inhibit in vitro and in vivo production and activity of interleukin-1 (IL-1) and other proinflammatory cytokines involved in the pathology of EB. In total, 17 patients were randomized to either placebo or Diacerein for a 4-week treatment and a 3-month follow-up in period 1. After a washout, patients were crossed over during period 2. Both periods consisted of four measurement points each (see Figure 1).

The EBStatMax project's aims are to reanalyze the data using various state-of-the-art methodologies, investigate the impact that certain characteristics of the trial have on the statistical analysis, develop strategies and design recommendations for future trials in rare diseases, and, as a means to ensure transferability and high dissemination of the results, devise computational tools for practitioners in order to implement results in concrete trial analysis, and provide educational material.

In this manuscript, the focus is on ordinal outcomes. Although the study protocol of the reference study considered count outcomes as primary end point (Wally et al., 2018), there were also additional outcomes referring to pain and pruritus (and QOL), which were assessed at each visit using a VAS. The VAS ranges from 0 (no pain/pruritus) to 10 (worst pain/pruritus imaginable) with an increment of 0.5, and QOL was assessed by a questionnaire (Wally et al., 2018). For the sake of simplicity, only the former will be considered in the present manuscript.

Note that the VAS score is usually assessed by using a line of 100 mm in length, with anchor descriptors such as "no pain/pruritus" and "worst pain/pruritus imaginable." The measurement can be done in 1-mm accuracy or 1-cm accuracy. Thus, this is seemingly a metric scale; however, the VAS score is a specific type of a continuous ordinal scale, because differences cannot be interpreted in a uniform way throughout the range of the VAS score. For example, in clinical practice, there is an important qualitative difference between a decrease in VAS from 8 to 6 and a decrease from 3 to 1 (Heller et al., 2016). Hence, since the VAS score is ordinal, this might imply the formulation of appropriate statistical models that allow for analyzing longitudinally collected ordinal outcome variables. In the sequel, the above-mentioned variables will be considered as truly ordinal, and therefore, we will not resort to quasi-metric modeling approaches (e.g., latent variable modeling). Our aim is to examine the performance of rank-based nonparametric methods in comparison to different GPC approaches, including a detailed evaluation of these methods in small sample size settings in a neutral way.

Apart from that, the EBStatMax project is also focused on evaluating the advantages and drawbacks of statistical analysis approaches for other types of outcomes (i.e., count and binary outcomes). Furthermore, one of the project's main aims is to provide advice regarding a study design that finds a trade-off between ensuring statistical power on the one hand, and decreasing the burden due to frequent study visits for the EB patients on the other hand, by optimal design methodology. However, the latter two aspects are outside the scope of this manuscript.

3 | METHODS

In the present work, wherever appropriate and possible, we have paid attention to the key quality criteria of "neutral comparisons" along the lines of Boulesteix et al. (2013). The methods that are subsequently employed in the simulation studies (4) are first introduced theoretically, differences and similarities are examined neutrally, and a common basic framework is defined. The main focus is then on the simulation study itself where the methods are compared. The emphasis is on already existing and rationally well justifiable methods, without having a preferred analysis method specified in advance. Moreover, a reference method is explicitly omitted in order to direct the focus only to the methods presented, and not to compare them to conventional methods that are expected to be inferior "per construction" (e.g., because they make overly simplifying assumptions).

In addition, all simulation results are reported, even negative results, in order to be transparent, and to point out possible problems and limitations in the application of the methods. Finally, it should be emphasized that the coauthors represent a consortium with well-balanced research interests and areas of expertise. The fact that they are collaborating in the EBStatMax project indicates their willingness to learn from each other, and to openly discuss advantages and drawbacks of the respective methods.

In the motivating example on EB, we have a longitudinal crossover design (Section 2). Hence, every subject is observed repeatedly at t time points (e.g., in the motivating example, we have $t = 4$ time points per period, namely, baseline, after 2 weeks, after 4 weeks—end of treatment—and after 3 months—end of follow-up) within each of the two periods. In the first treatment period, the subjects were randomly assigned to either placebo or verum (= Diacerin treatment); in the second period, the treatments were switched. To simplify notation, we assume without loss of generality that the first n_1 subjects were randomized to placebo in the first period, and the remaining $n - n_1$ subjects received verum. So, formally, we have pairs $(\mathbf{X}_{1k}^{(1)'}, \mathbf{X}_{2k}^{(2)'})'$, $k \in \{1, 2, \dots, n_1\}$, and $(\mathbf{X}_{2k}^{(1)'}, \mathbf{X}_{1k}^{(2)'})$, $k \in \{n_1 + 1, n_1 + 2, \dots, n\}$ of random vectors $\mathbf{X}_{ik}^{(j)} = (X_{ik1}^{(j)}, \dots, X_{ikt}^{(j)})'$, where $i \in \{1 = \text{placebo}, 2 = \text{verum}\}$ denotes the group within a particular period $j \in \{1, 2\}$, and $k \in \{1, 2, \dots, n\}$ is the subject index. 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 $s \in \{1, \dots, t\}$ by $F_{is}^{(j)}$. It should be noted that no specific parametric assumptions are made on $F_{is}^{(j)}$, other than that F is nondegenerate. Moreover, the total number of observations is denoted by $N = 2nt$. In the EB example (Section 2), we have $t = 4$; however, we would like to emphasize that all methodologies presented in the remainder of this manuscript are also applicable to the more general case of an arbitrary number of t repeated measures per period.

3.1 | A nonparametric rank-based approach—nparLD

The `R` package `nparLD` provides easy and user-friendly access to robust rank-based methods for the analysis of longitudinal data in factorial settings. For model classification purposes, nparLD uses a notation system for frequently used factorial designs depending on the number of factors. To this end, the factor that stratifies samples into independent groups is called a *whole-plot factor*, while the factor coding for repeated measurements is called a *subplot-factor* (Noguchi et al., 2012).

Accordingly, the particular designs are denoted by a name of the form

$$Fx - LD - Fy,$$

where x and y are the number of whole- and subplot factors, respectively, while "LD" stands for "longitudinal data." In the EB example, there is the subplot factor *time* (with four levels) and the whole-plot factor *treatment group* (with two levels). This results in the "F1–LD–F1" design, and the corresponding `R` function is `f1.ld.f1`. Formally, it is an implementation of different rank-based hypothesis tests for the hypotheses of no main effect A (i.e., the between-subject factor "group" within a particular period), no main time effect T, and no interaction effect AT between A and T. These hypotheses are expressed in terms of marginal distribution functions as follows:

- $H_0(A) : \bar{F}_{1\cdot}^{(j)} = \bar{F}_{2\cdot}^{(j)}$,
- $H_0(T) : \bar{F}_{\cdot 1}^{(j)} = \cdots = \bar{F}_{\cdot t}^{(j)}$,
- $H_0(AT) : F_{is}^{(j)} = \bar{F}_{i\cdot}^{(j)} - \bar{F}_{\cdot s}^{(j)} + \bar{F}_{\cdot\cdot}^{(j)}\ , i = 1, 2\ ; \ s = 1, .., t,$

where $\bar{F}_{i\cdot}^{(j)}$ denotes the cumulative distribution function (CDF) for group i averaged across the time points, $\bar{F}_{\cdot s}^{(j)}$ the average CDF at time point s across the two groups within a particular period, and $\bar{F}_{\cdot\cdot}^{(j)}$ the overall average CDF across all group and time levels. In the context of the motivating epidermolysis bullosa example, we are primarily interested in answering the research question whether the longitudinal profiles of the VAS scores differ between verum and placebo; therefore, we will only consider $H_0(AT)$ in Sections 4 and 5 below. It should be noted that within this framework, the period $j \in \{1, 2\}$ cannot be included as an additional factor into the model. Consequently, the empirical evaluations in Sections 4 and 5 have been conducted for each period $j \in \{1, 2\}$ separately. However, since the analyses have adopted exactly the same approach for both periods, we will drop the period index in the sequel for sake of simplicity.

