Qian *et al.*
Health and Quality of Life Outcomes (2025) 23:45
https://doi.org/10.1186/s12955-025-02373-z

Health and Quality of Life Outcomes

RESEARCH

Open Access

Comparison of statistical methods for the analysis of patient-reported outcomes (PROs), particularly the Short-Form 36 (SF-36), in randomised controlled trials (RCTs) using standardised effect size (SES): an empirical analysis

Yirui Qian[1*], Stephen J. Walters[2], Richard M. Jacques[2] and Laura Flight[2]

Abstract

Background The Short-Form 36 (SF-36), a widely used patient-reported outcome (PRO), is a questionnaire completed by patients measuring health outcomes in clinical trials. The PRO scores can be discrete, bounded, and skewed. Various statistical methods have been suggested to analyse PRO data, but their results may not be presented on the same scale as the original score, making it difficult to interpret and compare different approaches. This study aims to unify and compare the estimates from different statistical methods for analysing PROs, particularly the SF-36, in randomised controlled trials (RCTs), using standardised effect size (SES) summary measure.

Methods SF-36 outcomes were analysed using ten statistical methods: multiple linear regression (MLR), median regression (Median), Tobit regression (Tobit), censored absolute least deviation regression (CLAD), beta-binomial regression (BB), binomial-logit-normal regression (BLN), ordered logit model (OL), ordered probit model (OP), fractional logistic regression (Frac), and beta regression (BR). Each SF-36 domain score at a specific follow-up in three clinical trials was analysed. The estimated treatment coefficients and SESs were generated, compared, and interpreted. Model fit was evaluated using the Akaike information criterion.

Results Estimated treatment coefficients from the untransformed scale-based methods (Tobit, Median, & CLAD) deviated from MLR, whereas the SESs from Tobit produced almost identical values. Transformed scale-based methods (OL, OP, BB, BLN, Frac, and BR) shared a similar pattern, except that OL generated higher absolute coefficients and BLN produced higher SESs than other methods. The SESs from Tobit, BB, OP, and Frac had better agreement against MLR than other included methods.

Conclusions The SES is a simple method to unify and compare estimates produced from various statistical methods on different scales. As these methods did not produce identical SES values, it is crucial to comprehensively understand and carefully select appropriate statistical methods, especially for analysing PROs like SF-36, to avoid drawing

*Correspondence:
Yirui Qian
yirui.qian@york.ac.uk
Full list of author information is available at the end of the article

wrong estimates and conclusions using clinical trial data. Future research will focus on simulation analysis to compare the estimation accuracy and robustness of these methods.

Keywords Standardised effect size, Statistical methods, SF-36, Patient-reported outcome, Randomised controlled trial

Background

The Short-Form 36 (SF-36) is a widely used patient-reported outcome (PRO) to measure quality-of-life from patients' perspectives in clinical trials. The SF-36 consists of eight domain scores and one health transition item using 36 items on different ordinal categorical scales. The domain scores include physical functioning (PF), role limitation – physical (RP), bodily pain (BP), general health (GH), vitality (VT), social functioning (SF), role limitation – emotional (RE), and mental health (MH) [1].

The original version of the SF-36 was initially released in 1992 [2], with its validity and reliability tested in the subsequent two years [3, 4]. Modifications of the original SF-36 (SF-36v1) include the RAND 36-item [5], a publicly available version with slightly different scoring methods of the original version [1]; the SF-36 version 2 (SF-36v2) [6], an upgraded version with improvements in wording and in different ordinal categorical scales of some items to enhance the internal reliability consistency and reduce ceiling and floor effects [7]; the Short-Form 6-Dimension (SF-6D), a popular preference-based measure producing utility scores for use in the economic evaluation [8]; and other shorter versions that use only 8 or 12 items instead of all 36 items [9].

Two types of scoring mechanisms are commonly seen to produce SF-36 domain scores: the original scoring and the norm-based scoring. The original scoring anchors each scale from 0 (worst score on all items) to 100 (best score on all items). Norm-based scoring linearly rescales the eight domain scores to achieve a mean score of 50 and a standard deviation of 10 in the reference population (i.e. the US general population) [10]. While the described methods apply to both scoring approaches, we use the original, 0 to 100, scoring here for simplicity.

