Received: 7 March 2023 | Accepted: 26 July 2023

DOI: 10.1111/biom.13920

BIOMETRIC PRACTICE

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

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

Johan Verbeeck[1] | **Martin Geroldinger**[2,3] | **Konstantin Thiel**[2,3] | **Andrew Craig Hooker**[4] | **Sebastian Ueckert**[4] | **Mats Karlsson**[4] | **Arne Cornelius Bathke**[5] | **Johann Wolfgang Bauer**[6] | **Geert Molenberghs**[1,7] | **Georg Zimmermann**[2,3]

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

[2]Team Biostatistics and Big Medical Data, Intelligent Data Analytics (IDA) Lab Salzburg, Paracelsus Medical University, Salzburg, Austria

[3]Research and Innovation Management, Paracelsus Medical University Salzburg, Salzburg, Austria

[4]Department of Pharmacy, Uppsala University, Uppsala, Sweden

[5]Intelligent Data Analytics (IDA) Lab Salzburg, Department of Artificial Intelligence and Human Interfaces, University of Salzburg, Salzburg, Austria

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

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

Correspondence
Johan Verbeeck, Data Science Institute (DSI), Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat), Hasselt University, BE-3500 Hasselt, Belgium.
Email: johan.verbeeck@uhasselt.be

Funding information
WISS 2025 project 'IDA-Lab Salzburg', Grant/Award Numbers: 20102-F1901166-KZP, 20204-WISS/225/197-2019; European Joint Programme on Rare Diseases (EJP RD), EU Horizon 2020, Grant/Award Number: grant agreement no. 825575

Abstract

To optimize the use of data from a small number of subjects in rare disease trials, an at first sight advantageous design is the repeated measures cross-over design. However, it is unclear how these within-treatment period and within-subject clustered data are best analyzed in small-sample trials. In a real-data simulation study based upon a recent epidermolysis bullosa simplex trial using this design, we compare non-parametric marginal models, generalized pairwise comparison models, GEE-type models and parametric model averaging for both repeated binary and count data. The recommendation of which methodology to use in rare disease trials with a repeated measures cross-over design depends on the type of outcome and the number of time points the treatment has an effect on. The non-parametric marginal model testing the treatment–time-interaction effect is suitable for detecting between group differences in the shapes of the longitudinal profiles. For binary outcomes with the treatment effect on a single time point, the parametric model averaging method is recommended, while in the other cases the unmatched generalized pairwise comparison methodology is recommended. Both provide an easily interpretable effect size measure, and do not require exclusion of periods or subjects due to incompleteness.

KEYWORDS
Barnard test, cross-over, epidermolysis bullosa simplex, GEE, generalized pairwise comparison, model averaging, non-parametric marginal model, rare diseases, repeated measures

1 | INTRODUCTION

Epidermolysis bullosa simplex (EBS) is a rare, genetic disease that primarily affects the skin and is characterized by formation of blisters following mechanical stress (Coulombe & Lee, 2012). While current treatments are limited to alleviation and conventional wound care, a growing number of innovative therapeutic compounds are evaluated in clinical trials. One of these trials was a randomized, placebo-controlled, double-blind, two-period cross-over phase 2/3 trial, which assessed the reduction in blisters of an immunomodulatory 1% diacerein cream versus placebo (Wally et al., 2018) (Figure 1). The 16 patients in this trial were randomly assigned to either the placebo or the diacerein treatment, and were treated daily for 4 weeks, followed by a follow-up at 16 weeks. After a washout period, patients were crossed over to the opposite treatment, following an identical treatment schedule. In each treatment period, blisters in the treated body surface area were counted at the start of the treatment period, after 2 and 4 weeks of treatment, and after follow-up. The primary endpoint was the proportion of patients with more than 40% reduction from baseline in the number of blisters after 4 weeks of treatment. This was considered more meaningful from a clinical perspective than the raw blister counts.

Despite the recommendations of the CONSORT to analyze cross-over designs by a paired test (Dwan et al., 2019), the primary endpoint was tested with a one-sided Barnard test (Barnard, 1947), an exact test for two independent binomials. This test, however, requires separate analyses for each treatment period and showed an inconclusive result (Wally et al., 2018). During the first treatment period, 86% of patients receiving diacerein and 14% of the placebo-treated patients achieved a reduction in blister counts of more than 40% ($p = 0.007$). While in the second period 37.5% of all diacerein-treated patients and 17% of all placebo-treated patients achieved a reduction in blister counts of more than 40% ($p = 0.32$).

Unfortunately, the Barnard test ignores the cross-over design of the study, does not account for repeated measures, and requires a dichotomized outcome of the count data. Therefore, it only uses a fraction of the available information in a repeated-measures cross-over trial. Moreover, the choice of the time point for the primary endpoint analysis may influence the results. Indeed, there is an indication that the diacerein treatment may have a therapeutic effect beyond the 4-week treatment period. At the end of follow-up in the first treatment period, all diacerein-treated patients and only 57% of the placebo patients showed a reduction of more than 40% ($p = 0.038$). At the end of follow-up in the second treatment period, 75% of all diacerein-treated patients and 17% of all placebo-treated patients, achieved the 40% reduction ($p = 0.022$). Rather than evaluating the outcome on a single time point, separately per treatment period, an analysis that uses all information in a single analysis would be preferable to evaluate the treatment effect. This would evade difficulties in interpreting conflicting results from separate analyses of each treatment period, and would not require a correction for multiple testing if the analysis is repeated for several time points.

The work presented in this paper is embedded within the EBStatMax demonstration project of the European Joint Programme of Rare Diseases, which has the overarching aim, among others, of improving statistical methodology in rare diseases in general. Our goal is to evaluate how the repeated measurement and cross-over information in a small-sample trial can be used most efficiently to test for a treatment effect. Current methodologies recommended for the analysis of cross-over trials, such as paired parametric or non-parametric tests and meta-analytic approaches, are useful for evaluating the treatment effect of one time point. However, they require

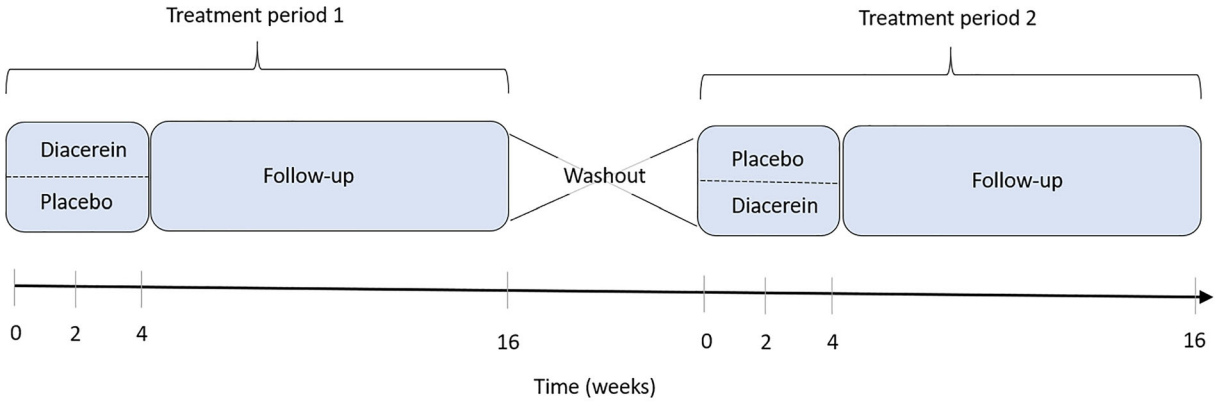

FIGURE 1 Design of the EBS trial. EBS, Epidermolysis bullosa simplex.

a summary measure for the repeated time measurements, which is less sensitive to longitudinal changes and we will therefore not evaluate them. Instead, we will focus on methodologies capable of using the information of the longitudinal profile more effectively, which are at the same time applicable to both count and binary outcomes. Using the information of the longitudinal profile will likely increase the precision of the treatment effect estimation and hence the power of the statistical test. Although most of the methods we will discuss are not new, they have not yet been systematically evaluated with respect to the specifics of a small-sample, repeated-measures cross-over study. Moreover, as missing observations are a common problem in EBS trials, since every transfer to the trial center for a study visit may be accompanied with a high physical burden for the patients, the methodologies are additionally evaluated for their ability to avoid loss of information due to missing data, as exclusion of data due to missing observations might lead to a decrease in power and an increased risk of bias, especially in a small-sample trial.

