Title

Simulation study comparing analytical methods for single-item longitudinal patient-reported outcomes data

Permalink
https://escholarship.org/uc/item/9b6375gs

Journal
Quality of Life Research, 32(3)

ISSN
0962-9343

Authors
Calsavara, Vinicius F
Diniz, Márcio A
Tighiouart, Mourad
et al.

Publication Date
2023-03-01

DOI
10.1007/s11136-022-03267-z

Simulation study comparing analytical methods for single-item longitudinal patient-reported outcomes data

Vinicius F. Calsavara[1] · Márcio A. Diniz[1] · Mourad Tighiouart[1] · Patricia A. Ganz[2] · N. Lynn Henry[3] · Ron D. Hays[4] · Greg Yothers[5] · André Rogatko[1]

Accepted: 29 September 2022 / Published online: 17 October 2022

Abstract

Purpose Efficient analytical methods are necessary to make reproducible inferences on single-item longitudinal ordinal patient-reported outcome (PRO) data. A thorough simulation study was performed to compare the performance of the semiparametric probabilistic index models (PIM) with a longitudinal analysis using parametric cumulative logit mixed models (CLMM).

Methods In the setting of a control and intervention arm, we compared the power of the PIM and CLMM to detect differences in PRO adverse event (AE) between these groups using several existing and novel summary scores of PROs. For each scenario, PRO data were simulated using copula multinomial models. Comparisons were also exemplified using clinical trial data.

Results On average, CLMM provided substantially greater power than the PIM to detect differences in PRO-AEs between the groups when the baseline-adjusted method was used, and a small advantage in power when using the baseline symptom as a covariate.

Conclusion Although the CLMM showed the best performance among analytical methods, it relies on assumptions difficult to verify and that might not be fulfilled in the real world, therefore our recommendation is the use of PIM models with baseline symptom as a covariate.

Keywords Adverse event · Cumulative logit mixed model · Probabilistic index model · Patient-reported outcome · Type I and II errors · PRO-CTCAE

Introduction

In 2016, Congress passed the 21st Century Cures Act, authorizing $1.8 billion in funding for the Cancer Moonshot over 7 years. This effort brought together a large community of investigators and clinicians dedicated to expedite research to improve the lives of people with cancer and their loved ones. A Blue Ribbon Panel was assembled to make recommendations related to high priority areas for research funding, and one of the areas identified was the acceleration of research that can identify approaches to monitor and manage patient-reported symptoms [1]. This goal has been addressed in part through an NIH funding announcement in 2017 whose stated purpose was to stimulate the development of methods for better understanding the treatment tolerability by analyzing data from clinical trials using adverse event data through the use of the Common Terminology Criteria for Adverse Events (CTCAE), the Patient-Reported Outcome (PRO) version of the CTCAE (PRO-CTCAE™), and other clinically relevant data (e.g., stage of disease, laboratory findings, co-morbidities, concurrent medications).

CTCAE data that are assessed by the clinical investigator reflect diverse sources of data: laboratory studies, physical

✉ Vinicius F. Calsavara
vinicius.calsavara@cshs.org

[1] Cedars-Sinai Medical Center, Samuel Oschin Comprehensive Cancer Institute, Los Angeles, CA, USA

[2] University of California Los Angeles Jonsson Comprehensive Cancer Center, Los Angeles, CA, USA

[3] University of Michigan Rogel Cancer Center, Ann Arbor, MI, USA

[4] Department of Medicine, University of California Los Angeles, Los Angeles, CA, USA

[5] University of Pittsburgh, Pittsburgh, PA, USA

examination findings, as well as symptoms elicited during history-taking by the clinician. CTCAE data are usually collected after exposure to treatment (e.g., after the first cycle of treatment and subsequent cycles, or at timed intervals such as every three months). In contrast, PRO data reflect the patient's own assessment of a symptom or toxicity (e.g., pain, fatigue, nausea), and are generally collected prior to treatment (baseline), and then at some regular interval after treatment initiation. The development of the PRO-CTCAE item bank [2] has been a major addition to the toolbox for assessment cancer treatment toxicity, along with existing symptom measures used to assess cancer treatment-related toxicities (e.g., MD Anderson Symptom Inventory [3], the Edmonton Symptom Assessment System [4], Patient-Reported Outcomes Measurement Information System [5]). PRO data complement CTCAE data, especially when a pretreatment assessment is included. In addition, clinicians' ratings of symptoms or toxicities often underestimate the severity of the symptom [2, 6].

Traditionally, longitudinal CTCAE data have been summarized and reported as the maximum (Max) toxicity grade experienced by patients receiving a particular treatment. In our work to date, we have shown that an alternative summary, the Toxicity Index (TI) [7], has greater power to detect differences between treatments than max and average (Avg) as summary measures [8]. Moreover, we have shown that summary measures modeled using a semiparametric regression model, the probabilistic index model (PIM), allow incorporation of patients' characteristics in the model, increasing the understanding of toxicity as recently reported in two separate clinical trials [8, 9]. The TI may be particularly valuable when considering cancer therapies that are taken chronically, and where patients may be at risk for discontinuation due to the accumulation of multiple low-grade toxicities.

Longitudinal data analysis of PROs is traditionally performed considering an overall measure (e.g., sum, average) for a group of single items over time using standard regression models such as linear mixed models (LMM) [10, 11]. However, there is no standard approach for the analysis of PRO-CTCAE items, which vary from 0 to 4 that can be interpreted by themselves and they are not summarized into an overall measure. Publications to date related to the analysis of PRO-CTCAE items have been analyzed as the proportion of scores greater than a given cut-off value (e.g., score ≥ 3) with comparisons between treatments based on the Fisher's exact test or chi-squared test [12–14]. To incorporate the baseline PRO-CTCAE score in those analyses, a modified version of this approach has been proposed based on the change from baseline recording the maximum scores only for those scores that were worse than the baseline grade [12, 15, 16]. Recently, this approach to adjust by baseline was extended when calculating TI with comparisons between treatments based on the Wilcoxon rank-sum test [15]. Nonetheless, these methods do not allow incorporation of patients' covariates in the analysis of PRO-CTCAE items and modeling the change has been shown as statistically inefficient [17].

In this manuscript, we analyze PRO single items—such as PRO-CTCAE items—longitudinally using the cumulative logit mixed model (CLMM) as a direct extension of the LMM used for traditional PRO outcomes. We also summarize repeated single items over time with commonly used summary measures in CTCAE data—TI, Max and Avg—with the baseline score items incorporated following two modeling strategies: (i) calculating the change from baseline and (ii) considering baseline as a covariate. Simulation studies were used to compare these methods in terms of type I and II error rates in testing the presence of treatments effect. No study to date has compared the performance of these approaches for single items that are ordinal longitudinal data. To reflect the reality of most clinical trials, the generated datasets were also simulated according to the observed data from the randomized, double-blind NSABP B-35 clinical trial [18].

Methods

Clinical trial data

The case study consists of individual patient data from NSABP B-35, a Phase III double-blind randomized, placebo-controlled trial comparing 5 years of treatment with tamoxifen versus anastrozole in postmenopausal women with hormone receptor-positive ductal carcinoma in situ [19]. Patient-reported symptom data were collected at baseline and every 6 months after treatment initiation, and a list of predefined 30 CTCAEs was selected to evaluate the patient's toxicity experience during the trial. Each item was scored from 0 (least severe) to 4 (most severe). The CTCAE data provide detailed longitudinal information about the severity and types of toxicity. The trial included 3104 patients, with patient-reported symptom data available for the first 1194 patients. The statistical analysis methods were applied to three time points (baseline, and 6- and 12 months during therapy) for each CTCAE. Additional details of the trial are reported elsewhere [18, 19]. Observed proportions for each AE category by arm over time are summarized in the Online Supplement Section 1.