The domain scores generated from the original scoring mechanism tend to be discrete, bounded, and skewed [11]. When analysing these scores with classical statistical methods (such as linear regression or *t*-test), the model assumptions (such as Normality of residuals or heteroscedasticity) are likely to be violated. An inappropriate analysis can result in unreliable estimates, with wider confidence intervals (CIs) or larger standard errors, and accordingly fail to provide accurate and robust results for decision-making [12].

Various statistical methods have been developed for the analysis of SF-36 data, or PRO data in general, and a few comparisons of these methods have been conducted [13–16]. However, some estimated treatment coefficients from different statistical methods may not be comparable in these studies. For example, the estimate of an ordered logit model is interpreted as an odds ratio after inverse log-transformation, whereas the estimate of a linear regression is a mean. The two types of estimates – odds ratio and mean – are not comparable. The standardised effect size (SES), which is calculated by dividing the group difference by the pooled standard deviation, can produce estimates with no units of measurement and therefore allows the comparison of estimates that are based on different scales [17–19]. We introduce the SES summary measure to unify the estimated treatment coefficients from different statistical methods for analysing PROs.

This study aims to apply various statistical methods for the analysis of SF-36 domain scores using data from randomised controlled trials (RCTs) and compare estimates from different methods using the SES summary measure of the treatment effect. In the rest of the paper, we first describe the included statistical methods for analysing SF-36; then produce and interpret the estimated treatment coefficients and SESs from these statistical methods; and finally present the agreement of the SESs from these statistical methods.

Methods

A series of secondary analyses of the SF-36v2 outcome using the original scoring mechanism at a single follow-up timepoint were conducted using data from three RCTs. The scoring strategies for eight domain scores in SF-36v2 are presented in Table 1.

Description of statistical methods

Ten statistical methods included for comparison were multiple linear regression (MLR), median regression (Median), Tobit regression (Tobit), censored absolute least deviation regression (CLAD), beta-binomial regression (BB), binomial-logit-normal regression (BLN), ordered logit model (OL), ordered probit model (OP), fractional logistic regression (Frac) and beta regression (BR). The selection of these ten statistical methods were based on their widespread use, ability to accommodate different distributions for the outcome, applicability to various scenarios, and recommendations from previous studies [14, 20], originating from our previous review that

Table 1 Scoring strategies for eight domains in SF-36 version 2

SF-36 Domain	Abbr	No. items	No. levels	Raw score range	No. possible values after recoding
Physical functioning	**PF**	10	21	10 to 30	21
Role limitation – physical	**RP**	4	17	4 to 20	17
Bodily pain	**BP**	2	27	2 to 11	10
General health	**GH**	5	39	5 to 25	21
Vitality	**VT**	4	17	4 to 20	17
Social functioning	**SF**	2	9	2 to 10	9
Role limitation – emotional	**RE**	3	13	3 to 15	13
Mental health	**MH**	5	21	5 to 25	21

Abbr abbreviation, *SF-36* Short-Form 36

summarised a list of statistical methods for PRO analysis [21]. These methods are described under the generalized linear model (GLM) framework [22] and presented in Table 2. The SF-36 domain score at a specific post-randomisation timepoint in each trial was analysed by these ten statistical methods, comparing the treatment groups and adjusting for the corresponding baseline scores, as the treatment effect is regarded as the main outcome of clinical trials, and a PRO score is likely to correlate with its baseline score [20, 23].

MLR is the most commonly used method for the analysis of PROs with some appealing features: it requires no transformation of the response variable, it produces point estimates that are based on the untransformed scale of measurement and are easy to interpret, and it is a robust method when faced with the violation of model assumptions [24]. Tobit is known to be consistent and efficient under the Normality assumption and homoscedasticity [25, 26]. When these model assumptions are violated, CLAD can be used as a substitute for Tobit [27]. It is worth noting that the use of Tobit or CLAD relies on the premise that a PRO score exceeding the low or/and high boundaries is possible and meaningful [28]. Two ordinal regression methods, OL and OP, assume proportional odds. Compared to OL, OP may be less preferable to use, since the estimated treatment coefficient from an OP cannot be explained in odds ratio as the OL does. BB and BLN are built upon binomial regression, with their probability parameter being a random variable and following beta distribution and logit-Normal distribution respectively [13, 14, 29]. They both can account for the ordinal and discrete feature of the PRO scores without the requirement for distributional assumptions. Fractional regression is suitable to analyse bounded data falling between 0 and 1 [30], such that recoding the SF-36 domain score that are on a [0, 100] scale is needed to apply fractional regression. Frac can account for scores at boundaries of 0 and 100, whereas BR cannot, and it therefore requires the 'squeezing' of the domain scores [16].