The evaluated methodologies are briefly described in Section 2 and include, apart from the Barnard test, the non-parametric marginal model (Brunner et al., 2002), the non-parametric general pairwise comparison method (Buyse, 2010; Verbeeck et al., 2019), generalized estimating equations-type models (Arnold & Strauss, 1991; Beunckens et al., 2008; Molenberghs & Verbeke, 2005), and a parametric model averaging technique (Bretz et al., 2005; Chatfield, 1995). More details on the methodologies are available in the Supporting information (Web Appendix A). The type I error control and power of these tests are compared in a real-data simulation study, based on the blister count data of the EBS trial in Section 3 and finally applied to the EBS trial in Section 4. The last section contains some conclusions, recommendations, and reflections regarding limitations and open questions for future research.

2 | METHODOLOGY

2.1 | Barnard test

Barnard's unconditional exact test considers the equality of two binomial proportions p_1 and p_2 (Barnard, 1947) of an observed 2×2 contingency table (Supporting Information, Web Appendix A, Section 1). Barnard's test allows for one- and two-sided hypothesis tests. The treatment effect is expressed as a risk difference of the relative proportions. As it is based on a 2×2 table, the Barnard test requires binary outcomes and ignores the longitudinal and cross-over aspect of the design. Moreover, as it does not allow for incomplete data, subjects without an observation at the analysis time point need to be excluded from analysis. Standard programs for the Barnard test are available in many statistical software environments, including SAS (PROC FREQ-EXACT BARNARD statement) (SAS Help Center, 2020) and R (package Barnard) (Erguler, 2016). Despite not being recommended for repeated measures cross-over trials, the Barnard test is included in the simulation study to demonstrate the limitations of such a choice.

2.2 | Non-parametric marginal model

A non-parametric alternative approach capable of addressing the longitudinal aspect of a repeated measures trial is the rank-based non-parametric marginal model (Brunner et al., 2002) (Supporting information, Web Appendix A, Section 2). It tests the (two-sided) hypothesis of no interaction effect between treatment and time and expresses the treatment effect by time point as the *relative marginal effect*, which can be interpreted as the probability that a random observation in the treatment group and at time point t results in a larger value than an observation randomly chosen from *all* observations in the study.

Although the non-parametric marginal model can take account of the longitudinal aspect of both the blister count and the dichotomized reduction in blister counts of more than 40%, it currently cannot address the cross-over nature. Hence, the analysis needs to be applied for each treatment period separately. Nevertheless, the non-parametric marginal modeling is included in the simulation study as it is a first step toward improving the Barnard analysis for a longitudinal data analysis. Additionally, as the method requires non-missing data at all time points, only subjects with fully observed longitudinal profiles can be included in the analysis. The *R* package *nparLD* (Noguchi et al., 2012) provides access to analyses with non-parametric marginal models.

2.3 | Generalized pairwise comparison

The Generalized pairwise comparison (GPC) method (Buyse, 2010; Finkelstein & Schoenfeld, 1999; Pocock et al., 2012; Verbeeck et al., 2019) is a non-parametric approach that evaluates outcomes between pairs of subjects, one from each treatment arm and assigns a score per pair. Although originally designed to evaluate independent outcomes, the GPC method can be adapted for the analysis of correlated repeated measures and the cross-over design (Supporting information, Web Appendix A, Section 3).

The exact permutation test for the unmatched GPC method (Anderson & Verbeeck, 2023; Verbeeck et al., 2020) evaluates the equality of distributions between groups. The conditional sign test of the matched GPC (Matsouaka, 2022), which accounts for the cross-over design, tests whether a randomly chosen subject is doing equally well on treatment as on placebo. Both a one-sided and a two-sided test can be constructed with the GPC method.

We will evaluate the conditional sign test or the exact permutation test (Anderson & Verbeeck, 2023; Verbeeck et al., 2020) for five GPC variants (matched univariate, unmatched univariate, matched prioritized, unmatched prioritized and unmatched non-prioritized) for both the standardized difference of the blister count outcome and the dichotomized blister outcome. The longitudinal summary measure for the univariate GPCs does not allow for missing observations per treatment period. Therefore, periods with at least one missing observation need to be excluded from the unmatched analyses, while the matched GPC (conditional sign test) additionally requires complete observations for both treatment periods.

As no standard programs are available for GPC analyses, programs in both SAS and R have been developed by the authors.

2.4 | GEE-type models

In a repeated-measures cross-over design, two types of covariance patterns or correlations between measurements within a subject can be distinguished, namely dependencies between measurements in the same treatment period and dependencies between measurements from different treatment periods. In the presence of repeated measures and when interest lies in marginal (a.k.a. population-averaged) effects, it may be difficult to work with full likelihoods, certainly considering non-Gaussian data, which commonly occur in rare disease trials. A possible solution to circumvent the need for full likelihood is presented by pseudo-likelihood estimation (le Cessie & van Houwelingen, 1994) in GEE-type models, avoiding the need to fully model the association structure, while still leading to valid inferences. In small samples, bias-corrected sandwich estimators may be required (Li & Redden, 2015; Long & Ervin, 2000; MacKinnon & White, 1985). Considering $i = 1, 2$ for the treatment assignment, $k = 1, \ldots, N$ for the subjects and $t = 1, \ldots, 4$ for the time points, the following GEE-type model will be evaluated for both the blister count $X_{ikt} \sim \text{Poisson}(\lambda_{ikt})$ and the dichotomized blister outcome $Y_{ikt} \sim \text{Bernoulli}(\pi_{ikt})$:

$$\begin{aligned}\text{logit}(\theta_{ikt}) &= \beta_0 + \beta_1 G_{ik} + \beta_2 P_k \\ &+ \sum_{j=3}^{5} \beta_j T_{ijkt} + \beta_6 G_{ik} P_k \\ &+ \sum_{j=7}^{9} \beta_j G_{ik} T_{ijkt} \\ &+ \sum_{j=10}^{12} \beta_j P_k T_{ijkt}, \end{aligned} \quad (1)$$

where $\theta_{ikt} = \lambda_{ikt}$ or π_{ikt}, and with G_{ik} being a treatment group indicator, P_k a period indicator, and T_{ijkt} a discrete time indicator. We will evaluate several working correlation variance–covariance structures (exchangeable, heterogeneous autocorrelation and unstructured), as well as several corrections for small-sample bias (no correction, Kauermann & Carroll, Fay & Graubard and Mancl & DeRouen) (Supporting information, Web Appendix A, Section 4). These GEE-type models infer, with a one- or two-sided Wald test, whether the linear combination of parameters involving the treatment group indicator equals zero, that is, whether there is an overall treatment effect.

In SAS, the procedure GLIMMIX allows for GEE-type models with several small-sample corrections for the sandwich estimator, as well as several working correlation structures. As the inference is valid under missing completely at random and can be corrected under missing at random (Molenberghs et al., 2011), no information needs to be excluded due to missing observations.