We briefly review the statistical methods used in the analysis of the clinical trial data and the simulation study for comparing toxicity between arms. Table 1 illustrates statistical analysis methods for single-item longitudinal data.

Table 1 Analysis methods for single-item longitudinal data

Score aggregated	Summary measure[c]	Method to summarize in a single measure	Selected scores[a]	Fitted model	Baseline as covariate in fitted model	Comparison between arms performed with statistical test	Statistical analysis method
Yes	TI	Baseline-adjusted	Follow-up scores that had at least a 1-score increase in AE score from baseline[b]	PIM	No	Wald test	PIM Baseline-adjusted TI
	Avg						PIM Baseline-adjusted Avg
	Max						PIM Baseline-adjusted Max
	TI	Post-baseline	Only follow-up scores	PIM	Yes	Wald test	PIM Baseline as covariate TI
	Avg						PIM Baseline as covariate Avg
	Max						PIM Baseline as covariate Max
No	None	No method applied	Only follow-up scores	CLMM	Yes	LRT	CLMM
						LRT by parametric bootstrap	CLMM (Bootstrap)

[a]Denotes how each method summarizes scores into a single value

[b]If the follow-up scores are less than or equal to the baseline score, then the summary measure is zero

[c]Summary measure is computed based on selected scores. LRT denotes likelihood ratio test

Repeated measures and incorporating baseline items

Each longitudinal single item can be analyzed considering two general approaches: (i) describe single items as ordinal repeated scores using an appropriate mixed regression model; (ii) summarize single items into a single measure, then apply an appropriate regression model for the single measure as the response variable. Several single measures have been proposed for the CTCAE data analysis—TI, Avg, and Max—therefore, we consider them as candidates as summary indexes for PRO data analysis as well.

In both approaches, baseline item can be incorporated following two reasonings: (a) as a covariate with post-baseline toxicity scores as the response variable; and (b) calculating a change in toxicity from baseline, denoted baseline-adjusted following previous work [12, 15, 16], such that change in toxicity score is the response variable. In this framework, toxicity summary indexes can be calculated either considering post-baseline scores and baseline-adjusted scores. Mathematical definitions of TI, Max, and Avg for post-baseline and baseline-adjusted scores with calculations are presented in the Online Supplement Section 2.

Probabilistic index model (PIM)

When summarizing repeated measures into a single score such as TI and Max, standard regression models may not be appropriate, and therefore, rank-based methods are applied. The PIMs are a class of semiparametric regression models [8, 20–24] in which the probability index (PI) is modeled as a function of covariates. The PI is a relative effect measure that corresponds to the probability that the outcome (e.g., toxicity) of a patient randomly sampled from group 1 is greater than the outcome of another patient randomly sampled from group 0, conditional on the covariate values of both patients [25]. In other words, let S_0 and S_1 be the random variables representing toxicity scores for groups 0 and 1, respectively: when PI is statistically significantly greater than 0.5 (i.e., PI=$P\left(S_0 < S_1\right) >$ 0.5), then patients in group 0 are more likely to have lower scores compared to patients in group 1; when PI is statistically significantly less than 0.5, then patients in group 0 are more likely to have higher scores than in group 1, and a PI equal to 0.5 means that both groups have similar toxicity score distribution.

In the single-item longitudinal data analysis, the PIM was fitted considering the toxicity burden summaries—TI, Avg, and Max—for (a) post-baseline scores as the response variable with baseline score and treatment as covariates (denoted as PIM Baseline as covariate) and (b) baseline-adjusted scores as the response variable with treatment as a covariate (denoted as PIM Baseline-adjusted). The PI with Wald-type 95% confidence interval was reported with p values based on the Wald statistic.

Cumulative logit mixed model (CLMM)

The CLMM allows modeling ordinal-scale data, i.e., observations that fall in an ordered finite set of categories, while taking into account the correlation structure of the data over time. It is a powerful model for such data since observations are treated correctly as categorical but the ordered nature of the data is exploited [26–28].

Although the CLMM model is appropriate to analyze ordinal-scale data over time, tests for statistical significance are problematic for both fixed and random effects. The likelihood ratio tests (LRT) are widely used to test fixed effects, but these tests can result in significant p values when the null hypothesis is true [29]. The LRT is usually appropriate for inference about random effects, but corrections are needed to address boundary problems [30, 31]. The reliable inferences for fixed and random effects in generalized linear mixed models can be done using alternative methods such as parametric bootstrap. Rather than relying on a chi-square distribution, the distribution of the LR statistic is constructed using a parametric bootstrap approach, which does not make any assumptions about degrees of freedom and the p value is estimated by using repeated sampling [32]. It is computationally intensive and requires longer run time than the usual LRT, but allows for better control of the type I error.

Longitudinal data analysis was performed with CLMM using maximum likelihood estimation with Laplace approximation. Each CLMM included the follow-up scores as a dependent variable, the intercept, an independent variable representing time points, baseline measurement and treatment variable as covariates, and a random patient effect. Odds ratios along with 95% confidence intervals (95%CI) are reported. The *p* value associated with the treatment effect was obtained by using the usual LRT and LRT by parametric bootstrap with 500 bootstrap resamples (see the Online Supplement Section 3 for a description of the CLMM and algorithm's implementation to perform the LRT by parametric bootstrap).

Throughout this article, we will denote CLMM and CLMM (Bootstrap) as the statistical analysis methods when LRT and LRT by parametric bootstrap were applied to single-item longitudinal data analysis, respectively.

Simulation studies

We conducted simulation studies based on a two-arm randomized-controlled design with ordinal response to mimic the B-35 trial adverse events data. Without loss of generality, we assume that both arms contain an equal number of 100 patients. The simulation studies were carried out considering only a single adverse event for each patient with five different ordinal categories (ranging from 0–4) evaluated at three time points. Correlated responses were generated using a copula multinomial model [33] by fixing the marginal probabilities for the categories at each time point and the correlation among the repeated measures in each arm.

For each generated dataset, the comparison between arms was performed using the following eight statistical analysis methods: (1 PIM Baseline-adjusted (1a) TI, (1b) Avg, and (1c) Max; 2) PIM Baseline as covariate (2a) TI, (2b) Avg, and (2c) Max; 3a) CLMM and 3b) CLMM (Bootstrap). For each scenario, 1000 independent datasets were generated to estimate the type I and II error rates (power = 1-type II error).

All simulations and analyses were conducted using the R software version 4.0 [34] with packages pim [35] and ordinal [36]. All hypotheses were two-tailed with 5% significance level. If the generated sample did not present data in at least two categories over time or any of these fitted models did not converge (i.e., to find a solution) within a reasonable number of iterations, the sample was to be redrawn. Thus, the simulation results are not influenced by numerical convergence issues. The convergence time of CLMM statistical analysis method tends to increase with the complexity of the model, especially the random effects structure. The CLMM (Bootstrap) analysis method is computationally intensive and requires longer run times, and so 500 resamples bootstrap was chosen to make the simulations computationally feasible. The results of the analysis methods for each simulated scenario took about 10 h of total CPU time using an Intel(R) Xeon(R) Gold 2.90 GHz [32 CPUs], 32 GB RAM, and Microsoft Windows 10 Enterprise.

Results

We briefly illustrate the results of a single-item longitudinal PRO data analysis (headache) from the B-35 trial using the statistical analysis methods aforementioned.

Furthermore, we report the results of the estimated type I error rates and power estimates obtained in the simulation study. Type I error rate is estimated as the proportion of rejection of the null hypothesis (H_0) when it is true, i.e., when there is not difference between arms, and the power is computed as the proportion of rejection of the when it is false, i.e., when the two groups are different over time.

Clinical trial data

The violin plots display the patient-reported headache symptom distribution summarized by baseline-adjusted and post-baseline methods, where higher headache score was observed for anastrozole treatment when the post-baseline method is applied, while for the baseline-adjusted method, the summary measure distribution shape was close in both treatments (see the Online Supplement Section 4).