For testing the above-mentioned hypotheses, one may use the Wald-type statistic (WTS) or the ANOVA-type statistic (ATS). In the manuscript, we only focus on the latter, because the WTS would need considerably larger sample sizes to maintain the prespecified type I error level (Brunner et al., 2002). The ATS is defined as

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

where C is a suitable contrast matrix, $T = C^T[CC^T]^- C$ the projection matrix, and tr denotes the trace of the respective matrices. A suitable contrast matrix C for this setting (i.e., for testing for an interaction effect) would thereby be $C = \mathbf{P}_a \otimes \mathbf{P}_b$, where $\mathbf{P}_a = \mathbf{I}_a - \frac{1}{a}\mathbf{J}_a$ and $\mathbf{P}_b = \mathbf{I}_b - \frac{1}{b}\mathbf{J}_b$ denote the a- and b-dimensional centering matrices, respectively, and $\otimes$ denoting the Kronecker product (for more details on contrast matrices, see Brunner et al., 2019). The quantity $\hat{\mathbf{p}}$ represents the vector of *estimated relative effects* $\hat{p}_{11}, \dots, \hat{p}_{1t}, \hat{p}_{21}, \dots, \hat{p}_{2t}$, and $\hat{V}$ is the corresponding empirical covariance matrix. These estimators are obtained by replacing the CDFs by their empirical counterparts in the following definition of the *relative effect*:

$$p_{is} := \int H dF_{is} \tag{2}$$

for $i \in \{1, 2\}, s \in \{1, \dots, t\}$, where H denotes the average over the marginal distribution functions $F_{11}, \dots, F_{2t}$. Hence, the period index $j \in \{1, 2\}$ has been omitted, since the periods have to be considered separately for the methods under consideration. By rewriting (2) using probabilities, one can show that p_{is} quantifies the tendency that a subject in group i at time point s has a higher VAS score (i.e., feels more severe pain/pruritus) than on average. Moreover, by doing some algebra, an expression of the above-mentioned estimators $\hat{p}_{is}$ of p_{is} that is based on ranks can be derived. Therefore, the relative effect estimators and the corresponding empirical covariance matrix $\hat{V}$ are rank-based estimators. Consequently, also, the ATS defined in (1) is considered a rank-based hypothesis testing approach. Note that in this manuscript, emphasis rests on specifying appropriate hypotheses to simulate the power and the type I error. That is, estimation is not discussed in the present manuscript; for more details on the relative effect, its estimators, and corresponding theoretical results, see Brunner et al. (2019, Ch. 2). In order to obtain a p-value, an approximation to the sampling distribution of $A_n(C)$ by a $F_{(\hat{f},\infty)}$ distribution is used, where

$$\hat{f} = \frac{(tr(T\hat{V}))^2}{tr(T\hat{V}T\hat{V})}.$$

For theoretical details on longitudinal rank-based estimation and hypothesis testing approaches, see Brunner et al. (2002) and Noguchi et al. (2012).

3.2 | Generalized pairwise comparisons

The nonparametric GPC method is an extension of the Gehan test (Gehan, 1965) for multivariate outcomes (Buyse, 2010; Finkelstein & Schoenfeld, 1999; Pocock et al., 2012; Verbeeck et al., 2019). With a single outcome and no missing data, the GPC test is a linear transformation of the Mann–Whitney test (Mann & Whitney, 1947; Verbeeck et al., 2021). Using the notations introduced at the beginning of Section 3, we consider within subject pairs $(\mathbf{X}_{11}^{(1)'}, \mathbf{X}_{21}^{(2)'})', \dots (\mathbf{X}_{1n_1}^{(1)'}, \mathbf{X}_{2n_1}^{(2)'})', (\mathbf{X}_{2n_1+1}^{(1)'}, \mathbf{X}_{1n_1+1}^{(2)'})', \dots (\mathbf{X}_{2n}^{(1)'}, \mathbf{X}_{1n}^{(2)'})'$ of random vectors $\mathbf{X}_{ik}^{(j)} = (X_{ik1}^{(j)}, \dots, X_{ikt}^{(j)})'$, where $i \in \{1, 2\}$ denotes the group within a particular period, k is the subject index, and $j \in \{1, 2\}$ denotes the period.

The GPC method evaluates the longitudinally collected VAS scores, $\mathbf{X}_{ik}^{(j)}$, by constructing all possible pairs between and within subjects, thus taking one subject from each treatment group, and subsequently assigning a score to each pair. As pairs are constructed between treatment arms, period effects are ignored in GPC and the period index j is omitted in the notation related to GPC. So, formally, we set $F_{is} := F_{is}^{(1)} = F_{is}^{(2)}$ for $i \in \{1, 2\}, s \in \{1, 2, \dots, t\}$. In the longitudinal crossover design of the EB example (Section 2), there are several options of constructing pairs of the longitudinally collected VAS scores. A summary measure per period can be constructed, which is compared per pair (for more details, see Section 3.2.1). Alternatively to this univariate evaluation, the longitudinal VAS scores can be compared in a multivariate way by comparing the VAS scores per time point between pairs (for more details, see Section 3.2.2). Often, in GPC, the components within a pair are considered as independent, but this ignores the crossover design. Alternatively to this unmatched GPC, pairs can be restricted to compare treatment arms only within subjects, in a matched GPC (Pocock et al., 2012). This matched GPC approach takes the crossover effect into account in the sense of incorporating the within-subject association. Either way, per pair, a score $U_{k\ell}$ corresponding to the comparison of the univariate or multivariate 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 $k, \ell \in \{1, \dots, n\}$ for the unmatched GPC and $k = \ell$ for the matched GPC):

$$U_{k\ell} = \begin{cases} 1, & \text{if } V_{1k} > V_{2\ell} \\ -1, & \text{if } V_{1k} < V_{2\ell} \\ 0, & \text{if } V_{1k} = V_{2\ell}. \end{cases} \tag{3}$$

Using the calculated scores, various statistics can be constructed: For example, the net benefit (Δ) (Buyse, 2010), win ratio (Pocock et al., 2012), or win odds (Brunner et al., 2021). The unmatched net benefit, for example, is the sum of the scores divided by the total number of pairs:

$$\Delta_{unm} = \frac{1}{n^2} \sum_{k=1}^{n} \sum_{\ell=1}^{n} U_{k\ell},$$

and has values between −1 and 1, which is easy to interpret as the difference of probability that a random subject will do better on active treatment than on placebo minus the reverse. Many inferential tests have been proposed for unmatched GPC statistics (Verbeeck et al., 2020). Most inference is based on testing the null hypothesis:

$$H_0 : P(V_{1k} < V_{2\ell}) = P(V_{1k} > V_{2\ell}),$$

but only the exact permutation test has good small sample properties (Anderson & Verbeeck, 2019). The inference of the closed-form exact permutation test is based on the more restrictive null hypothesis:

$$H_0 : \bar{F}_{1.} = \bar{F}_{2.},$$

with $\bar{F}_{i.}$ the distribution of the VAS in treatment group i. The test is based on the permutation distribution of the GPC statistic, which has been empirically shown to be close to the standard normal distribution, even for data with as little as five observations per arm (Verbeeck et al., 2020). Therefore, the test statistic $z = \Delta/\sqrt{\mathrm{var}(\Delta)}$ might as well be compared to the appropriate standard normal quantile for obtaining a test decision.