2.5 | Model averaging

Finally, a parametric method, generalized linear mixed models (GLMM), is evaluated, as often parametric models are the more powerful methods in hypothesis testing. However, they require the definition of both a mean and a correct variance structure, the latter being difficult to assess in small samples (Bartlett, 1937; Hurvich & Tsai, 1989). Moreover, parameter estimation in GLMM may become intractable. To circumvent these limitations of GLMM in small samples, model averaging (Aoki et al., 2017) may be a convenient alternative. In model averaging, rather than a single GLMM, the analysis is based on a pool of GLMM models that are weighted according to their fit to the data. The hypothesis test of no treatment effect in the model averaging framework is described in detail in the Supporting information, Web Appendix A, Section 4. We compute the treatment effect in two different ways. First as the placebo-corrected change from baseline after 4 weeks of treatment, averaged across the two

potential sequences ($\Delta\Delta_q$), and second as the average difference between placebo and treatment at the observation time point for each of the two potential sequences (EMC, expected mean change from placebo).

We consider a pool of eight GLMM models with two random effects, allowing for repeated measures and the cross-over design:

$$\mathcal{M}_1: \quad g(E(y_{kt}|\boldsymbol{b}_k)) = \beta_0 + \beta_1 G_{ik} + \beta_2 t_{kt} + \beta_3 P_i + \boldsymbol{z}'_{kt}\boldsymbol{b}_k$$

$$\mathcal{M}_2: \quad g(E(y_{kt}|\boldsymbol{b}_k)) = \beta_0 + \beta_1 G_{ik} + \beta_2 t_{kt} + \beta_3 P_i + \beta_4 G_{ik} t_{kt} + \boldsymbol{z}'_{kt}\boldsymbol{b}_k$$

$$\mathcal{M}_3: \quad g(E(y_{kt}|\boldsymbol{b}_k)) = \beta_0 + \beta_1 G_{ik} + \beta_2 t_{kt} + \beta_3 P_i + \beta_4 G_{ik} t_{kt} + \beta_5 P_i t_{kt} + \boldsymbol{z}'_{kt}\boldsymbol{b}_k$$

$$\mathcal{M}_4: \quad g(E(y_{kt}|\boldsymbol{b}_k)) = \beta_0 + \beta_1 G_{ik} + \beta_2 t_{kt} + \beta_3 P_i + \beta_4 G_{ik} t_{kt} + \beta_5 P_i t_{kt} + \beta_6 P_i G_{ik} + \boldsymbol{z}'_{kt}\boldsymbol{b}_k$$

$$\mathcal{M}_5: \quad g(E(y_{kt}|\boldsymbol{b}_k)) = \beta_0 + \beta_1 G_{ik} + \sum_t \beta_t T_{ikt} + \beta_3 P_i + \boldsymbol{z}'_{kt}\boldsymbol{b}_k$$

$$\mathcal{M}_6: \quad g(E(y_{kt}|\boldsymbol{b}_k)) = \beta_0 + \beta_1 G_{ik} + \sum_t \beta_t T_{ikt} + \beta_3 P_i + \sum_{t_i} \beta_{t_i} T_{ikt} G_{ik} + \boldsymbol{z}'_{kt}\boldsymbol{b}_k$$

$$\mathcal{M}_7: \quad g(E(y_{kt}|\boldsymbol{b}_k)) = \beta_0 + \beta_1 G_{ik} + \sum_t \beta_t T_{ikt} + \beta_3 P_i + \sum_{t_G} \beta_{t_G} T_{ikt} G_{ik} + \sum_{t_P} \beta_{t_P} T_{ikt} P_i + \boldsymbol{z}'_{kt}\boldsymbol{b}_k$$

$$\mathcal{M}_8: \quad g(E(y_{kt}|\boldsymbol{b}_k)) = \beta_0 + \beta_1 G_{ik} + \sum_t \beta_t T_{ikt} + \beta_3 P_i + \sum_{t_G} \beta_{t_G} T_{ikt} G_{ik} + \sum_{t_P} \beta_{t_P} T_{ikt} P_i + \beta_6 P_i G_{ik} + \boldsymbol{z}'_{kt}\boldsymbol{b}_k$$

Model $\mathcal{M}_1$ includes a baseline effect β_0 and effect differences (β_1, β_2, and β_3) based on treatment (G_{ik}), continuous time (t_{kt}), and period (P_i). Models $\mathcal{M}_2$ through $\mathcal{M}_4$ add interaction terms between these descriptors. Models $\mathcal{M}_5$ through $\mathcal{M}_8$ investigate the use of observation number (T_{ikt} where $t = 1, \ldots, 4$) instead of continuous time in the models. The corresponding link function, g, was Poisson for count data and binomial for binary data. For each model above, two random-effects models were used, for a total of 16 models in the predefined model set.

$$\mathcal{RE}_1: \quad \boldsymbol{z}'_{kt}\boldsymbol{b}_k = b_{0,k}$$

$$\mathcal{RE}_2: \quad \boldsymbol{z}'_{kt}\boldsymbol{b}_k = b_{0,k}P_i + b_{1,k}(1 - P_i)$$

As there are no standard programs available for the model averaging, user-defined R code was developed. Additionally, as GLMM implicitly handles missing data, no data need to be excluded in the analysis.

3 | SIMULATIONS

For each of the methods described in Section 2, the type I error and power are evaluated in a real-data simulation study resembling the EBS trial. The blister count outcome of the EBS trial is grouped into blocks per subject and period. Thus, subjects have a maximum of two blocks in this cross-over trial, with four time points in each block. These blocks are subsequently permuted across all subjects and both periods, while holding the observations within the block unchanged. This preserves the within-subject correlation per treatment period, while eliminating any treatment period effect. A treatment time effect is simulated under two scenarios, one with a treatment effect at a single visit and another with a treatment effect at multiple visits, to account for a gradual onset of effect, and then decrease in effect as treatment is stopped. In a first scenario, for each subject in the placebo arm, a random count from a Poisson($\lambda = 3$) distribution was added to the 4-week time point. The $\lambda = 3$ is based on the observed difference in blister count between placebo and treatment in the EBS trial at week 4 (2.5) and an expected 30%–50% difference in blister count at week 4 by the clinicians. In a sensitivity analysis, other Poisson distributions led to similar relative positions in the comparison of the methodologies. In the second scenario, half of the random count at week 4 was added to the week 2 visit (rounded to the nearest integer) and for the month 3 visits the treatment effect at week 4 was varied by adding a draw from a random standard normal (rounded to the nearest integer). Analogously to the primary analysis of the original study (Wally et al., 2018), a binary indicator was obtained by evaluating whether the blister count at weeks 2 and 4, and at the follow-up visit was reduced by more than 40% compared to the baseline count. The type I error and power of each method is evaluated in 5,000 permuted samples, using $\alpha = 0.05$ as the one- or two-sided level, as appropriate. For the parametric model averaging approach 200 permuted samples were used because of runtime considerations. Uncertainty in the resulting power and type 1 error calculations are reported using binomial confidence intervals (Wilson (1927) method).

The data in the original analysis comprise 7 placebo and 7 Diacerein periods in the first treatment period and 6 placebo and 8 Diacerein periods in the second treatment period. The matched GPC method requires full observations in both treatment periods and includes only 12 subjects, since 4 of the 16 subjects were either treated with only one treatment or had missing blister counts in a treatment period.

The simulations with the dichotomized blister outcome show that the Barnard test is a rather conservative test (Table 1). In spite of this, the one-sided tests show a 24%–35% power for both treatment periods, which is, as expected, higher than the power for the two-sided Barnard test (9%–15%). Recall that the Barnard test only evaluates binary outcomes at week 4. Consequently, the count outcome cannot be evaluated at all and the scenario with a single treatment effect at week 4 and the scenario with additional visits with treatment effects will lead to one and the same result.

The non-parametric marginal model is a rather liberal test for the binary outcome (Table 1), while it controls the type I error well with the count outcome (Table 2). The power reaches 14%–16% for the binary outcome and 24%–25% for the count outcome with a treatment effect on a single time point, and is lower when there are multiple time points with a treatment effect (Tables 1 and 2).