Figure 1 illustrates the results of each statistical analysis method for the item headache. The hypothesis of equality between the distributions of toxicity scores of tamoxifen and anastrozole groups was not rejected by the PIM Baseline-adjusted TI, Avg, and Max. However, it was rejected (statistically significant difference between arms) for the PIM Baseline as covariate TI, Avg (PI = 0.453, 95%CI 0.419 to 0.488), and Max (PI = 0.456, 95%CI 0.422 to 0.490) statistical analysis methods, indicating a worse (higher) headache score for anastrozole than tamoxifen. Similar results from the approaches with PIM Baseline as covariate were obtained using the CLMM, where anastrozole is associated with a worse (higher) headache score (OR = 1.391, 95%CI 1.040 to 1.861).

As different statistical analysis methods did not lead to the same conclusion, it is not clear whether the differences between arms are true (power = 1–type II error) or false positives (type I error) when the CLMM and PIM baseline as covariate are applied. Inferences should be carried out with the method that provides greatest power to detect differences between treatments, while controlling type I error, therefore, we performed a thorough simulation study to compare the performance of the CLMM and CLMM (Bootstrap) with PIM Baseline-adjusted and PIM Baseline as covariate.

Simulation study

We performed a simulation study to evaluate the type I and II error rates of CLMM, CLMM (Bootstrap), PIM Baseline-adjusted, and PIM Baseline as covariate analysis methods in evaluating the treatment effects.

The marginal probabilities were set according to the observed proportions of each patient-reported symptom from the B-35 trial and three values for the correlations ($\rho = 0.2$, 0.5, and 0.9) among items over time. These values reflect the range of correlation levels observed in the B-35 trial for different items. We also simulate datasets with different values for the marginal probabilities under the assumption that the treatment groups are different or equal. A total of 270 scenarios were simulated with 225 scenarios showing differences between arms over time. A detailed algorithm on how to generate the datasets is described in Online Supplement Section 5. The R syntax code is available from the corresponding author upon request.

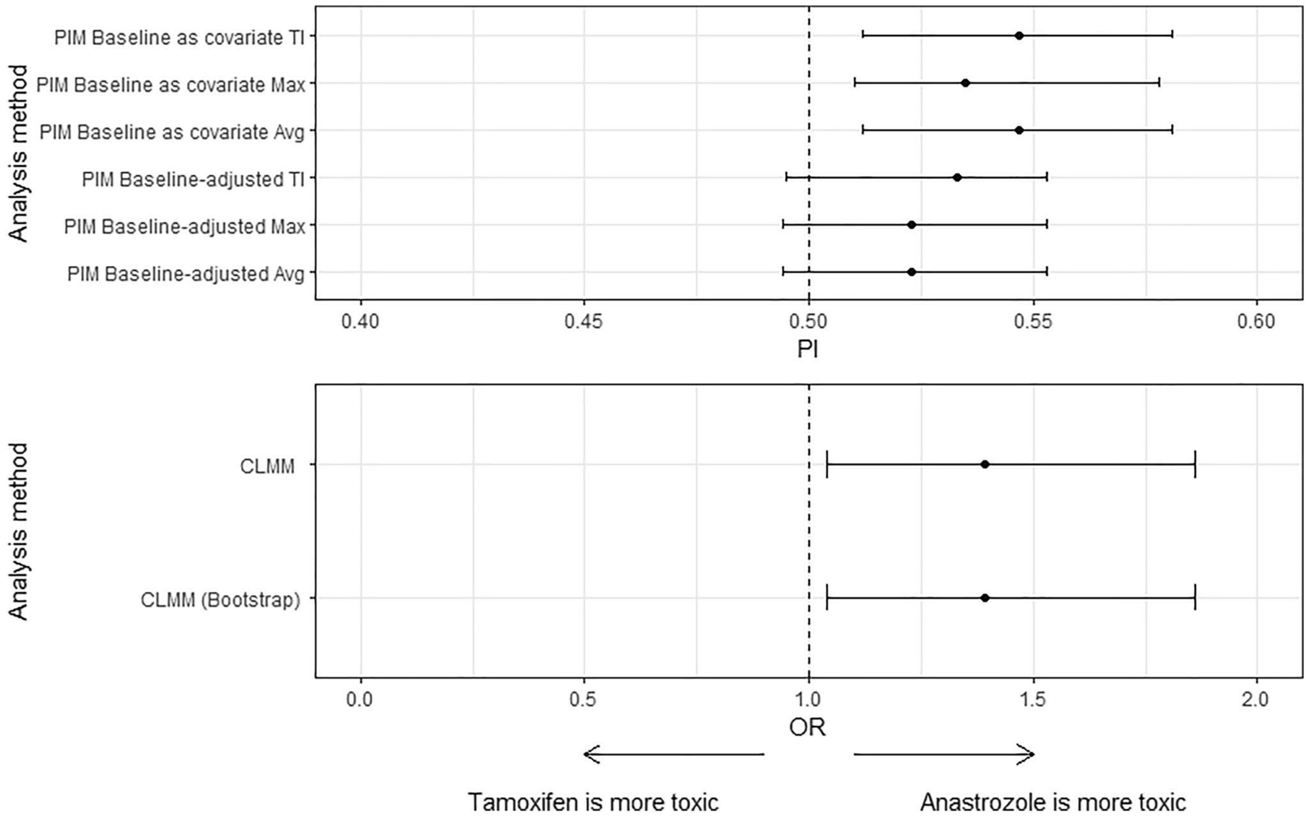

Fig. 1 Forest plots demonstrating results of PIM and CLMM for the headache. If the effect size (probability index) is less than 0.5, then probability of the toxicity for Tamoxifen is greater than Anastrozole is small, indicating that Anastrozole has higher headaches score. Odds ratio of the event $Y \geq k$, i.e., Anastrozole treatment increases (39.1%) the chance of higher headaches score compared to Tamoxifen treatment. Similar findings were observed for other B-35 trial items, where of the 28 items analyzed using the same analysis methods, four (14%) reached different conclusions

Initially, we evaluate the rejection rates of the null hypothesis when it is true, i.e., no difference between the two arms over time. At a 5% nominal significance level and 1000 trial replicates, we expect the simulated type I error rate to be between 3.65% and 6.35% (95% confidence interval of the fixed nominal level), and any procedure with type I error rate below this range will be considered conservative, and above this range will be considered anti-conservative. The estimated type I error rates based on 45 different scenarios are shown in Fig. 2. For the PIM Baseline-adjusted TI, Avg, Max, and PIM Baseline as covariate TI, Avg, and Max statistical analysis methods, the estimated type I error is close to the nominal level 5%, and in most scenarios, the rates are between 3.65% and 6.35%. On the other hand, the estimated type I error rates associated to the CLMM statistical analysis method are inflated when testing the null hypothesis of treatment effect, yielding minimum type I error rate of 8% (on average it was 12%). These results show that *p* values calculated from usual LRT are somewhat anti-conservative. Further, these *p* values appear to be more anti-conservative for higher correlation among repeated measures. On average, the estimated type I error rates were 10.7%, 11.6%, and 13.9% for correlation values 0.2, 0.5, and 0.9, respectively. However, when the CLMM (Bootstrap) analysis method is applied, it produced smaller type I error rates than CLMM. The empirical rejection rates are reasonably close to the 5% nominal level, regardless of fixed correlation among repeated measures. Overall, the type I error rate was close to 5% (on average it was 5.6%) when