Inference for the matched GPC was originally proposed for the win ratio summary measure and is based on the asymptotic normality of a binomial distribution (Pocock et al., 2012). However, this inference method ignores ties, is inappropriate for small samples, and is based on a variance estimation under the alternative hypothesis. It is more appropriate for small samples to resort to extensions of the exact paired test for ties. Of the extensions, the conditional sign test was found to be the uniformly most powerful test (Coakley & Heise, 1996; Dixon & Massey, 1951; Fagerland, 2012; Wittkowski, 1998), for sample sizes of 15–20 subjects. Since the matched net benefit, Δ_m, can be expressed as the difference of two probabilities, namely, the probability of "wins," π_w, estimated by $\widehat{\pi}_w = \frac{1}{n}N_w = \frac{1}{n}\sum_{k=1}^{n} U_k \mathbb{1}_{(U_k=1)}$, and the probability of "losses," π_l, estimated by $\widehat{\pi}_l = \frac{1}{n}N_l = \frac{1}{n}\sum_{k=1}^{n} U_k \mathbb{1}_{(U_k=-1)}$, the variance of this difference is:

$$\begin{aligned}\operatorname{var}(\Delta_m) &= \operatorname{var}(\pi_w) + \operatorname{var}(\pi_l) - 2\operatorname{cov}(\pi_w, \pi_l)\\ &= \frac{1}{n}[\pi_w(1-\pi_w) + \pi_l(1-\pi_l) + 2\pi_w\pi_l]\\ &= \frac{1}{n}[\pi_w + \pi_l - (\pi_w - \pi_l)^2].\end{aligned}$$

Under the null hypothesis of the conditional sign test, $H_0 : \pi_w = \pi_l$, the variance can be estimated by:

$$\widehat{\sigma^2}(\Delta_m) = \frac{1}{n}(\widehat{\pi}_w + \widehat{\pi}_l) = \frac{1}{n^2}(N_w + N_l).$$

Using the normal approximation to the multinomial distribution, a large-sample test for H_0 is derived as:

$$Z = \frac{\Delta_m}{\widehat{\sigma^2}(\Delta_m)} = \frac{N_w - N_l}{\sqrt{(N_w + N_l)}},$$

which is compared to the standard normal distribution.

3.2.1 | Univariate generalized pairwise comparisons

One option to evaluate the repeated measures in the EB trial via GPC is by constructing one summary measure per subject and per treatment period and compare (univariately) these summary measures between pairs. So, using the notation introduced in Section 3, we consider independent pairs $(S_{11}^{(1)}, S_{21}^{(2)}), \dots, (S_{1n_1}^{(1)}, S_{2n_1}^{(2)}), (S_{2\,n_1+1}^{(1)}, S_{1\,n_1+1}^{(2)}), \dots, (S_{2\,n}^{(1)}, S_{1\,n}^{(2)})$ of random variables, where $S_{ik}^{(j)}$ is an appropriate summary measure of the random vector $\mathbf{X}_{ik}^{(j)} = (X_{ik1}^{(j)}, \dots, X_{ikt}^{(j)})'$. For example, in this manuscript, we consider $S_{ik}^{(j)} = \sum_{s=1}^{t} X_{iks}^{(j)}$, because the sum of all VAS scores of a particular subject in period j might be a good indication of the severity of pain/pruritus in that period. Of course, one may argue that one must not take the sum of ordinally scaled measurements; yet, on the other hand, the analysis method that is described in what follows indeed considers the outcome (i.e., the sum of the VAS scores) as truly ordinal. Moreover, taking the sum of ordinally scaled variables as a new outcome measure does not do any harm as long as the resulting variable is not interpreted as a metric quantity—actually, taking sums of ordinally scaled variables is frequently done with clinical scales (e.g., a clinical score is often just the sum of different subitems or subscales).

Apart from that, analogously to the outline of the general GPC framework (see above), we drop the period index j in the sequel. So, in the pairwise comparison in (3), V_{ik} is replaced by the summary measures S_{ik}, $i \in \{1, 2\}, k \in \{1, 2, \dots, n\}$, and both the matched and unmatched GPC can be applied to the univariate summary measure. It is important to note that for a particular subject k, the summary measure S_{ik} has to be considered as missing if there is at least one missing observation in this treatment group i (clearly, taking only the sum of the remaining observations would be inappropriate). Moreover, the matched GPC (i.e., the conditional sign test for the VAS summary measures) additionally requires complete observations for *both* treatment conditions $i \in \{1, 2\}$ for a particular subject. The unmatched GPC thus allows for more data to be used than the matched GPC, yet with the obvious drawback that the correlations between periods resulting from the crossover design are not accounted for.

3.2.2 | Multivariate generalized pairwise comparisons

Rather than summarizing repeated measures in a single summary measure, GPC also allows for evaluating repeated measures by time point t in a prioritized and nonprioritized analysis.

In a prioritized GPC, which is the most commonly applied method in a multiple outcome setting, outcomes are prioritized by severity (the worst outcome ranked first). Per pair a score is assigned according to (3) on the first ranked outcome. Only when the score results in a tie, the next ranked outcome is evaluated in the pair, continuing until the last outcome (Buyse, 2010; Finkelstein & Schoenfeld, 1999; Pocock et al., 2012; Verbeeck et al., 2019). For repeated measures, the time points can be prioritized according to the clinical importance of a treatment effect. Specifically for the EBStatMax study, the VAS score after 4 weeks of treatment is most important, followed by the 3 months follow-up visit, the 2 week, and the baseline visit. Although the prioritized GPC handles missing data in a natural way, by assigning a score 0 for the missing outcome comparison and moving to the next outcome, we will use the same data for all methods in our simulations, thus excluding periods with missing VAS scores. Both the unmatched and the matched analysis can be applied to the prioritized GPC, which tests the null hypotheses and uses the test statistics as described in Section 3.2, with the important difference that the distributions being compared are now multivariate distributions (i.e., the first component corresponds to the distribution of the outcome at the top-prioritized time point, the second component is the conditional distribution of the outcome at the second-ranked time point, given that there is a tie at the top-ranked time point, and so forth).

In a nonprioritized GPC, all pairs are evaluated according to (3) on each repeated measure s, resulting in scores $U_{k\ell s}$. The unmatched net benefit is then the sum of these scores divided by the total number of pairs and the number of repeated measures:

$$\Delta = \frac{1}{n^2 s} \sum_{k=1}^{n} \sum_{\ell=1}^{n} \sum_{s=1}^{t} U_{k\ell s},$$

which again has an easily interpretable value between −1 and 1 (Ramchandani et al., 2016; Verbeeck et al., 2019). While the univariate GPC for repeated measures first sums up observations and then compares, the nonprioritized GPC first compares and then sums. Since the probability of “wins” and “losses” is not straightforward to calculate from $U_{k\ell s}$, we will not perform matched analyses for the nonprioritized GPC. Inference of the unmatched nonprioritized GPC is based on the exact permutation test, where the null hypothesis is formulated using the distribution functions under verum and placebo, that is, the joint distributions of the outcomes at the different time points (Ramchandani et al., 2016; Verbeeck et al., 2019).

The correlation between repeated measures within a period is accounted for by both the prioritized as well as the nonprioritized GPC in a different way. The correlation is reflected in the net benefit statistic in the prioritized GPC, whereas it is reflected in the variance of the net benefit distribution in the nonprioritized GPC (Verbeeck et al., 2019).