Both the one- and two-sided GPC tests control the Type I error well in the small-sample simulations, except for the one-sided matched GPC with the count outcome (Tables 1 and 2).

In contrast to the marginal model, the power of GPC tests increases with more time points having a treatment effect, in some cases reaching up to 70%. As expected, dichotomizing the blister counts, lead to a loss of granularity in the data and thus a lower power compared to the count outcome (Tables 1 and 2). Recall that the matched GPC is using less data than the other methods, which may explain the lower power compared to the unmatched GPC tests.

When considering an exchangeable or heterogeneous autocorrelation variance structure, the small-sample corrections in the GEE-type model lead to an over-conservative type I error for the dichotomized count, while the Mancl & DeRouen correction controls the type I error well for the count outcome (Appendix Table A.1). This is not surprising, given that the use of the small-sample corrections lies primarily with heteroscedasticity and the variance for the dichotomized count is expected to be closer to homoscedasticity compared to the count outcome. While the unstructured working correlation structure is the most flexible, it consists of too many elements to be estimated efficiently in the count outcome. In our simulated samples, the heterogeneous autocorrelation working correlation structure is deemed most likely and therefore considered for further details. In the simulations, the Mancl & DeRouen correction for the count outcome and no correction for the dichotomized count leads to a power close to the power in the matched GPC tests, but lower compared to the unmatched GPC tests (Tables 1 and 2).

Finally, the parametric model averaging is a very liberal test using the EMC as a treatment effect measure. In contrast, the $\Delta\Delta$ controls the type I error better (Tables 1 and 2). The power of the $\Delta\Delta$ is similar to GPC for the binary outcome in the scenario with treatment effects on multiple time points, while it is higher with a treatment effect in a single time point. For the count outcome, it is similar to or lower than GPC (Tables 1 and 2).

In conclusion, the (standardized difference of the) blister count, rather than the dichotomized count, reaches higher power for most methods. GPC results most often in the highest power, especially the prioritized GPC, except for the binary outcome in the scenario with a treatment effect on a single time point. In the latter case, the model averaging with the $\Delta\Delta$ is the most powerful test.

4 | APPLICATION TO EPIDERMOLYSIS BULLOSA SIMPLEX TRIAL

The EBS blister count outcome and the dichotomized count outcome were re-analyzed with each of the methods discussed in Section 2, where each method takes the maximum amount of data it can use. The repeated measures of the binary outcome in the EBS trial suggest a treatment difference between Diacerein and placebo, mainly in the first period (Figure 2).

The repeated measures of the blister counts suggest a treatment effect on all visits during the first treatment period, while the second treatment period indicates a late treatment effect (Figure 3).

Recall that the one-sided Barnard test showed an inconclusive result when analyzing the effect on week 4, with evidence for a treatment effect in period 1 ($p = 0.007$), but no evidence in period 2 ($p = 0.32$) (Wally et al., 2018).

When using the complete blocks of repeated measures per treatment period in the non-parametric marginal models, there is, surprisingly, evidence for a treatment effect in the second treatment period for both the blister counts ($p = 0.26$ and 0.01, respectively, for periods 1 and 2) and the dichotomized blister counts ($p = 0.52$ and 0.02, respectively, for periods 1 and 2) (Table 3). The disadvantage of the non-parametric marginal model is that an overall treatment effect measure is not available.

Additionally including the cross-over design in the analysis with the matched univariate and prioritized GPC, 8

TABLE 1 Type I error (95% CI) and power (95% CI) of the scenario with only a treatment effect on week 4 and the scenario with treatment effect on all visits in 5,000 permuted samples of the dichotomized outcome, with the number of permuted samples that converged or could be used in both the type I error and power evaluation.

	One-sided				Two-sided			
	Samples	Type I	Power W4 effect	Power all W effects	Samples	Type I	Power W4 effect	Power all W effects
Barnard period 1	5000/4968	0.036	0.348	NA	5000/4968	0.023	0.153	NA
		(0.031;0.042)	(0.333; 0.359)			(0.019;0.028)	(0.142;0.162)	
Barnard period 2	5000/4978	0.033	0.243	NA	5000/4978	0.044	0.085	NA
		(0.031;0.034)	(0.246;0.255)			(0.039;0.043)	(0.078;0.083)	
Marginal model period 1	NA	NA	NA	NA	4882/4966/4923	0.069	0.158	0.099
						(0.067;0.072)	(0.147;0.154)	(0.091;0.108)
Marginal model period 2	NA	NA	NA	NA	4999/4994/4992	0.066	0.141	0.098
						(0.064;0.069)	(0.137;0.144)	(0.090;0.107)
Matched univariate GPC	5000/5000/5000	0.056	0.136	0.522	5000/5000/5000	0.046	0.065	0.319
		(0.050;0.063)	(0.127;0.146)	(0.508;0.536)		(0.041;0.052)	(0.058;0.072)	(0.306;0.332)
Unmatched univariate GPC	5000/5000/5000	0.055	0.178	0.724	5000/5000/5000	0.055	0.108	0.585
		(0.053;0.058)	(0.176;0.184)	(0.711;0.736)		(0.053;0.058)	(0.100;0.117)	(0.571;0.599)
Matched prioritized GPC	5000/5000/5000	0.058	0.125	0.515	5000/5000/5000	0.044	0.061	0.310
		(0.052;0.065)	(0.116;0.134)	(0.501;0.529)		(0.039;0.050)	(0.055;0.068)	(0.297;0.323)
Unmatched prioritized GPC	5000/5000/5000	0.052	0.171	0.725	5000/5000/5000	0.050	0.100	0.589
		(0.050;0.055)	(0.167;0.174)	(0.712;0.737)		(0.048;0.053)	(0.097;0.103)	(0.575;0.603)
Unmatched non-prioritized GPC	5000/5000/5000	0.055	0.204	0.681	5000/5000/5000	0.063	0.121	0.541
		(0.053;0.058)	(0.196;0.204)	(0.668;0.694)		(0.063;0.066)	(0.117;0.124)	(0.527;0.555)
GEE-no correction	4917/3435/2928	0.059	0.147	0.514	4917/3435:2928	0.055	0.070	0.343
		(0.053;0.066)	(0.137;0.157)	(0.500;0.528)		(0.049;0.062)	(0.063;0.077)	(0.330;0.356)
GEE-Kauermann & Carroll	4917/3435/2928	0.045	0.116	0.449	4917/3435/2928	0.038	0.049	0.262
		(0.040;0.051)	(0.107;0.125)	(0.435;0.463)		(0.033;0.044)	(0.034;0.055)	(0.250;0.274)
GEE-Fay & Graubard	4917/3435/2923	0.045	0.113	0.449	4917/3435/2923	0.039	0.047	0.264
		(0.040;0.051)	(0.105;0.122)	(0.435;0.463)		(0.034;0.045)	(0.041;0.053)	(0.252;0.276)
GEE-Mancl & DeRouen	4917/3435/2928	0.033	0.080	0.374	4917/3435/2928	0.027	0.028	0.187
		(0.028;0.038)	(0.073;0.088)	(0.361;0.388)		(0.023;0.032)	(0.024;0.033)	(0.176;0.198)
Model averaging ΔΔ	200/200/106	0.070	0.325	0.642	200/200/106	0.070	0.205	0.538
		(0.042;0.114)	(0.264;0.393)	(0.547;0.726)		(0.042;0.114)	(0.155;0.266)	(0.443;0.630)
Model averaging-EMC	200/200/106	0.090	0.30	0.755	200/200/106	0.125	0.210	0.642
		(0.058;0.138)	(0.241;0.367)	(0.665;0.827)		(0.086;0.178)	(0.159;0.272)	(0.547;0.726)

Abbreviation: NA, not applicable. EMC, expected mean change from placebo; GPC, Generalized pairwise comparison; GEE, Generalized estimating equations and EMC, Expected mean change from placebo.