Model assumptions were tested where necessary, including the Normality of residuals and the homoscedasticity assumptions for MLR and Tobit, and proportional odds assumptions for ordinal regression. To run statistical methods that require outcome variables on transformed scales, different recoding techniques were used to transform the SF-36 domain scores to appropriate forms [31–33]. Technical details of these methods and the recoding strategies for the SF-36 domain scores are summarised in the Supplementary Material.

Estimated treatment coefficient and standardised effect size

An estimand is a well-defined and explicit description of precisely what treatment effect is to be estimated in an RCT [34]. It consists of five connected elements [35]: the treatments to be compared, the target population, the outcome or endpoint, intercurrent event handling, and a population-level summary measures of how outcomes between the different randomised groups will be compared. Common population-level summary measures include the difference in means, risk ratios and odds ratios.

If the treatment coefficient is chosen as the summary measure to compare outcomes between the different randomised groups, then it may not make sense to compare the ten estimators (i.e. the statistical methods) and their associated estimates as they are different estimands. However, some of the methods, such as MLR, Tobit, CLAD, and Median produce estimates that have similar population summary measures that look at differences in location or central tendency e.g. differences in means, or medians. Therefore, in these models it may be sensible to compare the treatment coefficient estimates of difference

Table 2 Summary of ten statistical methods for the analysis of SF-36 dimension scores under the GLM framework

Statistical methods	Distribution of the outcome/dependent variable (Y)	Link function $g(\bullet)$	Model assumption	Recoding of PRO needed	Interpretation	Estimation method	Stata command
Classical model							
MLR	Continuous	Identity $g(\mu) = \mu = \boldsymbol{X\beta}$	Normality (of residuals); homoscedasticity; linearity; independence of outcomes	No	Mean	OLS or MLE	regress
Median	Continuous	$g\left(Q_{Y\|X}(median)\right) = Q_{Y\|X}(median) = \boldsymbol{X\beta}_{median}$	Linearity; independence of outcomes	No	Median	LAD	qreg
Censored regression							
Tobit	Observed$Y$: Continuous and censored; LatentY^*: Continuous	Identity $g(\mu^*) = \mu^* = \boldsymbol{X\beta}$	Normality (of residuals); homoscedasticity; linearity; independence of outcomes	No	Latent Mean	MLE	tobit
CLAD	Observed$Y$: Continuous and censored; LatentY^*: Continuous	$g\left(Q_{Y^*\|X}(median)\right) = Q_{Y^*\|X}(median) = \boldsymbol{X\beta}_{median}$	Linearity; independence of outcomes	No	Latent Median	CLAD	clad
Ordinal regression							
OL	Ordinal	Logit $g(\theta_{il}) = \ln(\frac{\theta_{il}}{1-\theta_{il}})$	Proportional-odds; linearity; independence of outcomes	Yes (to [0,k−1]$\subseteq$N)	Odds ratio	MLE	ologit
OP	Ordinal	Probit $g(\theta_{il}) = \Phi^{-1}(\theta_{il})$	Proportional odds; linearity; independence of outcomes	Yes (to [0,k−1]$\subseteq$N)	Probability	MLE	oprobit
Binomial regression							
BB	Beta-binomial i.e.$Y_i \sim Bin(k,\theta_i)$ $\theta_i \sim Beta(\boldsymbol{\alpha},\boldsymbol{\gamma})$	Logit $g(\theta_i) = \ln(\frac{\theta_i}{1-\theta_i})$	Linearity; independence of outcomes; beta distribution of probability of success	Yes (to [0,k−1]$\subseteq$N)	Odds ratio	MLE	betabin
BLN	Binomial-logit-Normal i.e.$Y_i \sim Bin(k,\theta_i)$ $\theta_i \sim LN(0,1)$	Logit $g(\theta_i) = \ln(\frac{\theta_i}{1-\theta_i})$	Linearity; independence of outcomes; logit-Normal distribution of probability of success	Yes (to [0,k−1]$\subseteq$N)	Odds ratio	MLE	glm…link (logit) family (binomial N)
Fractional regression							
Frac (logit link)	Recoded$Y\prime$: continuous and bounded in [0, 1]	Logit $g(\mu_{Y\prime}) = \ln(\frac{\mu_{Y\prime}}{1-\mu_{Y\prime}})$	Linearity; independence of outcomes	Yes (to [0, 1] scale)	Odds ratio	Quasi-likelihood estimation	fracreg logit
BR (logit link)	Recoded$Y\prime\prime$: continuous and bounded in (0, 1) $Y\prime\prime \sim Beta(\mu\varphi,(1-\mu)\varphi)$	Logit $g(\mu_{Y\prime\prime}) = \ln(\frac{\mu_{Y\prime}}{1-\mu_{Y\prime\prime}})$	Linearity; independence of outcomes	Yes (to (0, 1) scale)	Odds ratio	MLE	betareg