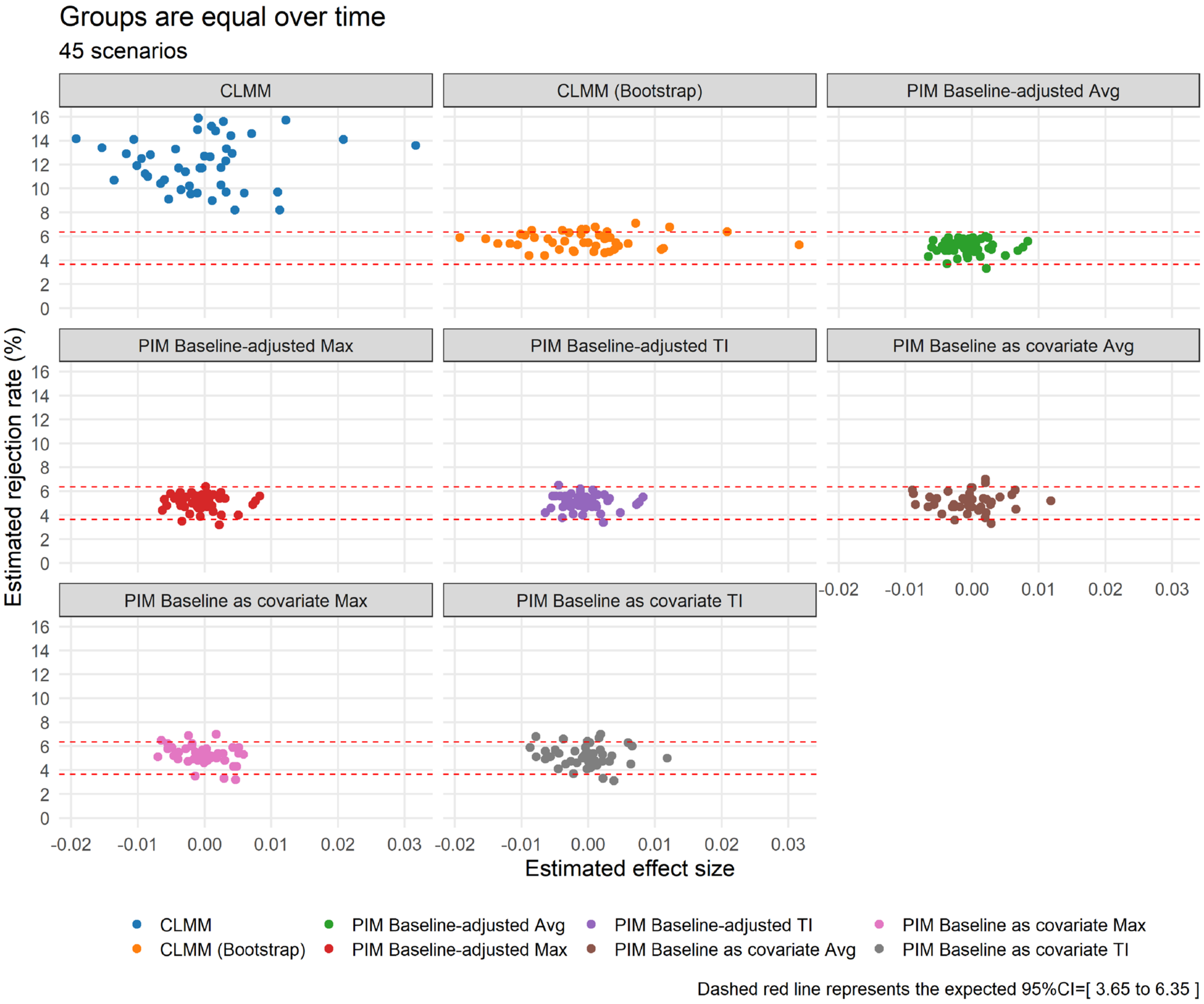

Fig. 2 Estimated type I error rates for detecting a significant treatment effect when the groups are equal over time, based on 45 simulated scenarios. Estimated effect size represents the logit of estimated probabilistic index when PIM is used, and under CLMM, it denotes the estimated regression coefficient associated with the group variable

CLMM (Bootstrap) was applied, a difference of 6.4% in estimated type I error rate compared to CLMM analysis method. The results suggest that the CLMM (Bootstrap) can produce acceptable type I error rates.

As illustrated in Fig. 2, the statistical analysis methods used in these simulation study produced type I error rates of 5%, except for CLMM. In this case, using the CLMM statistical analysis method may lead to an increase in statistical power because of the inflated type I error rate. Therefore, we considered CLMM (Bootstrap) for fair comparisons.

Due to the large number of scenarios under study, the power estimates comparisons were performed using a graphical representation and the area under the curve was used to summarize the differences. For each statistical analysis method, the empirical powers were plotted as one minus the empirical cumulative distribution function (1-ECDF). Figure 3 illustrates the power estimates comparisons among the statistical analysis methods, and the area under each curve is shown in Table 2. The y-axis can be interpreted as the probability of a given statistical analysis method to have power

Table 2 Overall power estimates and stratified by fixed correlation among repeated measures by analysis method

Statistical analysis method	Overall*			
	Power[a]	95%CI of power	Mean difference (MD)	95%CI of MD
PIM Baseline-adjusted Max	27.54%	23.80% to 31.28%	8.84%	5.86% to 11.83%
PIM Baseline-adjusted Avg	26.95%	23.25% to 30.65%	9.43%	6.43% to 12.44%
PIM Baseline-adjusted TI	28.14%	24.38% to 31.90%	8.25%	5.32% to 11.17%
PIM Baseline as covariate Max	34.12%	29.78% to 38.46%	2.26%	0.89% to 3.62%
PIM Baseline as covariate Avg	34.90%	30.53% to 39.27%	1.48%	0.90% to 2.08%
PIM Baseline as covariate TI	34.78%	30.41% to 39.15%	1.60%	0.88% to 2.33%
CLMM (Bootstrap)	36.38%	32.01% to 40.75%	Ref	Ref
	$\rho = 0.2$**			
PIM Baseline-adjusted Max	24.54%	18.66% to 30.42%	11.18%	5.3% to 17.06%
PIM Baseline-adjusted Avg	23.72%	17.96% to 29.48%	12.00%	6.08% to 17.92%
PIM Baseline-adjusted TI	25.33%	19.39% to 31.27%	10.40%	4.65% to 16.15%
PIM Baseline as covariate Max	33.40%	26.17% to 40.63%	2.32%	0.31% to 4.35%
PIM Baseline as covariate Avg	34.84%	27.35% to 42.33%	0.88%	0.34% to 1.42%
PIM Baseline as covariate TI	34.76%	27.31% to 42.21%	0.96%	− 0.16% to 2.08%
CLMM (Bootstrap)	35.72%	28.27% to 43.17%	Ref	Ref
	$\rho = 0.5$**			
PIM Baseline-adjusted Max	25.89%	19.65% to 32.13%	8.48%	3.98% to 12.99%
PIM Baseline-adjusted Avg	25.24%	19.08% to 31.40%	9.13%	4.60% to 13.66%
PIM Baseline-adjusted TI	26.42%	20.16% to 32.68%	7.95%	3.55% to 12.33%
PIM Baseline as covariate Max	32.58%	25.09% to 40.07%	1.79%	0.13% to 3.45%
PIM Baseline as covariate Avg	33.38%	25.83% to 40.93%	0.99%	0.39% to 1.60%
PIM Baseline as covariate TI	33.38%	25.77% to 40.99%	0.99%	0.003% to 1.98%
CLMM (Bootstrap)	34.37%	26.90% to 41.84%	Ref	Ref
	$\rho = 0.9$**			
PIM Baseline-adjusted Max	32.19%	24.78% to 39.60%	6.87%	1.63% to 12.1%
PIM Baseline-adjusted Avg	31.88%	24.51% to 39.25%	7.18%	1.90% to 12.4%
PIM Baseline-adjusted TI	32.66%	25.19% to 40.13%	6.40%	1.22% to 11.6%
PIM Baseline as covariate Max	36.39%	28.30% to 44.48%	2.67%	− 0.57% to 5.90%
PIM Baseline as covariate Avg	36.47%	28.48% to 44.46%	2.59%	1.00% to 4.17%
PIM Baseline as covariate TI	36.19%	28.16% to 44.22%	2.87%	1.28% to 4.46%
CLMM (Bootstrap)	39.06%	30.97% to 47.15%	Ref	Ref

[a] Mean power estimates obtained under area 1-ECDF curve of each statistical analysis method. Mean difference (MD) represents the mean difference of power estimates by CLMM compared to other statistical analysis methods. 95%CI was computed using the paired *t* test considering the sample of power estimates. 95%CIs that exclude zero indicate statistically significant differences at the 5% significance level (two-sided hypothesis test)

*Based on 225 simulated scenarios

**Based on 75 simulated scenarios

greater than a given threshold among 225 scenarios and the x-axis represents possible thresholds based on the power estimates in 225 scenarios. Curves at the top indicate that a larger number of scenarios showed power greater than a given threshold than curves at the bottom.