TABLE 2 Type I error (95% CI) and power (95% CI) of the scenario with only a treatment effect on week 4 and the scenario with treatment effect on all visits in 5,000 permuted samples of the count outcome, with the number of permuted samples that converged or could be used in both the type I error and power evaluation.

	One-sided				Two-sided			
	Samples	Type I	Power W4 effect	Power all W effects	Power Samples	Power Type I	W4 effect	all W effects
Marginal model period 1	NA	NA	NA	NA	5000/5000/5000	0.050	0.252	0.197
						(0.044;0.056)	(0.240;0.264)	(0.186;0.208)
Marginal model period 2	NA	NA	NA	NA	5000/5000/5000	0.051	0.237	0.202
						(0.045;0.057)	(0.225;0.249)	(0.191;0.213)
Matched univariate GPC	5000/5000/5000	0.066	0.258	0.596	5000/5000/5000	0.043	0.114	0.351
		(0.059;0.073)	(0.246; 0.270)	(0.582; 0.610)		(0.038;0.049)	(0.106;0.124)	(0.338; 0.364)
Unmatched univariate GPC	5000/5000/5000	0.051	0.284	0.721	5000/5000/5000	0.048	0.180	0.583
		(0.045;0.057)	(0.272; 0.297)	(0.708; 0.733)		(0.042;0.054)	(0.170;0.191)	(0.569; 0.597)
Matched prioritized GPC	5000/5000/5000	0.070	0.560	0.586	5000/5000/5000	0.039	0.304	0.324
		(0.063;0.077)	(0.546; 0.574)	(0.572; 0.600)		(0.034;0.045)	(0.291;0.317)	(0.311; 0.337)
Unmatched prioritized GPC	5000/5000/5000	0.048	0.673	0.692	5000/5000/5000	0.046	0.525	0.553
		(0.042;0.054)	(0.660; 0.686)	(0.679; 0.705)		(0.041;0.052)	(0.511;0.539)	(0.539; 0.567)
Unmatched non-prioritized GPC	5000/5000/5000	0.051	0.228	0.830	5000/5000/5000	0.046	0.138	0.708
		(0.045;0.057)	(0.217; 0.240)	(0.819; 0.840)		(0.041;0.052)	(0.129;0.148)	(0.695; 0.720)
GEE-no correction	4616/4628/4683	0.077	0.166	0.507	4616/4628/4683	0.094	0.125	0.390
		(0.070;0.085)	(0.156; 0.177)	(0.493; 0.521)		(0.086;0.102)	(0.116;0.134)	(0.377; 0.404)
GEE-Kauermann & Carroll	4616/4628/4683	0.064	0.141	0.465	4616/4628/4683	0.079	0.099	0.343
		(0.058;0.071)	(0.131; 0.151)	(0.451; 0.479)		(0.072;0.087)	(0.091;0.108)	(0.330; 0.356)
GEE-Fay & Graubard	4620/4628/4679	0.067	0.148	0.477	4620/4628/4679	0.082	0.102	0.355
		(0.060;0.074)	(0.138; 0.158)	(0.463; 0.491)		(0.075;0.090)	(0.094;0.111)	(0.342; 0.368)
GEE-Mancl & DeRouen	4616/4628/4683	0.050	0.119	0.414	4616/4628/4683	0.058	0.077	0.284
		(0.044;0.056)	(0.110; 0.128)	(0.400; 0.428)		(0.052;0.065)	(0.070;0.085)	(0.272; 0.297)
Model averaging ΔΔ	200/200/200	0.045	0.490	0.560	200/200/200	0.055	0.375	0.470
		(0.024;0.083)	(0.422;0.559)	(0.491;0.627)		(0.031;0.096)	(0.311;0.444)	(0.402;0.539)
Model averaging EMC	200/200/200	0.120	0.23	0.570	200/200/200	0.140	0.175	0.505
		(0.082;0.172)	(0.177;0.293)	(0.501;0.637)		(0.099;0.195)	(0.129;0.234)	(0.436;0.574)

Abbreviation: NA, not applicable. EMC, expected mean change from placebo; GPC, Generalized pairwise comparison; GEE, Generalized estimating equations and EMC, Expected mean change from placebo.

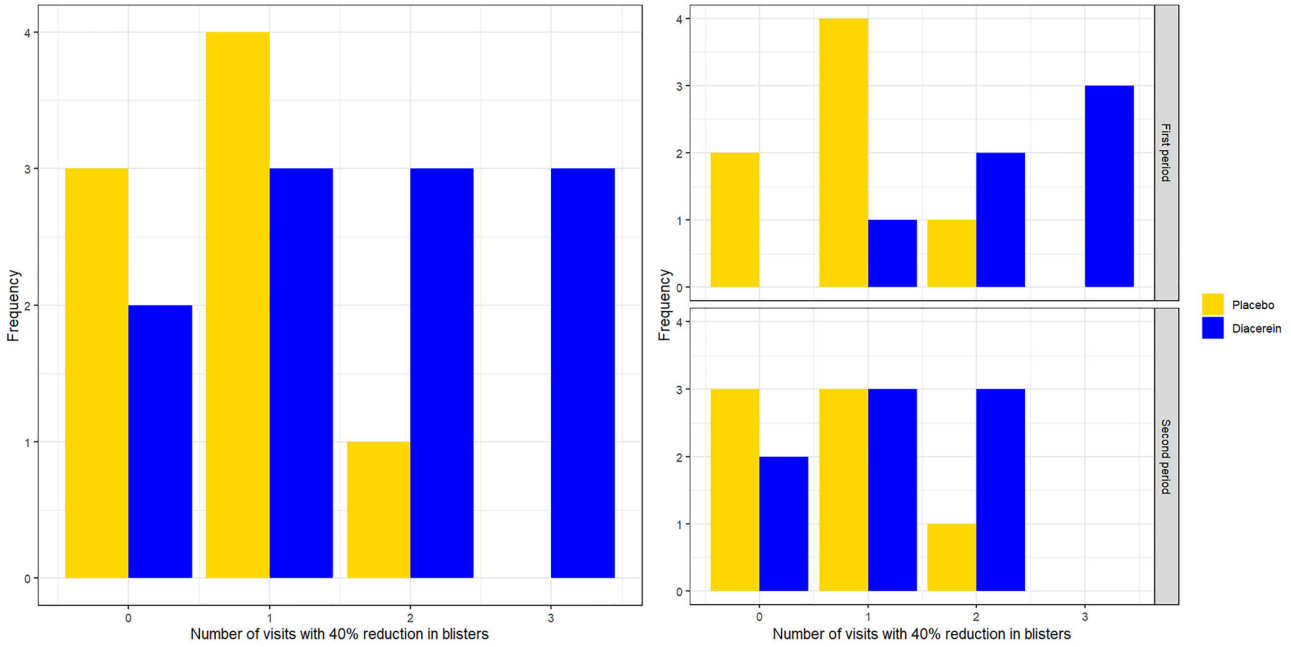

FIGURE 2 Frequency of number of visits with 40% reduction in blisters overall (left) and by both periods (right). This figure appears in color in the electronic version of this paper, and any mention of color refers to that version.