BB Beta-binomial regression, *BLN* binomial-logit-Normal regression, *CI* confidence interval, *CLAD* censored least absolute deviations, *Frac* fractional logistic regression, *GLM* generalized linear model, *LAD* least absolute deviations, *Median* median regression, *MCAR* missing completely at random, *MCDA* multi-criteria decision analysis, *MLE* maximum likelihood estimation, *MLR* multiple linear regression, *SF-36* Short-Form 36, *Tobit* Tobit regression, *OL* ordered logit model, *OLS* ordinary least squared, *OP* ordered probit model, *PRO* patient-reported outcomes.k, represents the number of possible categorical values in a domain;μdenotes the mean;μ^* denotes the latent mean; N, denotes the non-zero positive natural numbers i.e. 1,2,3…k−1;$\varphi$is the precision parameter for beta distribution;Φstands for the standard Normal cumulative distribution function.θ, denotes the probability of success or the cumulative response probabilities i.e.$\theta_{il} = P(Y \leq l)$the probability of a response in category l or below for OL, and$\theta_l = P(Y = l)$the probability of a response in category l for BB, BLN, and BR. Note that clad and betabin are user-developed packages in Stata, and therefore installation of the corresponding package is required to run these two commands

in means or medians between two treatment groups produced by these models as they are similar estimands. Again, some of the methods, such as OL, BB, BLN, Frac, and BR, also have similar population summary measures that estimate the (log) odds ratio for the treatment effect. Therefore, in these models it may be sensible to compare the estimates of (log) odd ratios between two treatment groups produced by these models, as they are similar estimands.

The estimates of MLR, Tobit, CLAD and Median are generated using SF-36 based on the untransformed scale. As the logit link is used for OL, BB, BLN, BR, and Frac, their estimated treatment coefficients can be interpreted as odds ratio through the exponential transformation, where $coef(TE)$ denotes the estimated values for the treatment effect parameter.

$$OddsRatio_{TE} = \exp(coef(TE))$$

A special case of these methods is the OP, which uses a probit link. The probit is the inverse of the cumulative standard Normal distribution that is denoted by Φ. The estimates of an OP cannot be transformed to odds ratios as other methods do, but it can be interpreted as the probability or the effect size in the response.

Under the estimand framework, we can compare the estimates of the treatment difference between the treatment groups from the different statistical methods, using the SES as the population level summary measure, as the other four attributes for the estimand (the treatment, the target population, the outcome, and intercurrent event handling) are the same regardless of the statistical methods used. The ten statistical methods that we compared in this study all fit under the GLM framework, and their Z-statistics are calculated using the same formula, i.e. the point estimate, of the treatment effect parameter, divided by its standard error. Therefore, in these circumstances, the estimand, and its population summary measure, the SES, is the same, but the estimators (e.g. the ten statistical models) will be different and may produce different estimates that can be compared and contrasted.