4 | SIMULATION STUDY

4.1 | Description of the simulation setting

Different approaches were considered for the simulations. Primarily, it was decided to compare the longitudinal measurements of the raw values between the treatment groups. However, an alternative approach using the change of baseline was also considered and is briefly described in the end of Section 4.2. Further results of this second simulation approach can be found in the Supporting Information.

The EB trial data set (see Section 2) was used for the simulations. Recall that this is a longitudinal data set from a crossover study with n subjects and 2 periods, with $t = 4$ time points (baseline, 2 weeks, 4 weeks (= posttreatment) and 3 months (= follow-up)) for VAS measures per period. In our simulations, observations were not considered at individual time point levels, but rather grouped into blocks of $t = 4$ time points each. Hence, adopting the notation from Section 3, for each subject $k \in \{1, 2, \ldots, n\}$, we have a pair $(\mathbf{X}_{1k}, \mathbf{X}_{2k})$ of vectors with four components each, corresponding to the repeated measures under treatment conditions 1 = placebo and 2 = verum (for notational simplicity, without loss of generality, we have dropped the index j denoting the trial period, because we did not simulate any period effects, see below), that is, every patient has a “placebo” and a “verum” block. Now, in each simulation run, the blocks $\mathbf{X}_{ik}$ were randomly permuted across all subjects and treatment conditions. The reason for adopting this approach was that by sampling from the EB data set, we kept original data characteristics, thus avoiding any additional (parametric) assumptions. For power simulations, however, we, of course, had to implement some additional steps to the above-mentioned permutation approach, in order to simulate different effect scenarios. These steps were implemented as follows:

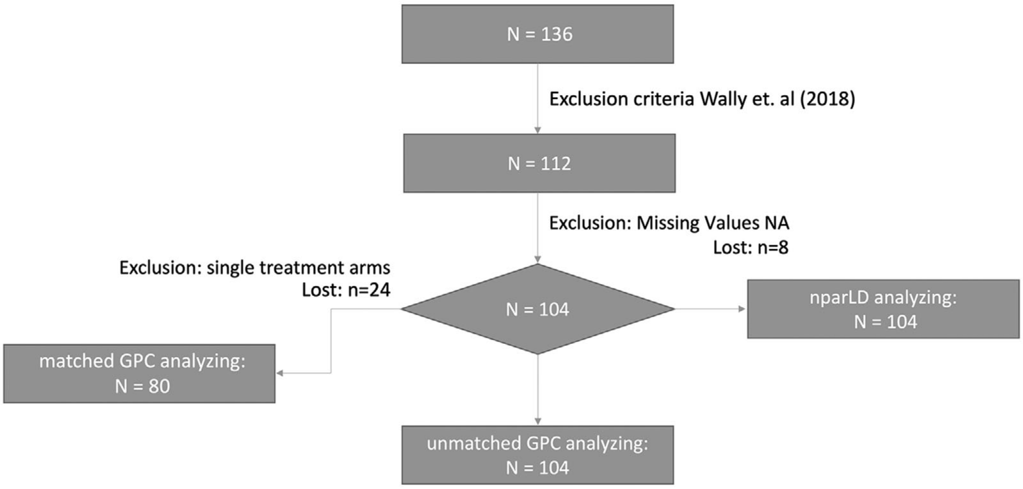

FIGURE 2 Flowchart showing numbers of subjects who were analyzed in the simulation study..

1. Random variables $Z_k \overset{iid}{\sim} \mathcal{D}$, $k \in \{1, 2, \dots, n\}$, 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 $\mu_{\text{log}} \in \{0.2, 0.6, 0.9\}$, respectively. Negative values (which possibly occur with normally distributed Z_k) were replaced with zeroes.
2. These random variables $(Z_k)_{k=1}^n$ were subsequently added to the observations from the placebo group. More precisely, 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 posttreatment visit) only.
   - Scenario 2: The random variables were added to the VAS scores under placebo at the third time point (i.e., the post-treatment visit), and additionally, $(Z_k/2)_{k=1}^n$ were added to the VAS scores under placebo at the fourth time point (i.e., the follow-up visit).
3. The corresponding “new” observations resulting from step 2 were appropriately cut off (maximum VAS value is 10.0) and rounded to one decimal place, if required, in order to adequately represent VAS scores.

This setup for the power simulations is closely aligned with clinical expertise in several ways. First, the distributions have been selected such that a symmetric (i.e., normal) as well as an asymmetric setting (i.e., lognormal) is covered. Second, the parameters of these distributions have been chosen such that the expected values (i.e., the shift effects) are equal to 2, 3, and 4, respectively (or at least very close to these values after truncation and rounding). An average difference in VAS scores of 3 between placebo and verum conditions would be regarded as a clinically meaningful effect; the scenarios with shift effects 2 and 4 have been added in order to have a broader range of different settings for the power simulations at hand. Third, the two above-mentioned scenarios in step 2 correspond to the assumptions of no and some long-term effects, respectively; both scenarios are clinically plausible, taking into account that the Diacerein cream is an intervention—like many others in EB—which has no curative, but mainly a symptom-relieving effect.

For each combination of the distribution/parameter settings and effect scenarios described above, $R = 5000$ simulation runs were performed. The resulting empirical power values are based on using the two-sided level $\alpha = 0.05$. All simulations were carried out using the statistical software R (Version: 4.0.3, R Core Team (2021)).

It is important to note that the inclusion and exclusion criteria (see Figure 2) for the data underlying the simulations were intended to resemble the setting of the original study of Wally et al. (2018) as closely as possible, and to avoid preoptimizing effects for a specific setting or sample size in which the methods may work better than existing approaches, as Boulesteix et al. (2013) point out as a criterion for a neutral comparison. Therefore, 24 observations (i.e., six blocks of four longitudinally collected VAS measurements each) were excluded from the initial set of 136 observations (4 time points × 2 periods × 17 subjects) for clinical and study-related reasons. In addition, as our proposed methods require complete observations within periods, blocks with NAs had to be excluded as well. This led to a further reduction of the number of observations to $112 - 8 = 104$ observations, which formed the core data set for the simulations. Furthermore, since for the matched GPC method, those subjects that have one single block only (i.e., either the placebo or the verum block is missing) had to be excluded completely, the actual number of observations for that particular method was reduced to 80

TABLE 1 Power simulation result for the ordinal outcome “pruritus” and “pain” with varying log-normal effects and normal effects (with σ_{log} and $\sigma_{\text{norm}} = 1$) and scenarios 1 and 2 using the method nparLD for treatment period 1.

	Power			
	Pain		**Pruritus**	
	Scenario 1	**Scenario 2**	**Scenario 1**	**Scenario 2**
nparLD				
$\mu_{\text{log}} = 0.2$	0.2402	0.2604	0.2846	0.3124
$\mu_{\text{log}} = 0.6$	0.3476	0.3566	0.3642	0.3796
$\mu_{\text{log}} = 0.9$	0.4522	0.4534	0.4418	0.4430
$\mu_{\text{norm}} = 2$	0.2800	0.2874	0.3000	0.3246
$\mu_{\text{norm}} = 3$	0.5112	0.4862	0.4532	0.4532
$\mu_{\text{norm}} = 4$	0.7322	0.6846	0.6040	0.5692

(4 time points × 2 periods × 10 subjects). Since the matched GPC variants had considerably fewer observations for the analysis, an additional simulation was also performed using the unmatched GPC variants based on the reduced number of 80 observations (see Supporting Information: Tables S7 and S8).