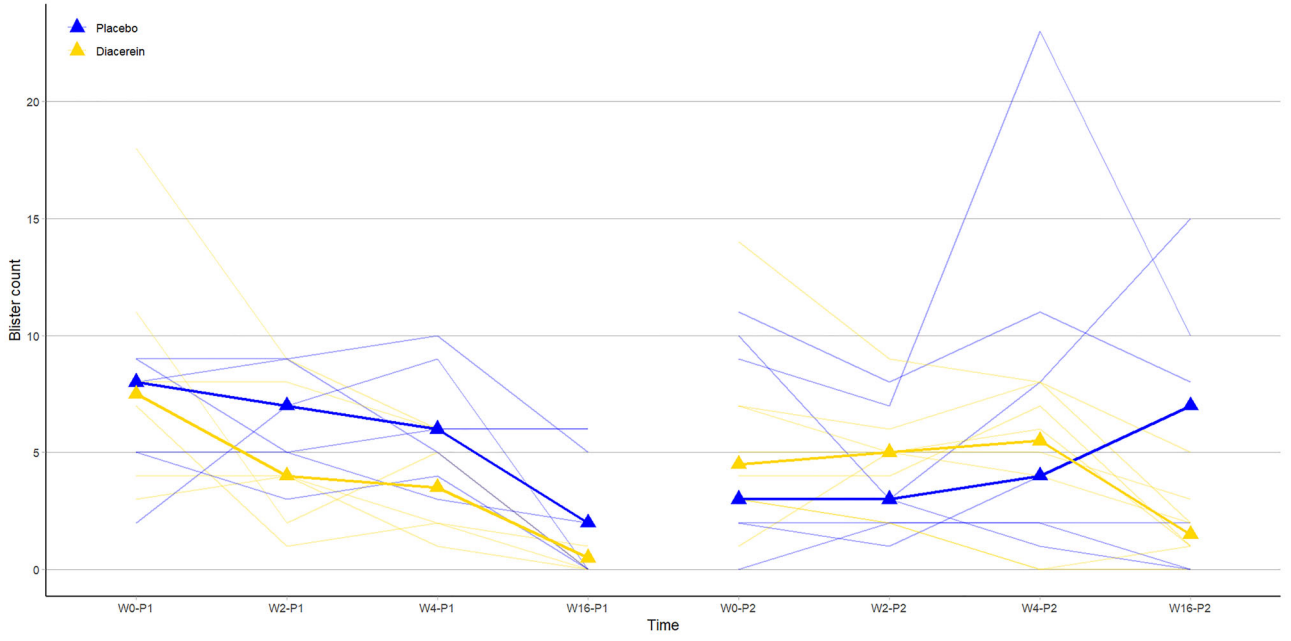

FIGURE 3 Longitudinal blister counts, with the median per time point (triangles and bold lines), per treatment and per treatment period (P1, P2). W = week. This figure appears in color in the electronic version of this paper, and any mention of color refers to that version.

out of 13 subjects have more visits with a 40% reduction in blister counts on Diacerein treatment than on placebo, while only 2 out of 13 have fewer visits with this reduction. Although a two-sided test is not significant ($p = 0.058$), the one-sided test does show evidence for a treatment effect of Diacerein ($p = 0.029$) (Table 3). The probability that a random subject will do better on Diacerein than on placebo minus the reverse is 0.46 with the Agresti—Coull 95% CI: −0.01 to 0.76 (Matsouaka, 2022). In contrast, both the matched univariate and prioritized GPC do not show a treatment effect when looking at the standardized blister count ($p = 1$ for the two-sided test and $p = 0.5$ for the

TABLE 3 Re-analysis of the EBS trial.

	Count outcome			Dichotomized count outcome		
	Treatment effect (95% CI)	One-sided *p*-value	Two-sided *p*-value	Treatment effect (95% CI)	One-sided *p*-value	Two-sided *p*-value
Marginal model period 1	/	/	0.260	/	/	0.520
Marginal model period 2	/	/	0.010	/	/	0.020
Matched univariate GPC	0 (−0.49: 0.49)	0.500	1.000	0.46 (−0.01; 0.76)	0.029	0.058
Unmatched univariate GPC	0.49 (0.06; 0.93)	0.013	0.027	0.55 (0.14; 0.96)	0.004	0.008
Matched prioritized GPC	0 (−0.49; 0.49)	0.500	1.000	0.46 (−0.01; 0.76)	0.029	0.058
Unmatched prioritized GPC	0.44 (0.00; 0.87)	0.025	0.050	0.61 (0.19; 1.00)	0.002	0.004
Unmatched non-prioritized GPC	0.31 (−0.02; 0.65)	0.031	0.062	0.34 (0.09; 0.59)	0.004	0.007
GEE-type model	1.75 (1.08; 2.82)	0.015	0.030	7.37 (1.65; 32.96)	0.006	0.011
Model averaging ΔΔ	−1.72 (−3.43; −0.30)	0.012	0.024	0.33 (0.12; 0.56)	0.006	0.011
Model averaging EMC	−1.61 (−3.77; −0.06)	0.023	0.046	0.34 (0.16; 0.50)	<0.005	<0.005

Abbreviations: EBS, Epidermolysis bullosa simplex; EMC, expected mean change from placebo; GPC, Generalized pairwise comparison; GEE, Generalized estimating equations.

one-sided test, with 6 out of 12 subjects having less blister counts on the Diacerein treatment than on placebo and 6 subjects with more blisters on Diacerein) (Table 3).

Ignoring the cross-over feature in the unmatched GPC results in evidence for improvement of Diacerein compared to placebo in both the one- and two-sided test of the univariate, prioritized and non-prioritized tests on the dichotomized blister count (p-value between 0.002 and 0.008) and in the one-sided test with the standardized blister count ($p = 0.013$, 0.025, and 0.031, respectively), but not always in the two-sided test ($p = 0.027$, 0.050, and 0.062, respectively) (Table 3). Out of the 210 possible pairs, the majority have more visits with a 40% reduction in blister counts on Diacerein treatment than on placebo, while only 24 have less visits with this reduction. The probability that a random subject will do better on Diacerein than a random subject on placebo is estimated between 0.34 (95% CI: 0.09–0.59) by the non-prioritized GPC and 0.61 by the prioritized GPC (95% CI: 0.19–1.00) on the dichotomized blister count and 0.31 (95% CI: −0.02 to 0.65) by the non-prioritized GPC and 0.44 by the prioritized GPC (95% CI: 0.00–0.87) on the standardized blister count (Table 3).

Explicitly modeling the dependence of the repeated dichotomized blister counts within a treatment period and within a subject in the GEE-type model again shows evidence for a treatment effect with all working correlation structures and all small-sample corrections (two-sided p-value between 0.011 and 0.036). Besides a clear treatment effect, there is also a period and a time effect in all models. As the heterogeneous autocorrelation working correlation structure is most plausible and simulations show that no small-sample correction controls the Type I error best, we explore the result of this model further, although they are fairly consistent over all models. The odds ratio of a 40% reduction in the number of blisters between Diacerein and placebo is 7.37 (95% CI: 1.65–32.96; $p = 0.011$), which is mainly due to the effect in the first period (Table 3). Indeed, the odds ratio of a 40% reduction in the number of blisters in period 1 versus period 2 is 6.71 (95% CI: 1.55–28.99). Using the repeated blister counts, only the GEE-type model with an independence variance–covariance structure converges and shows evidence for a treatment effect (two-sided p-value 0.030) (Table 3). Although the independence assumption is obviously incorrect, since measurements are clustered within subjects, the inference with an incorrect working correlation matrix should be valid.

The model averaging showed that the population average placebo-corrected probability of achieving 40% reduction in blister count from baseline after 4 weeks of treatment (ΔΔ) was 0.33 (95% CI: [0.12, 0.56], $p = 0.011$). The EMC for achieving 40% reduction in blister count from baseline was estimated to be 0.34 (95% CI: [0.16, 0.50], $p < 0.005$) (Table 3). The two effect measures for the repeated blister count found a significant drug effect with a population average placebo-corrected change from baseline after 4 weeks of treatment (ΔΔ) to be −1.72 blisters (95% CI: [−3.43, −0.30], $p = 0.024$) and the EMC to be −1.61 blisters (95% CI: [−3.77, −0.064], $p = 0.046$) (Table 3).