Our simulations suggest that the CLMM (Bootstrap) statistical analysis method is more powerful than the PIM Baseline-adjusted method to detect difference between treatments (Fig. 3). The power of the PIM Baseline as covariate statistical analysis method was superior to Baseline-adjusted but lower than that of the CLMM (Bootstrap) approach. Specifically, the CLMM (Bootstrap) provided on average higher power estimates (36.38%, 95%CI 32.01% to 40.75%) than PIM Baseline as covariate TI (34.78%, 95%CI 30.41% to 39.15%), Avg (34.9%, 95%CI 30.53% to 39.27%), and Max (34.12%, 95%CI 29.78% to 38.46%) as well as PIM Baseline-adjusted TI (28.14%, 95%CI 24.38% to 31.9%), Avg (26.95%, 23.25% to 30.65%), and Max (27.54%, 95%CI 23.8% to 31.28%).

When comparing statistical methods over scenarios, there was a statistically significant mean difference (MD) of 8.25% (95%CI 5.32% to 11.17%), 9.43% (95%CI 6.43% to 12.44%), and 8.84% (95%CI 5.86% to 11.83%) between the power estimates by CLMM (Bootstrap) analysis method and PIM Baseline-adjusted TI, Avg, and Max, respectively. A statistically significant MD was also observed between the power estimates by CLMM (Bootstrap) versus PIM Baseline as covariate TI (MD = 1.6%, 95%CI 0.88% to 2.33%), Avg (MD = 1.48%, 95%CI 0.9% to 2.08%), and Max (MD = 2.26%, 95%CI 0.89% to 3.62%). The same pattern was observed in the subset of simulations with low ($\rho = 0.2$), moderate ($\rho = 0.5$), and high ($\rho = 0.9$) correlation among the repeated measures, where higher MDs between CLMM (Bootstrap) and PIM Baseline-adjusted, and greater than PIM Baseline as covariate (Table 2 and Fig. 4).

When comparing the results excluding the CLMM statistical method, there was a statistically significant MD of 6.64% (95%CI 3.81% to 9.47%), 7.95% (95%CI 4.95% to 10.94%), and 6.58% (95%CI 3.97% to 9.2%) between the power estimates by PIM Baseline as covariate TI, Avg, and Max and PIM Baseline-adjusted TI, Avg, and Max, respectively. The same pattern was also observed in the subset of simulations with low and moderate correlation among the measurements. There were no differences between the power estimates by PIM Baseline as covariate and PIM Baseline-adjusted under the assumption of high correlation among the repeated measures (Online Supplement Section 5, Table S5).

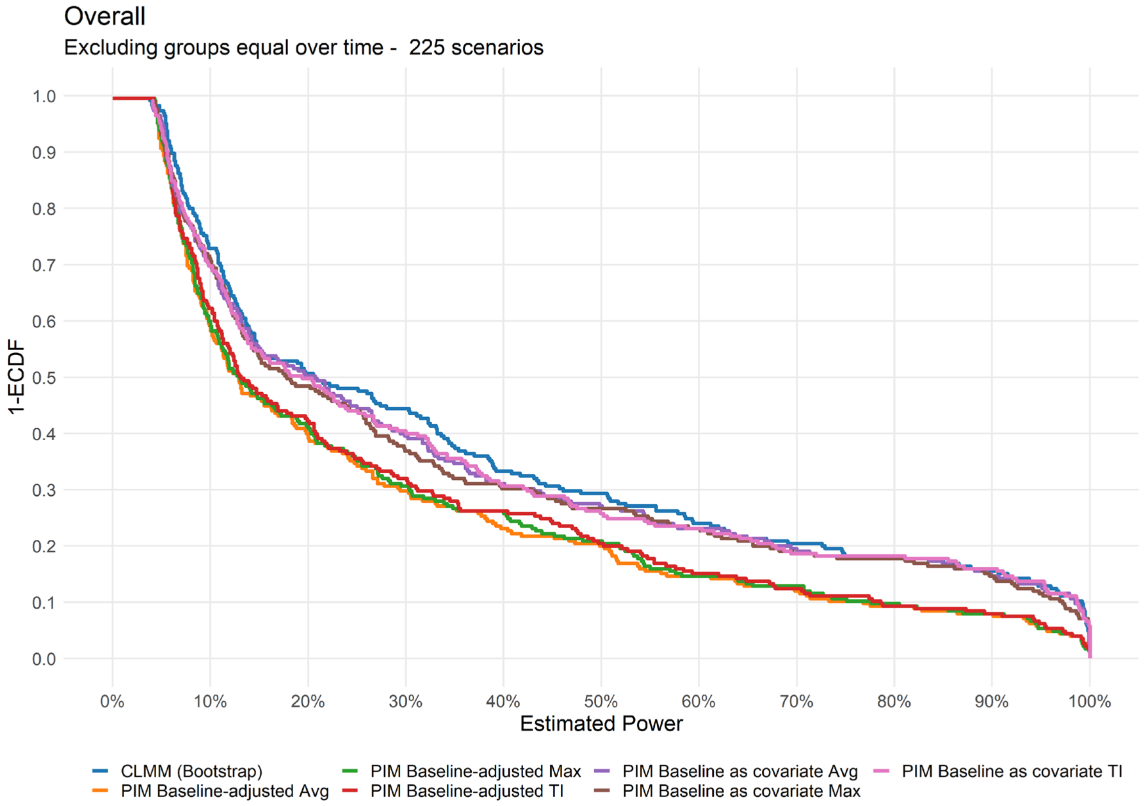

Fig. 3 One minus the cumulative distribution function of the overall power estimates by statistical analysis methods under the assumption that the groups are different