4.2 | Results

In what follows, the results of the above-mentioned simulation scenarios based on the EB trial data are reported for the nparLD approach as well as for five variants of the GPC method (univariate unmatched and matched, prioritized unmatched and matched, and nonprioritized unmatched). It should be mentioned that for nparLD, the hypothesis tests had to be conducted separately for periods 1 and 2, whereas the GPC methods did not require such a split. However, since the simulation setup has been designed such that there is no period effect, only the nparLD results for the first period are reported, and the analogous tables for period 2 (which are very similar, yet with a somewhat more liberal type I error) are provided in the Supporting Information (see Tables S9 and S10). Moreover, for nparLD, only the interaction effect is considered, and therefore, only the null hypothesis $H_0^F(AT)$ is used (see Section 3). Apart from that, note that the GPC variants allow for both one- and two-sided testing. In order to ensure consistency of the presentation and comparability of the methods under consideration, only the simulation results for the two-sided case are reported in the main body of the manuscript; their one-sided counterparts can be found in the Supporting Information (see Tables S9, S11, and S12). Additionally, as mentioned before, a second simulation scenario was also performed using a change from baseline approach. A short summary is provided at the end of this section, and the tables and numerical results can be found in the Online Appendix (see Supporting Information Tables S14–S18). Besides, the simulation results for the data set reduced to 80 observations for the unmatched GPC variants are also presented in the Online Appendix in order to provide additional information for comparison (see Supporting Information Tables S7 and S8). Finally, note that all simulations were run twice, considering the VAS score for pruritus and pain as outcomes, respectively.

Regarding the type I error, almost all considered methods exhibit an adequate type I error control (rank-based nonparametric ANOVA-type test from nparLD: $\alpha_{\text{pain}} = 0.0560$, $\alpha_{\text{pruritus}} = 0.0586$; univariate unmatched GPC: $\alpha_{\text{pain}} = 0.0492, \alpha_{\text{pruritus}} = 0.0468$; prioritized unmatched GPC: $\alpha_{\text{pain}} = 0.0490, \alpha_{\text{pruritus}} = 0.0472$; nonprioritized unmatched GPC: $\alpha_{\text{pain}} = 0.0508, \alpha_{\text{pruritus}} = 0.0496$). However, possibly due to the reduction in sample size in the analysis with matched GPC, conservative type I error rates occurred in both the univariate ($\alpha_{\text{pain}} = 0.0344, \alpha_{\text{pruritus}} = 0.0240$) and the prioritized matched GPC ($\alpha_{\text{pain}} = 0.0252, \alpha_{\text{pruritus}} = 0.0214$). It is also interesting to note here that due to the aim of neutral comparability, only two-sided tests have been used. However, for the one-sided counterpart of the matched GPC, one would see a somewhat liberal type I error rate (see Supporting Information Table S9). Violations of the target type I error level were also observed for nparLD applied to data from period 2 (see Supporting Information Table S9). This could be caused by the underlying original data, because there was a considerable effect in period 2 (see Section 5). Perhaps, due to the very limited sample size, maybe, this effect has not fully “disappeared” in the permutation-based simulations.

Regarding the power simulation with the method nparLD, as shown in Table 1, it can be observed that the power is larger with normal effects than with asymmetric log-normal effects. Moreover, it can be seen that for $\mu_{\text{norm}} = 4$, the power is already above 0.6 in almost every case. It is interesting to note that there is a pronounced difference in the magnitude of

TABLE 2 Power simulation result for the ordinal outcome "pruritus" and "pain" with varying log-normal effects and normal effects (with σ_{log} and $\sigma_{\mathrm{norm}} = 1$) and scenarios 1 and 2 using the two-sided univariate matched and unmatched GPC method (caution is needed in interpreting the matched GPC results, due to an uncontrolled type I error; therefore, these rates are marked with †).

	Power			
	Pain		Pruritus	
	Scenario 1	Scenario 2	Scenario 1	Scenario 2
Univariate matched GPC				
$\mu_{\mathrm{log}} = 0.2$	0.0450†	0.0650†	0.0516†	0.0764†
$\mu_{\mathrm{log}} = 0.6$	0.0614†	0.1074†	0.0786†	0.1318†
$\mu_{\mathrm{log}} = 0.9$	0.0948†	0.1628†	0.1084†	0.1910†
$\mu_{\mathrm{norm}} = 2$	0.0504†	0.0798†	0.0542†	0.0856†
$\mu_{\mathrm{norm}} = 3$	0.0814†	0.1536†	0.0914†	0.1688†
$\mu_{\mathrm{norm}} = 4$	0.1280†	0.2750†	0.1508†	0.2776†
Univariate unmatched GPC				
$\mu_{\mathrm{log}} = 0.2$	0.1106	0.1794	0.1398	0.2196
$\mu_{\mathrm{log}} = 0.6$	0.1768	0.3054	0.2024	0.3458
$\mu_{\mathrm{log}} = 0.9$	0.2638	0.4618	0.2902	0.4984
$\mu_{\mathrm{norm}} = 2$	0.1236	0.2218	0.1616	0.2636
$\mu_{\mathrm{norm}} = 3$	0.2284	0.4374	0.2626	0.4658
$\mu_{\mathrm{norm}} = 4$	0.3800	0.7016	0.4086	0.6752

the power between the effect sizes $\mu_{\mathrm{norm}} = 2$ and $\mu_{\mathrm{norm}} = 3$. However, the power is slightly smaller in some cases with normal effects for the second scenario, where in addition to the posttreatment effect, there is also a (smaller) effect at the follow-up visit (see above). At first sight, this seems counterintuitive; however, this is due to the fact that a group–time interaction is tested here: Recall that differences between treatment and placebo regarding the VAS trajectories over time are tested here. Now, since the change from posttreatment to follow-up is more pronounced in scenario 1, this leads to a more pronounced interaction between group and time. and thus, to a higher power than for scenario 2. In general, it is remarkable that the nparLD method yields quite large power values despite the fact that only one period, and thus, an even smaller sample is considered.

Table 2 illustrates the power results of the two-sided univariate matched and unmatched GPC method. It can be observed that in scenario 1, the power is always lower than in the nparLD method and the maximum value reaches only 0.4086. On the contrary, for scenario 2, while the power of the univariate matched GPC is always lower than nparLD, the power is almost always higher than nparLD for the univariate unmatched GPC. To sum up, comparing the results of the unmatched GPC with nparLD reveals that overall, it cannot be determined which method shows consistently higher power. The difference in power between scenarios 1 and 2 is larger in both univariate GPC methods compared to nparLD. Apart from that, it can be observed that the power for the unmatched GPC is much higher than for the matched GPC, which is possibly due to a larger sample size in the former case. Moreover, the conservative behavior of the univariate matched GPC regarding type I error control might also play a role. Therefore, the power rates of the matched GPC variant in Table 2 were marked with †, because one might have to be careful with the interpretation here.

Tables 3 and 4 illustrate the power results of the nonprioritized unmatched GPC as well as the two-sided prioritized matched and unmatched GPC method. It can be observed that the power is, particularly for the prioritized unmatched GPC, higher than for any other method. Especially for the simulation with asymmetric log-normal effects, the power is considerably higher in comparison to the other methods and already takes 0.6404 as the smallest value. However, this high power may also be due to the prioritization of the time points (see Section 6). Furthermore, again, the power is higher for the second scenario in each case, as already seen for the univariate GPC methods, but the difference between the two scenarios is now much smaller. Additionally, the power of the nonprioritized unmatched GPC is smaller in scenario 1, while it is much higher in scenario 2. Having multiple visits with a treatment effect thus obviously increases the power of the nonprioritized GPC much. Yet, again, any power values of the matched GPC variant were flagged using †, as one must be cautious in interpretation due to the conservative type I error rate.