5 | DISCUSSION

We have compared non-parametric, semi-parametric, and parametric statistical methods for repeated measures, 2-period cross-over small-sample trials with a binary and count outcome (Table 4). Although the non-parametric marginal model allows for the longitudinal repeated measures within a treatment period, it still ignores the between period correlation of the cross-over aspect. The latter

TABLE 4 Advantages and disadvantages of evaluated statistical methods for repeated measures, cross-over trials in small samples.

	Repeated measures	Cross-over	Missing data	Effect measure	One- and two-sided	Comments
Barnard	No	No	No	Risk difference[†]	Yes	
Marginal model	Yes	No	No	Relative effect*	No	
GPC	Yes	Yes	No	Net treatment benefit	Yes	Extension to multivariate outcomes
GEE-type	Yes	Yes	Yes	Odds/risk ratio	Yes	Can test carry-over effect
Model averaging	Yes	Yes	Yes	ΔΔ	Yes	Time consuming for simulations

Notes: [†]There is no overall treatment effect available for the Barnard test, only a per treatment period effect.
*There is no overall treatment effect available for the marginal models, only a per time point treatment effect. Abbreviation: GPC, Generalized pairwise comparison; GEE, Generalized estimating equations.

can be incorporated into the non-parametric GPC methods, GEE-type models, and model averaging. While the unmatched prioritized and non-prioritized GPC incorporates the within-subject correlation within a period, the matched GPC additionally incorporates the cross-over design. Ignoring the cross-over design in the unmatched GPC, however, leads to asymptotically valid results (Konietschke & Pauly, 2012) and controls the type I error better in sample sizes < 15 subjects. GPC has an obvious interpretable treatment effect measure, results in a single analysis rather than a per treatment period analysis and can be easily extended to multivariate outcomes, for example, combining pain, pruritus, and/or quality of life to the blister count outcome (Geroldinger et al., 2023).

Modeling the within-period and within-subject dependency of the repeated-measures cross-over trial in a GEE-type model does not show a clear advantage over the non-parametric methodologies. The power is similar to the matched GPC and lower for the unmatched GPC and may lead to convergence issues. On the other hand, it provides, besides a single analysis, also information concerning a possible period effect. Treatment or period effects can be expressed in an easily interpretable effect measure, the odds or rate ratio. Moreover, standard software programs are available that allow for several working correlation structures and small-sample corrections and no data need to be excluded due to missingness, as inference is valid under missing completely at random and can be corrected for missing at random (Molenberghs et al., 2011). Our simulations confirm that in the presence of homoscedasticity small-sample corrections are not required, while they are useful under heteroscedasticity (Long & Ervin, 2000; MacKinnon & White, 1985).

Averaging a pool of parametric GLMM models and expressing the treatment effect with the ΔΔ shows an increased power compared to all other tests when there is a single time point with a treatment effect on a binary outcome. In all other cases, it did not show an advantage in power. Although the model averaging method is computer intensive, the analysis of a single trial takes between 10 and 30 min, which is not prohibitive. On the other hand, a simulation study with 5,000 permuted samples requires a large computer cluster, while the most time-consuming non-parametric or semi-parametric method lasted only 1 h for 5,000 permuted samples. The model-averaging method shares with the GEE-type models that both within-period and within-subject dependencies can be modeled, and no data need to be excluded due to missingness, since the inference is valid under MCAR and MAR. The benefit of the GLMM is that they can be used to simulate future studies, making them a valuable tool for study design.

One may argue that the simulation setting for comparing the methodologies was favoring the prioritized GPC, as the largest treatment effect was added to the post-treatment visit, which was evaluated first in the prioritized analyses. However, the simulation setting and prioritization of the time points in the GPC analyses were based on clinical reasoning and not by intention to favor any method. A Diacerein treatment in EBS patients is clinically expected to have a delayed onset of treatment effect, followed by a gradually decline of effect after stopping the treatment. Similarly, it is clinically plausible to prioritize the visits by the same hierarchy in GPC.

In the comparison between the different methodologies, it should be stressed that the underlying null hypotheses being tested are not equivalent. This needs to be taken into account when the empirical power is compared between the methods. The null hypotheses of the non-parametric methods are for example more restricted compared to the GEE-type models and model averaging. Perhaps the choice of only testing the interaction effect in the non-parametric marginal models may not be optimal for our simulation setting. Potentially, in other simulation settings the non-parametric models might be recommended. The choice of methodology depends on the null hypothesis of interest, the treatment effect measure, and the respective operational characteristics (Table 4).

Most of the re-analyses of the EBS trial, taking account of the repeated-measures cross-over design, show a beneficial effect of the Diacerein cream over placebo on the blister count reduction. Additionally, the GEE-type model confirms that there is a carry-over or period effect. Indeed,

during a pilot study, investigators noted that after application of the Diacerein cream for 6 weeks, blisters did not recur during a 6-week placebo-controlled withdrawal (Wally et al., 2018). The effect of Diacerein is potentially longer than anticipated, hence the wash-out period in a future cross-over EBS trial should ideally be increased. The wash-out period in the EBS trial was 5.6 months (standard deviation, 1.7) (Wally et al., 2018).

Most methodologies are more sensitive to detect a treatment effect using the count outcome compared to the dichotomization of blister count. While it is tempting to prefer the blister count over its dichotomization, one should be careful not to ignore the uncertainty around the blister count. Indeed, counting blisters on a predefined area of the skin is not automated and hence variable between clinical assessors. Dichotomizing the blister count by achievement of 40% reduction in blisters is more robust against the assessment uncertainty and, importantly, it is considered clinically meaningful. It should be noted that the GPC method would allow for considering an alternative to these two extremes, namely to define a threshold (Buyse, 2010) in comparing pairs of subjects (e.g., a "win" is defined as a difference in blister counts of at least 3). It is expected that adding a threshold will result in an achieved power in between the two extreme scenarios evaluated in this manuscript, with on one end the blister counts and on the other end the dichotomized count.

The power to detect a treatment effect in all methods, but the non-parametric marginal model, increases when the treatment exhibits an effect on more time points. Potentially, the seemingly paradoxical behavior of the marginal model is a consequence of the fact that a time–treatment-interaction hypothesis is being tested. Consequently, the corresponding ANOVA-type test is in particular sensitive to abrupt changes in a longitudinal profile. In this sense, a profile with a single time point with a treatment effect manifests a more extreme change than a profile with multiple time points with a treatment effect. The same phenomena may play a role in the re-analysis of the EBS trial, where the non-parametric marginal model is the only methodology that showed a treatment effect in period 2 and not in period 1. Further research by separately analyzing both treatment periods of the EBS trial with the unmatched GPC, GEE-type model, and model averaging may explain this in more detail.

There is not a single methodology that was uniformly the best in our simulation. Which method is most powerful depends on the type of outcome and the number of time points with a treatment effect. For a single time point involving a treatment effect with respect to a binary outcome, the $\Delta\Delta$ model averaging was more powerful, while in most other settings, the prioritized or non-prioritized unmatched GPC was most powerful to detect a treatment effect in a repeated-measures and cross-over 2-arm trial in rare diseases or small-sample study. Although one would expect a parametric method to be superior in power compared to a non-parametric method, this was not always the case in our simulations. The ability to detect a treatment effect in the parametric methods however depends on the correct model specification. Potentially, the model averaging models can be improved in our simulation settings with models that better fit the data. The non-parametric marginal model testing the treatment–time-interaction effect is suitable for detecting between group differences in the shapes of the longitudinal profiles.

As anticipated, the matched GPC does not always control the type I error, because the size of the sample is just below the limit of what is required to maintain the nominal confidence level. While there is a large difference in power between a single and multiple time points with a treatment effect on a binary outcome for both the prioritized and non-prioritized GPC, this difference is not present for the prioritized GPC with a count outcome. This is a logical consequence of the GPC algorithm. Indeed, for a count outcome in much more pairs a "win" can be assigned on the first time point compared to a binary outcome. Hence, in a prioritized GPC with a count outcome, not many pairs will be evaluated on the subsequent time points and thus these time points will not add much to the power. In contrast, in a prioritized GPC with a binary outcome, a lot of ties are expected at the first time point and more information will be used from the subsequent time points, which will result in an increased power, if there is also a treatment effect on these subsequent time points. In the non-prioritized GPC with a count outcome, when there is only a single time point with a treatment effect, this effect is 'diluted' by the time points with no treatment effect.