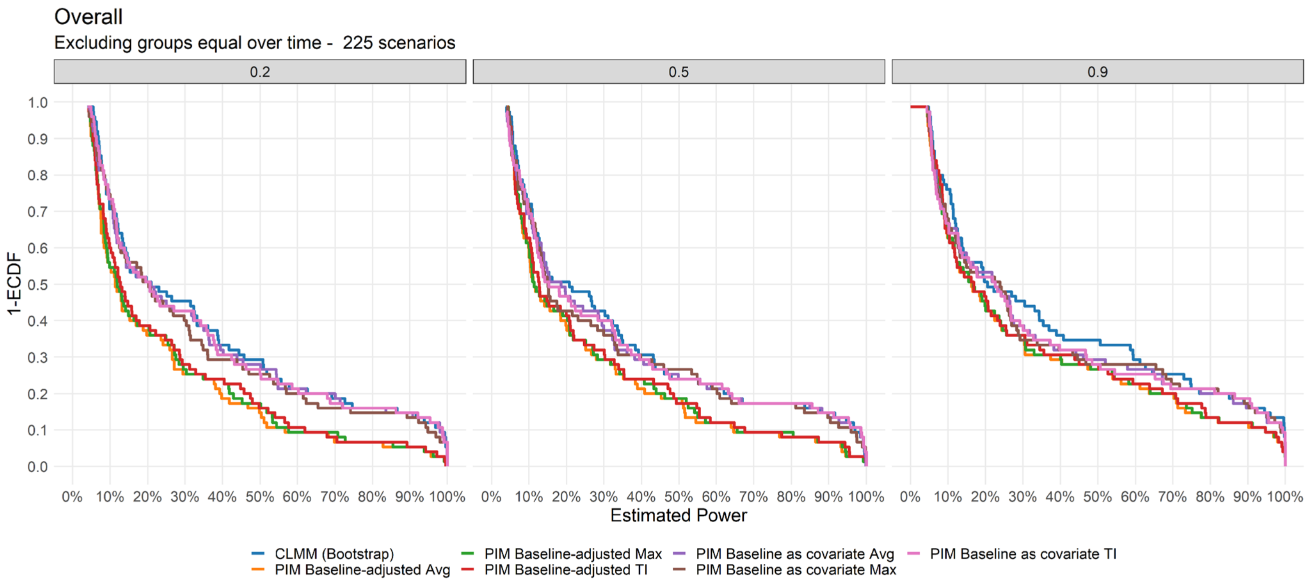

Fig. 4 One minus the cumulative distribution function of the overall power estimates by statistical analysis methods stratified by correlations ($\rho = 0.2$, 0.5, and 0.9) among the repeatead measures under the assumption that the groups are different

Similar results were also observed within each simulated data type (Online Supplement Section 1 describes all simulated scenarios). Higher MDs of power estimates occurred under the assumption that the groups are equal only at baseline and different at times 2 and 3. A statistically significant MD ranging from 16.51% to 17.93% was observed between CLMM (Bootstrap) and PIM Baseline-adjusted analysis methods, and from 2.81% to 3.98% between CLMM (Bootstrap) and PIM Baseline as covariate. When the simulated data followed the assumptions that the groups are different at baseline with same increments in times 2 and 3, a statistically significant MD of 5.37% to 5.78% and 1.75% to 2.33% was observed between CLMM (Bootstrap) and PIM Baseline-adjusted and between the CLMM (Bootstrap) and PIM Baseline as covariate, respectively. Under the B-35 trial scenarios, a statistically significant MD ranging from 5.41% to 6.72% occurred between the CLMM (Bootstrap) and PIM Baseline-adjusted, while a MD varying of 1.52% to 2.60% was observed between CLMM (Bootstrap) and PIM Baseline as covariate. There were no differences in power estimates among the methods under the assumption that the groups are different in all time points (see the Online Supplement Section 5, Table S4). When excluding the CLMM statistical analysis method, higher MDs of power estimates occurred between PIM Baseline as covariate and PIM Baseline-adjusted under the assumption that the groups are equal only at baseline and different at times 2 and 3. There was a statistically significant MD of 13.7% (95%CI 6.7% to 20.69%), 15.06% (95%CI 7.8% to 22.32%), and 13.01% (95%CI 6.44% to 19.59%) between the power estimates by PIM Baseline as covariate TI, Avg, and Max and PIM Baseline-adjusted TI, Avg, and Max, respectively. Under the assumptions that the groups are different at baseline with same increments in times 2 and 3, a statistically significant MD of 4.46% (95%CI 0.33% to 8.59%) and 3.45% (95%CI 0.2% to 6.7%) was observed between PIM Baseline as covariate Avg, and Max and PIM Baseline-adjusted Avg, and Max, respectively. When the simulated data followed the B-35 trial scenarios, a statistically significant MD of 3.89% (95%CI 2.49% to 5.29%), 5.36% (95%CI 3.53% to 7.18%), and 3.51% (95%CI 2.21% to 4.82%) occurred between PIM Baseline as covariate TI, Avg, Max and PIM Baseline-adjusted TI, Avg, Max, respectively. Under the assumption that the groups are different in all time points, there were no differences in power estimates between PIM Baseline as covariate and PIM Baseline-adjusted statistical analysis methods (see the Online Supplement Section 5, Table S5).

Discussion

Patient-reported symptom items (e.g., PRO-CTCAE) are increasingly considered as indicators of toxicity in [16, 37–40] clinical trials. Different from clinician-reported CTCAE data, PRO-CTCAE items are collected from patients at baseline to measure symptom burden before treatment, allowing for separation of burden caused from disease and treatment. While overall summary scores are often used to evaluate toxicity, single items provide information about specific adverse events. Therefore, robust statistical methods are needed for single items that incorporate baseline symptoms while capturing the overall patients' toxicity burden over time. This paper provides a comprehensive simulation study comparing analytical methods for single-item longitudinal PRO data from the B-35 clinical trial.

Our simulations indicate that the CLMM had an estimated type I error rates routinely higher than the 5% nominal significance level. In addition, higher estimated type I error rates were observed as the correlation among the repeated measures increased. In contrast, the LRT parametric bootstrap (CLMM (Bootstrap)) was able to produce acceptable type I error rates. Similar results close to 5% were also observed for the PIM Baseline-adjusted and PIM Baseline as covariate analysis methods.

Simulation results revealed important differences in power estimates among the analysis methods that maintained the type I error rate close to 5% (i.e., excluding CLMM): CLMM (Bootstrap) analysis method provided substantially higher power estimates than PIM Baseline-adjusted and a smaller but significant improvement in power compared to PIM Baseline as covariate. There was a statistically significant MD of approximately 9% between the power estimates of CLMM (Bootstrap) and PIM Baseline-adjusted analysis methods; and a statistically significant MD of roughly 2% between CLMM (Bootstrap) and PIM Baseline as covariate analysis methods. Among the PIM statistical methods, higher power estimates were observed for the PIM Baseline as covariate with a statistically significant MD of approximately 6.5% between the power estimates of PIM Baseline as covariate Max, TI and PIM Baseline-adjusted Max, TI; and a statistically significant MD about 8% between PIM Baseline as covariate TI and PIM Baseline as covariate TI. The same pattern was also observed for the simulated scenario that mimicked the observed proportions of B-35 trial items and the other different scenarios, except when the groups were different in all time points. Therefore, the difference between tamoxifen and anastrozole for headache detected by CLMM and PIM Baseline as covariate is plausible given both methods showed greater power than PIM Baseline-adjusted analysis method.

Although the PIM Baseline as covariate is a semiparametric approach, it provided average power close to the CLMM (Bootstrap) analysis method in most of the simulated scenarios. The small difference between average powers obtained by both analysis methods may be associated with the small number of assessment time points after baseline (times 2 and 3). The CLMM analysis method considers all patient's follow-up scores when modeling, while PIM Baseline as covariate summarizes them into a single value. With a greater number of assessments during follow-up, the difference between analytical methods might increase. Another study is needed to evaluate the performance of the analysis methods for a greater number of assessments. The current study found, consistent with previous work [17], that the PIM Baseline-adjusted analysis method shows lower power.

Several strengths are associated with the use of CLMM. It considers all the patient's experience over the course of the trial by accounting for the dependency structure among ordinal repeated measures, without losing interpretability. This additional information provides, on average, greater power to detect differences between treatments compared to the PIM Baseline-adjusted and PIM Baseline as covariate analysis methods, resulting in fewer number of patients needed to detect treatment differences. However, it shows unacceptable type I error rate without bootstrap, yielding minimum type I error rate of 8% (on average it was 12%), and when LRT by parametric bootstrap is applied, it is computationally intensive. In addition, it relies on assumptions of proportional odds, normality of the random effects and that the data generator process in the parametric bootstrap is the same as the data generator process of the observed data. On the other hand, PIM Baseline as covariate can provide results close to the CLMM (Bootstrap) method when the number of assessments is small, which is an advantage, because the PIM does not require any distribution assumption and it is computationally fast.

The limitations of this study are that these methods were applied in simulated datasets of 100 patients per group measured in three time points with same correlation among them (0.2, 0.5, and 0.9), and therefore, the impact of different sample sizes and/or repeated measures on type I error rates and power were not evaluated. In addition, the sample generator was based on copula multinomial model, an approach different from those considered in this study, which made it possible to generate datasets according to the item severities in both groups, and to vary correlations among the measures. The results for CLMM are based on simulations using random intercept mixed model with three fixed effects, so the observed type I error rates might not hold up for a model with a more complex structure and smaller sample sizes.