In summary, it can be concluded for the comparisons that the prioritized unmatched GPC method was consistently achieving a greater power, even with log-normal effects. Of course, this is not a surprise, because the prioritization is

TABLE 3 Power simulation result for the ordinal outcome "pruritus" and "pain" with varying log-normal effects and normal effects (with σ_{log} and $\sigma_{\text{norm}} = 1$) and scenarios 1 and 2 using the two-sided nonprioritized unmatched GPC method.

	Power			
	Pain		**Pruritus**	
	Scenario 1	**Scenario 2**	**Scenario 1**	**Scenario 2**
Nonprioritized unmatched GPC				
$\mu_{\text{log}} = 0.2$	0.1322	0.3244	0.2370	0.5854
$\mu_{\text{log}} = 0.6$	0.1672	0.4304	0.2590	0.6570
$\mu_{\text{log}} = 0.9$	0.1998	0.5376	0.2958	0.7236
$\mu_{\text{norm}} = 2$	0.1346	0.3256	0.2390	0.5868
$\mu_{\text{norm}} = 3$	0.1910	0.4858	0.2742	0.6942
$\mu_{\text{norm}} = 4$	0.2584	0.6724	0.3052	0.7718

TABLE 4 Power simulation result for the ordinal outcome "pruritus" and "pain" with varying log-normal effects and normal effects (with σ_{log} and $\sigma_{\text{norm}} = 1$) and scenarios 1 and 2 using the two-sided prioritized matched GPC method (caution is needed in interpreting the matched GPC results, due to an uncontrolled type I error; therefore, these rates are marked with †).

	Power			
	Pain		**Pruritus**	
	Scenario 1	**Scenario 2**	**Scenario 1**	**Scenario 2**
Prioritized matched GPC				
$\mu_{\text{log}} = 0.2$	0.2358†	0.2372†	0.4056†	0.4176†
$\mu_{\text{log}} = 0.6$	0.3266†	0.3280†	0.4776†	0.4936†
$\mu_{\text{log}} = 0.9$	0.4254†	0.4288†	0.5496†	0.5652†
$\mu_{\text{norm}} = 2$	0.2534†	0.2552†	0.4254†	0.4284†
$\mu_{\text{norm}} = 3$	0.4218†	0.4276†	0.5294†	0.5308†
$\mu_{\text{norm}} = 4$	0.6186†	0.6236†	0.6016†	0.6050†
Prioritized unmatched GPC				
$\mu_{\text{log}} = 0.2$	0.6404	0.6432	0.8808	0.8880
$\mu_{\text{log}} = 0.6$	0.7860	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	0.8890	0.8910
$\mu_{\text{norm}} = 3$	0.8834	0.8890	0.9500	0.9500
$\mu_{\text{norm}} = 4$	0.9778	0.9788	0.9528	0.9546

aligned with the simulation framework. Overall, however, it can be stated that each GPC method, except the univariate matched GPC, achieves a higher power for the second scenario (i.e., effects added at two time points), while nparLD mostly achieves higher power in the first scenario (i.e., effects added at one time point).

In addition, the unmatched GPC variants achieve higher power compared to the matched GPC counterparts, and prioritizing the time points (unmatched) has a big impact on power, which is close to 80% or 90% in most cases. When adding a normally distributed treatment effect, nparLD shows a higher power, even with a quite limited sample size, than the univariate GPC variants and the prioritized matched GPC method.

As mentioned at the beginning, an alternative simulation option was also considered by looking at the change from baseline. In practice, this is reasonable because the variability is reduced and, in theory, it may also increase the power. So, it was decided to apply this approach to all methods to see how the power values behave in this scenario. As a consequence of considering change of baseline, the baseline time point was excluded for each period in this analysis approach. As a result, it was found that the type I error rate remained approximately the same (see Supporting Information Table S14). However, for all GPC variants in all scenarios, the power analysis revealed lower power values than in the original setup using the raw values (see Supporting Information Table S6–S18). In comparison, the power of nparLD changed only slightly (see Supporting Information Table S15).

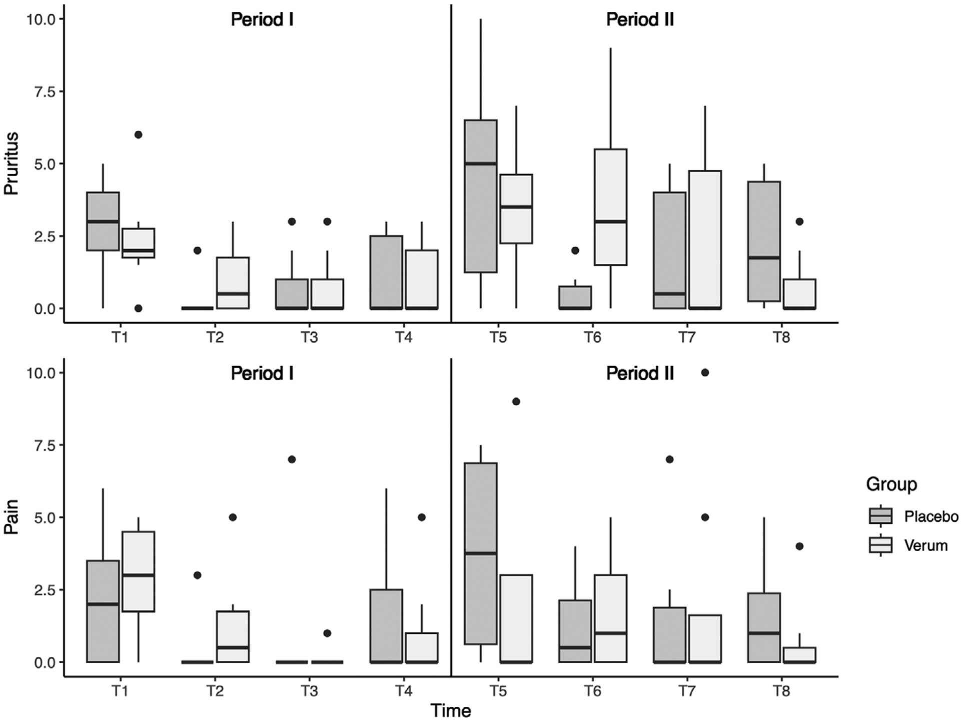

FIGURE 3 Profile plot showing boxplots for the outcomes “pain” and “pruritus” for treatment periods I and II.

TABLE 5 Resulting interaction effect of time and group for the ordinal outcome “pruritus” and “pain” in the original data set using nparLD with the ANOVA-type statistics.

	Test statistic	*p*-Value
Pruritus		
nparLD Period 1	0.7193	0.5018
nparLD Period 2	3.9737	0.0172
Pain		
nparLD Period 1	0.5769	0.5978
nparLD Period 2	0.5167	0.5596

5 | ANALYSIS OF EPIDERMOLYSIS BULLOSA TRIAL

To ensure a neutral evaluation of the methods presented for the real-life data example, Boulesteix et al. (2013) were used as a guidance again. Therefore, the data set used for the real-life data example fits into the application context of the described disease EB. This data set also meets the requirement of being as representative as possible of the application context and the domain of interest, given that crossover designs are frequently used as a remedy for small-sample size issues in Epidermolysis Bullosa trials in particular, as well as in rare diseases studies in general. In addition, any cases that were excluded in the original study are again explicitly excluded, to be in line with reality, and to avoid preoptimizing effects. To visualize the results the following two time-profile plots for the outcomes “pain” and “pruritus” are provided in Figure 3.

For the GPC variants, the *p*-values are calculated in the same way as described in Section 3.2, that is, based on calculating the wins, losses, and ties as well as the net benefit (with a 95% confidence interval). These more detailed results can be found in Supporting Information Table S13. In addition, the prioritization for the multivariate GPC is done as described in Section 3.2.2.