It is worth noting that with a single outcome and no missing data, the unmatched GPC is a linear transformation of the Mann–Whitney test (Mann & Whitney, 1947; Verbeeck et al., 2021) and can thus be equally constructed with ranks. However, the pairwise comparison notation has the advantage that it is easier to extend to the matched GPC and to multivariate outcomes with missing or censored data.

A limitation of our study is that we have not applied any model selection to the GEE-type models, which is cumbersome to automate in a simulation study. By applying model selection, it is possible that efficiency is gained in detecting a treatment effect in the GEE-type models. While the model averaging idea avoids the need to select any model, it cannot be applied as such to the GEE-type models, as model averaging is based on maximum likelihood.

Finally, while current versions of the non-parametric marginal models require fully observed longitudinal profiles, recently there have been some efforts to extend the models allowing for missing data (Rubarth et al., 2022). Further developments should also include the split-plot

design, so that potentially the non-parametric marginal model might also obtain higher power values within the setting discussed in this paper.

ACKNOWLEDGMENTS

We are grateful for the ability to use the EBS trial data. Additionally, we would like to thank Geert Verbeke for technical insights into SAS procedures.

The present work has been performed within the framework of the "EBStatMax Demonstration Project" funded by the European Joint Programme on Rare Diseases (EJP RD), EU Horizon 2020 grant agreement no. 825575. GZ gratefully acknowledges the support of the WISS 2025 project 'IDA-Lab Salzburg' (20204-WISS/225/197-2019 and 20102-F1901166-KZP).

DATA AVAILABILITY STATEMENT

The data that support the findings in this paper are available on request from the corresponding Johan Verbeeck (johan.verbeeck@uhasselt.be).

ORCID

Johan Verbeeck https://orcid.org/0000-0002-4923-1032
Andrew Craig Hooker https://orcid.org/0000-0002-2676-5912

REFERENCES

Anderson, W. & Verbeeck, J. (2023) Exact permutation and bootstrap distribution of generalized pairwise comparisons statistics. *Mathematics*, 11, 1502.

Aoki, Y., Röshammar, D., Hamrén, B. & Hooker, A.C. (2017) Model selection and averaging of nonlinear mixed-effect models for robust phase III dose selection. *Journal of Pharmacokinetics and Pharmacodynamics*, 44, 581–597.

Arnold, B. & Strauss, D. (1991) Pseudolikelihood estimation: some examples. *Sankhya: the Indian Journal of Statistics-Series B*, 53, 233–243.

Barnard, G. (1947) Significance tests for 2 2 tables. *Biometrika*, 34, 123–138.

Bartlett, M. (1937) Properties of sufficiency and statistical tests. *Proceedings of the Royal Society A*, 160, 268–282.

Beunckens, C., Sotto, C. & Molenberghs, G. (2008) A simulation study comparing weighted estimating equations with multiple imputation based estimating equations for longitudinal binary data. *Computational Statistics & Data Analysis*, 52, 1533–1548.

Bretz, F., Pinheiro, J.C. & Branson, M. (2005) Combining multiple comparisons and modeling techniques in dose–response studies. *Biometrics*, 61, 738–748.

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

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

Chatfield, C. (1995) Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society. Series A*, 158, 419–466.

Coulombe, P. & Lee, C. (2012) Defining keratin protein function in skin epithelia: epidermolysis bullosa simplex and its aftermath. *Journal of Investigative Dermatology*, 132, 763–775.

Dwan, K., Li, T., Altman, D.G. & Elbourne, D. (2019) Consort 2010 statement: extension to randomised crossover trials. *BMJ*, 366, l4378.

Erguler, K. (2016) *Package 'Barnard'*. https://github.com/kerguler/Barnard.

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

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*, e2200236. https://doi.org/10.1002/bimj.202200236. Epub ahead of print.

Hurvich, C. & Tsai, C. (1989) Regression and time series model selection in small samples. *Biometrika*, 76, 297–307.

Konietschke, F. & Pauly, M. (2012) A studentized permutation test for the nonparametric Behrens–Fisher problem in paired data. *Electronic Journal of Statistics*, 6, 1358–1372.

le Cessie, S. & van Houwelingen, J. (1994) Logistic regression for correlated binary data. *Applied Statistics*, 43, 95–108.

Li, P. & Redden, D. (2015) Small-sample performance of bias-corrected sandwich estimators for cluster-randomized trials with binary outcomes. *Statistics in Medicine*, 34, 281–96.

Long, J.S. & Ervin, L.H. (2000) Using heteroscedasticity consistent standard errors in the linear regression model. *American Statistician*, 54, 217–224.

MacKinnon, J.G. & White, H. (1985) Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29, 305–325.

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, 50–60.

Matsouaka, R. (2022) Robust statistical inference for matched win statistics. *Statistical Methods in Medical Research*, 31, 1423–1438.

Molenberghs, G., Kenward, M., Verbeke, G. & Birhanu, T. (2011) Pseudo-likelihood estimation for incomplete data. *Statistica Sinica*, 21, 187–206.

Molenberghs, G. & Verbeke, G. (2005) *Models for discrete longitudinal data*. New York: Springer.

Noguchi, K., Gel, Y., Brunner, E. & Konietschke, F. (2012) nparld: an r software package for the nonparametric analysis of longitudinal data in factorial experiments. *Statistical Software*, 50, 1–23.

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.

Rubarth, K., Pauly, M. & Konietschke, F. (2022) Ranking procedures for repeated measures designs with missing data: estimation, testing and asymptotic theory. *Statistical Methods in Medical Research*, 31, 105–118.

SAS Help Center (2020) *The FREQ procedures*. Available from: https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.5/procstat/procstat_freq_syntax03.htm [Accessed: 20 May 2021].

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, 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, 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, 892–901.

Wilson, E.B. (1927) Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22, 209–212.

SUPPORTING INFORMATION

The Web Appendix referenced in Section 2 is available with this paper at the Biometrics website on Wiley Online Library.

Table 1: 2 × 2 contingency table, with nij the observed subjects with the row and column characteristics, $n_{.j}$ and n_i. The column and row sum respectively and *N,* the total number of observations.

How to cite this article: Verbeeck, J., Geroldinger, M., Thiel, K., Hooker, A.C., Ueckert, S., Karlsson, M. et al. (2023) How to analyze continuous and discrete repeated measures in small-sample cross-over trials? *Biometrics*, 79, 3998–4011. https://doi.org/10.1111/biom.13920

APPENDIX A

TABLE A.1 GEE-type model—Type I error of the two-sided test on 5,000 permuted samples of the original EBS trial data, with different working correlation structure and with and without small-sample corrections.

	Binary		Count	
	Sample	Type I error	Sample	Type I error
Exchangeable				
No correction	4890	0.054	4980	0.094
Kauermann & Carroll	4890	0.037	4980	0.074
Fay & Graubard	4890	0.037	4981	0.074
Mancl & DeRouen	4890	0.025	4980	0.055
Heterogeneous autocorrelation				
no correction	4917	0.055	4616	0.094
Kauermann & Carroll	4616	0.038	4917	0.079
Fay & Graubard	4620	0.039	4917	0.082
Mancl & DeRouen	4616	0.027	4917	0.058
Unstructured				
no correction	4442	0.053	NC	NC
Kauermann & Carroll	4442	0.036	NC	NC
Fay & Graubard	4442	0.038	NC	NC
Mancl & DeRouen	4442	0.024	NC	NC

Abbreviation: NC, did not converge.