As the performance of the semiparametric approaches with baseline as covariate is close to the parametric approach, we hypothesize that parametric approaches using TI or other summary measures may be even more powerful methods to analyze longitudinal data with single items. Nevertheless, the quasi-continuous nature [41] of TI currently constrains the use of common parametric models. Further research is necessary, but the results of this study allow us to feel relatively confident that anastrozole treatment was associated with worse headache symptoms.

Conclusion

The CLMM showed unacceptable type I error rate without bootstrap, while the CLMM with bootstrap provided the best performance among analytical methods. However, it relies on assumptions difficult to verify and that might not be fulfilled in the real world, and therefore, our recommendation is the use of PIM models using Max or TI score with baseline as covariate considering the lack of interpretability of average AE score discussed in our previous work [41].

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11136-022-03267-z.

Acknowledgements The authors acknowledge members of the U01 Scientific Advisory Committee (Lari Wenzel, PhD; Claire Snyder, PhD; Michael Brundage, MD; and Elisa Long, PhD) and NRG Oncology statisticians (Hanna Bandos, PhD, and Reena Cecchini, PhD).

Author Contributions VFC, MAD, MT, and AR have design the study; GY has acquired the data; VFC and MAD have analyzed the data and performed the simulation studies; VFC, MAD, MT, and AR have interpreted the results; VFC has drafted the manuscript. All authors revised, read, and approved the final manuscript.

Funding Supported in part by the National Cancer Institute of the NIH (Grant No. 1U01CA232859-01; V.F.C., M.A.D., M.T., G.Y., R.D.H., P.A.G., A.R.); the NIH National Center for Advancing Translational Sciences UCLA CTSI (UL1 TR001881-01; M.A.D., M.T., A.R.); and NCI grants P01CA233452-02, U10CA180868, UG1CA189867, and U10CA180822. This program is supported by funding provided through the Cancer Moonshot.

Data availability The data that support the findings in the case study are available from NRG Oncology but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Requests for the data, however, can be made to NRG Oncology at https://www.nrgoncology.org/Resources/Ancillary-Projects-Data-Sharing-Application.

Declarations

Conflict of interest The authors have no relevant financal or non-financial interests to disclose.

Ethical approval This study was conducted with de-identified data obtained from the NRG Oncology Statistical Center for a completed clinical trial whose primary results have been published. This use of de-identified data from federally funded clinical is exempt from the requirements for IRB review and approval in accordance with federal regulations, 45 CFR 46.101(b).

Consent to participate Informed consent was obtained from participants in the study original study. This secondary analysis was exempt.

References

1. *Blue Ribbon Panel: Cancer Moonshot Blue Ribbon Panel Report*. (2016) Retrieved from https://www.cancer.gov/research/key-initiatives/moonshot-cancer-initiative/blue-ribbon-panel-report-2016.pdf.
2. Basch, E., Reeve, B. B., Mitchell, S. A., Clauser, S. B., Minasian, L. M., Dueck, A. C., Mendoza, T. R., Hay, J., Atkinson, T. M., Abernethy, A. P., Bruner, D. W., Cleeland, C. S., Sloan, J. A., Chilukuri, R., Baumgartner, P., Denicoff, A., St Germain, D., O'Mara, A. M., Chen, A., … Schrag, D. (2014). Development of the National Cancer Institute's patient-reported outcomes version of the common terminology criteria for adverse events (PRO-CTCAE). *Journal of the National Cancer Institute*. https://doi.org/10.1093/jnci/dju244
3. Cleeland, C. S., Mendoza, T. R., Wang, X. S., Chou, C., Harle, M. T., Morrissey, M., & Engstrom, M. C. (2000). Assessing symptom distress in cancer patients: the MD Anderson Symptom Inventory. *Cancer, 89*(7), 1634–46. https://doi.org/10.1002/1097-0142(20001001)89:7%3c1634::aid-cncr29%3e3.0.co;2-v
4. Bruera, E., Kuehn, N., Miller, M. J., Selmser, P., & Macmillan, K. (1991). The Edmonton Symptom Assessment System (ESAS): A simple method for the assessment of palliative care patients. *Journal of Palliative Care, 7*(2), 6–9.
5. Cella, D., Yount, S., Rothrock, N., Gershon, R., Cook, K., Reeve, B., Ader, D., Fries, J. F., Bruce, B., Rose, C., P.C. Group. (2007). The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years. *Medical Care, 45*(5 Suppl 1), S3–S11. https://doi.org/10.1097/01.mlr.0000258615.42478.55
6. Dueck, A. C., Mendoza, T. R., Mitchell, S. A., Reeve, B. B., Castro, K. M., Rogak, L. J., Atkinson, T. M., Bennett, A. V., Denicoff, A. M., O'Mara, A. M., Li, Y., Clauser, S. B., Bryant, D. M., Bearden, J. D., 3rd., Gillis, T. A., Harness, J. K., Siegel, R. D., Paul, D. B., Cleeland, C. S., … Basch, E. (2015). Validity and Reliability of the US National Cancer Institute's Patient-Reported Outcomes Version of the Common Terminology Criteria for Adverse Events (PRO-CTCAE). *JAMA Oncology, 1*(8), 1051–9. https://doi.org/10.1001/jamaoncol.2015.2639
7. Rogatko, A., Babb, J. S., Wang, H., Slifker, M. J., & Hudes, G. R. (2004). Patient characteristics compete with dose as predictors of acute treatment toxicity in early phase clinical trials. *Clinical Cancer Research, 10*(14), 4645–51. https://doi.org/10.1158/1078-0432.CCR-03-0535
8. Gresham, G., Diniz, M. A., Razaee, Z. S., Luu, M., Kim, S., Hays, R. D., Piantadosi, S., Tighiouart, M., Yothers, G., Ganz, P. A., & Rogatko, A. (2020). Evaluating treatment tolerability in cancer clinical trials using the toxicity index. *Journal of the National Cancer Institute, 112*(12), 1266–1274. https://doi.org/10.1093/jnci/djaa028
9. Henry, N. L., Kim, S., Hays, R. D., Diniz, M. A., Luu, M., Cecchini, R. S., Yothers, G., Rogatko, A., & Ganz, P. A. (2021). Toxicity index, patient-reported outcomes, and early discontinuation of endocrine therapy for breast cancer risk reduction in NRG oncology/NSABP B-35. *Journal of Clinical Oncology, 39*(34), 3800–3812. https://doi.org/10.1200/JCO.21.00910
10. Van Der Weijst, L., Surmont, V., Schrauwen, W., & Lievens, Y. (2021). Real life data on patient-reported outcomes and neurocognitive functioning of lung cancer patients: The PRO-long study. *Frontiers in Oncology, 11*, 685605. https://doi.org/10.3389/fonc.2021.685605
11. Li, J. B., Guo, S. S., Tang, L. Q., Guo, L., Mo, H. Y., Chen, Q. Y., & Mai, H. Q. (2020). Longitudinal trend of health-related quality of life during concurrent chemoradiotherapy and survival in

patients with stage II-IVb nasopharyngeal carcinoma. *Frontiers in Oncology, 10*, 579292. https://doi.org/10.3389/fonc.2020.579292