From Tables 5 and 6, several trends that have already been observed in the simulation study (e.g., matched GPC tends to have lower power than its unmatched counterpart; unmatched prioritized GPC is most powerful) are only partially seen in the case study. Maybe this is due to the particular dependency structure in the real-life data that are induced by the

TABLE 6 Resulting two-sided *p*-value and test statistic for the GPC variants applied to the original data set for the ordinal outcome “pruritus” and “pain.”

	Test statistic	*p*-Value
Pruritus		
matched univariate GPC	0.6325	0.5271
unmatched univariate GPC	0.3853	0.7000
matched prioritized GPC	0.6325	0.5271
unmatched prioritized GPC	0.8721	0.3832
unmatched nonprioritized GPC	0.6855	0.4931
Pain		
matched univariate GPC	1.0000	0.3173
unmatched univariate GPC	0.0773	0.9384
matched prioritized GPC	0.0000	1.0000
unmatched prioritized GPC	0.6418	0.5210
unmatched nonprioritized GPC	0.2237	0.8230

crossover design. Another remarkable finding concerns the small *p*-value of nparLD for period 2. This points to the fact that in a crossover design, one has to find a trade-off between taking specific effects within periods into account on the one hand, and ignoring periods by pooling the data together in order to increase power on the other hand.

6 | DISCUSSION AND OUTLOOK

The aim of this research was to neutrally compare different methods for longitudinally measured ordinal outcomes in rare diseases. Due to the rarity of the diseases, one frequently has to face challenges related to small sample sizes, and often, crossover designs are employed. One particular example is a data set from Epidermolysis Bullosa (EB) research, in which, among other outcomes, ordinal measures of pain and pruritus were assessed on a VAS. Based on these data, we have set up simulation scenarios enabling a neutral comparisons between a nonparametric rank-based approach using the R package nparLD and various GPCs, including a univariate unmatched and matched approach, as well as a multivariate prioritized and nonprioritized method. Key criteria for selecting these methods were that they could account for the longitudinal and the crossover aspects as well as for the small sample size as well as possible. In addition, the methods considered in this paper have not yet been investigated thoroughly for the present study design. Other potentially promising methods, such as ordinal random-effects regression models or the well-known Friedman test, might be investigated in future work, although defining a unified empirical framework for systematic, neutral comparisons might be substantially more difficult, then.

A challenge in the neutral comparison of these methods was presented by the different underlying approaches. Using the nonparametric nparLD method, which is based on the relative effect, analyses could only be conducted for each period separately. Thus, the crossover aspect was partially lost. This poses a problem for the analysis of crossover designs, which are commonly used in clinical studies with rare diseases. In contrast, the univariate GPC method is based on summary measures, and hence, leads to a partial loss of longitudinal information. Moreover, the GPC method has not been studied yet for longitudinal data. In addition, there were also differences regarding the cases that had to be excluded. For example, neither the univariate GPC method nor nparLD could handle missing values, and thus, any treatment arms containing missing values had to be excluded. Only the prioritized GPC method can handle missing values. It should be noted, however, that this problem may no longer exist in the future with an updated version of nparLD, which is currently being developed. Furthermore, for the matched GPC method, all single treatment arms had to be excluded, because the method is based on a pairwise comparison between both periods. Of course, this additionally decreased the size of the sample in our simulations for this method.

Due to the fact that the aim was to mimic the original data set as closely as possibly, we have only considered a limited number of different simulation scenarios. In particular, we have not employed any sample size configurations that are different from the ones used in the EB trial. However, on the other hand, we tried to be as exhaustive as possible, when including several combinations of different scenarios regarding the effects over time, and distributional assumptions. Yet, no “heavy-tailed” distribution, for example, t distribution with 3 or 4 degrees of freedom only, was included. However, it is questionable whether this heavy-tailed distribution would be clinically plausible, given that the VAS scores only take values between 0 and 10.

When considering the results of the simulation study, only type I error and power were assessed. Thereby, the type I error was well controlled by almost all methods, only for the matched GPC method, it appeared to be rather conservative; and, for nparLD, we found a somewhat liberal behavior in period 2. This may be explained by looking at the effectively used sample size for this matched GPC method: Since only $N = 10$ subjects could be used, due to the above-mentioned exclusion of the single treatment arm observations, this sample size setting could be possibly too small (for simulation evidence regarding the conditional sign test, see Coakley & Heise, 1996 and Fagerland et al., 2013). The appropriateness of inference with the matched GPC may thus be limited to about $N = 15$. These findings are also supported by the additional simulation for the unmatched GPC variants with the reduced data set of $N = 10$ subjects. The result for those unmatched variants showed that the power decreased, but still remained at a high level (see Supporting Information Table S7). Also, the type I error was not particularly conservative for the unmatched variants in this setting (see Supporting Information Table S8).

Regarding the nparLD results for period 2, the effect seen for period 2 in the original data might be an explanation. Since the main aim was to perform a neutral comparison, we deliberately refrained from a “cherry-picking search” for a specific data set or simulation setup where the superiority of one particular method was expected to be more pronounced. Yet, further research could investigate the impact of different sample size variations on the performance of the matched GPC variants and nparLD. When reflecting upon the power simulation for the ANOVA-type method implemented in nparLD, it seemed rather counterintuitive at first sight that scenario 2 yielded less power than scenario 1, because an additional “long-term effect” was present in the former. However, this might be explained by the fact that the ANOVA-type test was used to test for group–time interactions. Indeed, from the interaction perspective, a time profile that shows an effect at the posttreatment time point and completely returns to the baseline levels at the follow-up visit (i.e., scenario 1) may be regarded as a more pronounced change over time compared to a setting where some portion of the effect is still present at the follow-up visit (thus rendering the change of the time profile under treatment relative to placebo less markedly).

The highest power overall is achieved with the prioritized unmatched GPC method. The prioritization of the time points has therefore a big impact on power. This makes sense, because the prioritization that was specified in the GPC closely corresponds to the simulation settings (i.e., in the simulations, the effect was added in each scenario at the posttreatment visit, which was, on the other hand, evaluated first in the prioritized analyses). One may argue that this contradicts the neutral comparison principle; however, we would like to emphasize that the reason for setting up the simulations as well as the prioritization in the GPC method in this way was clinical reasoning, and not any intention to favor this particular method. It should be highlighted that a different prioritization could also result in lower power. Apart from that, it is noteworthy that nparLD has a high power that is quite close to the prioritized GPC at least in some cases, although the simulations were performed with a smaller sample size (group sizes of 6 and 7), due to the separate testing of the periods. Therefore, this indicates that nparLD provides good results in terms of power with very small sample sizes. Apart from that, it is noteworthy that the power of all methods stayed similar or even decreased when considering change from baseline as outcome. So, obviously, the reduction in baseline variability did not translate into a gain in power. Nevertheless, the fact that the power stayed quite the same for nparLD might be considered as an advantage of this method, because it is invariant of the choice of the outcome. Further investigations are warranted to evaluate the above-mentioned findings more thoroughly.

When considering the real-life data example, one notices that while several trends were observed in the simulation study, none of them are necessarily recognizable in the selected data set example, which might be due to the particular data structure and the dependencies in the crossover design. It seems remarkable that by splitting the periods in the nparLD method for the outcome pruritus, a very small p-value in period 2 is obtained as a result. This effect is not shown by using the GPC methods. As a conclusion, this indicates that especially for longitudinal data in a small sample size crossover study, a certain trade-off must be made between increasing power and analyzing period-specific effects. However, a modified approach of the GPC methods could also consider the periods separately, so that potential period-specific effects become evident. Further research regarding this issue is currently conducted by one of the coauthors.