12. Basch, E., Rogak, L. J., & Dueck, A. C. (2016). Methods for implementing and reporting patient-reported outcome (PRO) measures of symptomatic adverse events in cancer clinical trials. *Clinical Therapeutics, 38*(4), 821–30. https://doi.org/10.1016/j.clinthera.2016.03.011
13. Adelstein, D. J., Li, Y., Adams, G. L., Wagner, H., Jr., Kish, J. A., Ensley, J. F., Schuller, D. E., & Forastiere, A. A. (2003). An intergroup phase III comparison of standard radiation therapy and two schedules of concurrent chemoradiotherapy in patients with unresectable squamous cell head and neck cancer. *Journal of Clinical Oncology, 21*(1), 92–8. https://doi.org/10.1200/JCO.2003.01.008
14. Forastiere, A. A., Goepfert, H., Maor, M., Pajak, T. F., Weber, R., Morrison, W., Glisson, B., Trotti, A., Ridge, J. A., Chao, C., Peters, G., Lee, D. J., Leaf, A., Ensley, J., & Cooper, J. (2003). Concurrent chemotherapy and radiotherapy for organ preservation in advanced laryngeal cancer. *New England Journal of Medicine, 349*(22), 2091–2098. https://doi.org/10.1056/NEJMoa031317
15. Langlais, B., Mazza, G. L., Thanarajasingam, G., Rogak, L. J., Ginos, B., Heon, N., Scher, H. I., Schwab, G., Ganz, P. A., Basch, E., & Dueck, A. C. (2021). Evaluating treatment tolerability using the toxicity index with patient-reported outcomes data. *Journal of Pain and Symptom Management*. https://doi.org/10.1016/j.jpainsymman.2021.07.031
16. Dueck, A. C., Scher, H. I., Bennett, A. V., Mazza, G. L., Thanarajasingam, G., Schwab, G., Weitzman, A. L., Rogak, L. J., & Basch, E. (2020). Assessment of adverse events from the patient perspective in a phase 3 metastatic castration-resistant prostate cancer clinical trial. *JAMA Oncology, 6*(2), e193332. https://doi.org/10.1001/jamaoncol.2019.3332
17. Vickers, A. J. (2001). The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient: a simulation study. *BMC Medical Research Methodology, 1*, 6. https://doi.org/10.1186/1471-2288-1-6
18. Margolese, R. G., Cecchini, R. S., Julian, T. B., Ganz, P. A., Costantino, J. P., Vallow, L. A., Albain, K. S., Whitworth, P. W., Cianfrocca, M. E., Brufsky, A. M., Gross, H. M., Soori, G. S., Hopkins, J. O., Fehrenbacher, L., Sturtz, K., Wozniak, T. F., Seay, T. E., Mamounas, E. P., & Wolmark, N. (2016). Anastrozole versus tamoxifen in postmenopausal women with ductal carcinoma in situ undergoing lumpectomy plus radiotherapy (NSABP B-35): a randomised, double-blind, phase 3 clinical trial. *Lancet, 387*(10021), 849–56. https://doi.org/10.1016/S0140-6736(15)01168-X
19. Forbes, J. F., Sestak, I., Howell, A., Bonanni, B., Bundred, N., Levy, C., von Minckwitz, G., Eiermann, W., Neven, P., Stierer, M., Holcombe, C., Coleman, R. E., Jones, L., Ellis, I., Cuzick, J., IBIS-II investigators. (2016). Anastrozole versus tamoxifen for the prevention of locoregional and contralateral breast cancer in postmenopausal women with locally excised ductal carcinoma in situ (IBIS-II DCIS): a double-blind, randomised controlled trial. *Lancet, 387*(10021), 866–73. https://doi.org/10.1016/S0140-6736(15)01129-0
20. Fay, M. P., & Malinovsky, Y. (2018). Confidence intervals of the Mann-Whitney parameter that are compatible with the Wilcoxon-Mann-Whitney test. *Statistics in Medicine, 37*(27), 3991–4006. https://doi.org/10.1002/sim.7890
21. De Neve, J., & Thas, O. (2017). A Mann-Whitney type effect measure of interaction for factorial designs. *Communications in Statistics-Theory and Methods, 46*(22), 11243–11260. https://doi.org/10.1080/03610926.2016.1263739
22. De Neve, J., & Thas, O. (2015). A regression framework for rank tests based on the probabilistic index model. *Journal of the American Statistical Association, 110*(511), 1276–1283. https://doi.org/10.1080/01621459.2015.1016226
23. De Neve, J., Thas, O., & Ottoy, J.-P. (2013). Goodness-of-fit methods for probabilistic index models. *Communications in Statistics-Theory and Methods, 42*(7), 1193–1207. https://doi.org/10.1080/03610926.2012.695851
24. Thas, O., Neve, J. D., Clement, L., & Ottoy, J. P. (2012). Probabilistic index models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 74*(4), 623–671. https://doi.org/10.1111/j.1467-9868.2011.01020.x
25. De Schryver, M., & De Neve, J. (2019). A tutorial on probabilistic index models: Regression models for the effect size P(Y1 < Y2). *Psychological Methods, 24*(4), 403–418. https://doi.org/10.1037/met0000194
26. Christensen, R.H. (2019). Cumulative link models for ordinal regression with the R package ordinal. Technical University of Denmark & Christensen Statistics. https://cran.rproject.org/web/packages/ordinal/vignettes/clm_article.pdf
27. Christensen, R.H.B. (2019). Regression Models for Ordinal Data [R package ordinal version 2019.12–10]
28. Hedeker, D., & Gibbons, R. D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics, 50*(4), 933–944.
29. Pinheiro, J., & Bates, D. (2006). *Mixed-effects models in S and S-PLUS*. Springer science & business media.
30. Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H., & White, J. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution, 24*(3), 127–35. https://doi.org/10.1016/j.tree.2008.10.008
31. Molenberghs, G., & Verbeke, G. (2007). Likelihood ratio, score, and Wald tests in a constrained parameter space. *The American Statistician, 61*(1), 22–27. https://doi.org/10.1198/000313007X171322
32. Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC Press.
33. Razaee, Z. S., & Amini, A. A. (2020). The Potts-Ising model for discrete multivariate data. *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), 33*, 13727–13737.
34. R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.
35. Meys, J., De Neve, J., Sabbe, N., Amorim, G. G. (2020). pim: Fit Probabilistic Index Models. R package version 2.0.2. https://CRAN.R-project.org/package=pim.
36. Christensen, R. H. B. (2015). Ordinal—regression models for ordinal data. *R package version, 28*, 2015.
37. Kluetz, P. G., Chingos, D. T., Basch, E. M., & Mitchell, S. A. (2016). Patient-reported outcomes in cancer clinical trials: measuring symptomatic adverse events with the national cancer institute's patient-reported outcomes version of the common terminology criteria for adverse events (PRO-CTCAE). *American Society of Clinical Oncology Educational Book, 35*, 67–73. https://doi.org/10.14694/EDBK_159514
38. Kluetz, P. G., Kanapuru, B., Lemery, S., Johnson, L. L., Fiero, M. H., Arscott, K., Barbachano, Y., Basch, E., Campbell, M., & Cappelleri, J. C. (2018). Informing the tolerability of cancer treatments using patient-reported outcome measures: summary of an FDA and critical path institute workshop. *Value in Health, 21*(6), 742–747. https://doi.org/10.1016/j.jval.2017.09.009
39. Zhou, X., Eid, D., & Gnanasakthy, A. (2018). Methods for reporting the patient-reported outcomes version of the common terminology criteria for adverse events (PRO-CTCAE) data in cancer clinical trials. *Value in Health, 21*, S226.

40. Xia, H., Fourie, J. Z., Booth, B., Wang, Y., Fiero, M. H., Przepiorka, D., & Kluetz, P. (2017). Exploration of PRO-CTCAE data used for exposure-response relationships in an oncology clinical trial. *Blood., 130*(1 Supplement), 5670–5670. https://doi.org/10.1182/blood.V130.Suppl_1.5670.5670

41. Razaee, Z. S., Arash, A., Amini, M. A., Diniz, M., Tighiouart, G. Y., & Rogatko, A. (2021). On the properties of the toxicity index and its statistical efficiency. *Statistics in Medicine, 40*(6), 1535–1552. https://doi.org/10.1002/sim.8858

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.