Although the simulations in the present manuscript are based on one particular data set from Epidermolysis Bullosa Research, the results might be generalized to other rare diseases, given that the basic characteristics of the data (i.e., ordinal outcomes obtained longitudinally in a two-period crossover trial) are similar. For example, in clinical trials on treatments of rare epilepsies, the posttreatment seizure frequency is usually considered as the primary outcome. Seizure frequency is often quantified using the Engel scale, which is an ordinal variable ranging from class I (seizure freedom) to class IV (no worthwhile improvement). To take another example, in patients with spinal cord injury, ordinal outcome measures such as the ASIA classification or specific QOL scores are frequently used. Yet, to be on the safe side, before generalizing our methodological recommendations, the simulation results presented in this manuscript would have to be reproduced in several simulations based on data from the above-mentioned medical research areas.

ACKNOWLEDGMENTS

Georg Zimmermann gratefully acknowledges the support of the WISS 2025 project "IDA-Lab Salzburg" (20204-WISS/225/197-2019 and 20102-F1901166-KZP). All authors gratefully acknowledge the funding of the "EBStatMax Demonstration Project" by the European Joint Programme on Rare Diseases (EU Horizon 2020 research and innovation program, grant agreement no. 825575).

CONFLICT OF INTEREST STATEMENT

The authors have declared no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

ORCID

Martin Geroldinger https://orcid.org/0000-0002-7858-323X
Georg Zimmermann https://orcid.org/0000-0002-8282-1034

REFERENCES

Anderson, W., & Verbeeck, J. (2019). Exact bootstrap and permutation distribution of wins and losses in a hierarchical trial. https://arxiv.org/pdf/1901.10928.pdf

Böhm, R., Westermann, P., Gleim, M., Cascorbi, I., Gruenewald, M., Herdegen, T., & Ohnesorge, H. (2021). High-dose spironolactone lacks effectiveness in treatment of fibromyalgia (rct). *European Journal of Pain*, *25*(8), 1739–1750.

Boulesteix, A.-L., Binder, H., Abrahamowicz, M., Sauerbrei, W., & for the Simulation Panel of the STRATOS Initiative. (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.

Boulesteix, A.-L., Wilson, R., & Hapfelmeier, A. (2017). Towards evidence-based computational statistics: Lessons from clinical research on the role and design of real-data benchmark studies. *BMC Medical Research Methodology*, *17*(1), 138.

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., Domhof, S., & Langer, F. (2002). *Nonparametric analysis of longitudinal data in factorial experiments*. John Wiley & Sons.

Brunner, E., Vandemeulebroecke, M., & Mütze, T. (2021). Win odds: An adaptation of the win ratio to include ties. *Statistics in Medicine*, *40*, 3367–3384.

Buyse, M. (2010). Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. *Statistics in Medicine*, *29*, 3245–3257.

Coakley, C. W., & Heise, M. A. (1996). Versions of the sign test in the presence of ties. *Biometrics*, *52*, 1242–1251.

Dixon, W., & Massey, F. (1951). *An introduction to statistical analysis*. McGraw-Hill.

Fagerland, M., Lydersen, S., & Laake, P. (2013). The mcnemar test for binary matched-pairs data: Mid-p and asymptotic are better than exact conditional. *BMC Medical Research Methodology*, *13*, 91.

Fagerland, M. W. (2012). t-Tests, non-parametric tests, and large studies – A paradox of statistical practice? *BMC Medical Research Methodology*, *12*, 78.

Finkelstein, D., & Schoenfeld, D. (1999). Combining mortality and longitudinal measures in clinical trials. *Statistics in Medicine*, *18*, 1341–1354.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, *32*(200), 675–701.

Gehan, E. (1965). A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika*, *52*(1/2), 203–223.

Hedeker, D., & Gibbons, R. D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics*, *50*, 933–944.

Hedeker, D., & Mermelstein, R. J. (2000). Analysis of longitudinal substance use outcomes using ordinal random-effects regression models. *Addiction*, *95*(11s3), 381–394.

Heller, G. Z., Manuguerra, M., & Chow, R. (2016). How to analyze the visual analogue scale: Myths, truths and clinical relevance. *Scandinavian Journal of Pain*, *13*(1), 67–75.

Kristensen, S. D., Gormsen, J., Naver, L., Helgstrand, F., & Floyd, A. K. (2021). Randomized clinical trial on closure versus non-closure of mesenteric defects during laparoscopic gastric bypass surgery. *British Journal of Surgery*, *108*(2), 145–151.

Kruskal, W. H., & Wallis, W. A. (1952). The use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, *47*, 583–621.

Lange, T., Deventer, N., Gosheger, G., Lampe, L. P., Bockholt, S., Schulze Boevingloh, A., & Schulte, T. L. (2021). Effectiveness of radial extracorporeal shockwave therapy in patients with acute low back pain – Randomized controlled trial. *Journal of Clinical Medicine*, *10*(23), 5569.

Mann, H., & Whitney, D. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, *18*(1), 50–60.

Noguchi, K., Gel, Y. R., Brunner, E., & Konietschke, F. (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://www.jstatsoft.org/index.php/jss/article/view/v050i12

Pocock, S., Ariti, C., Collier, T., & Wang, D. (2012). The win ratio: A new approach to the analysis of composite endpoints in clinical trials based on clinical priorities. *European Heart Journal*, *33*, 176–182.

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Ramchandani, R., Schoenfeld, D., & Finkelstein, D. (2016). Global rank tests for multiple, possibly censored, outcomes. *Biometrics*, *72*(3), 926–935.

Verbeeck, J., Deltuvaite-Thomas, V., Berckmoes, B., Burzykowski, T., Aerts, M., Thas, O., Buyse, M., & Molenberghs, G. (2021). Unbiasedness and efficiency of non-parametric and umvue estimators of the probabilistic index and related statistics. *Statistical Methods in Medical Research*, *30*(3), 747–768.

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., Spitzer, E., de Vries, T., van Es, G., Anderson, W., Van Mieghem, N., Leon, M., Molenberghs, G., & Tijssen, J. (2019). Generalized pairwise comparison methods to analyze (non)prioritized composite endpoints. *Statistics in Medicine*, *38*(30), 5641–5656.

Wally, V., Hovnanian, A., Ly, J., Buckova, H., Brunner, V., Lettner, T., Ablinger, M., Felder, T., Hofbauer, P., Wolkersdorfer, M., Lagler, F., Hitzl, W., Laimer, M., Kitzmüller, S., Diem, A., & Bauer, J. (2018). Diacerein orphan drug development for epidermolysis bullosa simplex: A phase 2/3 randomized, placebo-controlled, double-blind clinical trial. *Journal of the American Academy of Dermatology*, *78*(5), 892–901.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, *1*, 80–83.

Wittkowski, K. (1998). Versions of the sign test in the presence of ties. *Biometrics*, *54*(2), 789–791.

Zimmermann, G., Bolter, L.-M., Sluka, R., Höller, Y., Bathke, A. C., Thomschewski, A., Leis, S., Lattanzi, S., Brigo, F., & Trinka, E. (2019). Sample sizes and statistical methods in interventional studies on individuals with spinal cord injury: A systematic review. *Journal of Evidence-Based Medicine*, *12*(3), 200–208.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Geroldinger, M., Verbeeck, J., Thiel, K. E., Molenberghs, G., Bathke, A. C., Laimer, M., & Zimmermann, G. (2023). A neutral comparison of statistical methods for analyzing longitudinally measured ordinal outcomes in rare diseases. *Biometrical Journal*, 2200236. https://doi.org/10.1002/bimj.202200236