After producing estimated treatment coefficients from the different statistical methods, their scale-invariant SES and its associated standard error were calculated using the following formula [18, 36]:

$$SES = Z \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = \frac{coef(TE)}{SE(TE)} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$SE(SES) = \sqrt{\frac{n_1 + n_2}{n_1 n_2} + \frac{SES^2}{2(n_1 + n_2)}}$$

where Z stands for the Z-statistics; TE stands for the treatment effect; $coef(TE)$ stands for the estimated values for the treatment effect parameter; $SE(TE)$ stands for the standard error of the treatment effect estimates; $SE(SES)$ represents the standard error of the SES; and n_1 and n_2 represents the sample size in each treatment group respectively.

The Akaike information criterion (AIC) [37] values were produced when fitting different statistical methods to compare the model fit, using the following equation:

$$AIC = 2c - 2\ln(\widehat{L})$$

where $\widehat{L}$ is the maximum likelihood for the model, and c is the number of estimated parameters.

A lower AIC value represents a better model fit. Note that the value c for different statistical methods with the same set of independent variables can be different, and the AIC values cannot be calculated for CLAD and Median because they both are quantile regression methods which use (censored) least absolute deviation for estimation rather than maximum likelihood.

The characteristics of the three trials included in this study were summarised in a table by the basic trial information on the study population, primary outcome, follow-up timepoints, samples size at the baseline, and number of patients analysed. Scatter plots were generated to compare the estimated treatment coefficients from different methods, using consistent markers for each method and consistent colours for each trial. Estimated treatment coefficients from the MLR and BB were used as the reference benchmark for statistical methods on the untransformed and transformed scales respectively, since MLR is the most commonly used methods for analysing PROs [38], and BB was reported to render satisfactory results in various situations for PRO analysis [14]. The SESs from different methods were displayed in scatter plots, using MLR as the reference benchmark for all included methods regardless of their scales, as the SES is believed comparable among statistical methods on different scales under the estimand framework. Effect size plots were graphed for the SESs with its associated CIs of estimated SES from ten statistical methods, together with two horizontal lines representing the clinical and statistical significance. The change of AICs against a number of possible categorical values (i.e. levels) of SF-36 domain scores was graphed using scatter plots.

We used the statistical package Stata/MP 17.0 for statistical analysis and data visualisation. Stata codes for applying the recoding techniques and conducting regression analyses are available in the Supplementary Material.

Results

Description of datasets

The three RCTs included in this study are Chronic Obstructive Pulmonary Disease (COPD) [39], Lifestyle Matters (LM) [40] and Putting Life IN Years (PLINY) [41] (Table 3). The follow-up timepoints for analysis were chosen as 2 months (acute phase endpoint) for COPD, and 6 months (primary endpoint) for both LM and PLINY. Each statistical method produced 24 estimates of treatment effect for eight SF-36 domain scores across the three RCTs. A total number of 684 patients were randomised in the three trials combined, and 492 patients that have SF-36 outcome data at follow-up were analysed.

In general, skewed distributions of residuals and heteroscedasticity are shown in SF-36 domain scores after the post estimation of MLR and Tobit, and 20.8% of the model outcomes using ordinal regression violated the proportional odds assumption. A summary of post-estimation plots for MLR and Tobit are available in the Supplementary Material (Figure S1-S4).

Estimated treatment coefficients and interpretation

Figure 1 and Fig. 2 show the scatter plots of estimated treatment coefficients from the untransformed scale-based methods against MLR and the transformed scale-based methods against BB respectively. Estimates from statistical methods on the untransformed scale (i.e. Tobit, CLAD, and Median) deviated from MLR. Especially when the magnitude of the estimated treatment effects was large, they tended to produce higher estimates than MLR. CLAD and Median shared a similar pattern when the estimates were small, i.e. they tended to produce estimates scattering at zero. Estimates from methods on the transformed scales were presented using odds ratios, except for OP. The odds ratios estimated from BLN and Frac are shown similar to BB. The OL produced higher absolute estimates than other methods, and the exponentiation procedure to generate odds ratios magnified this trend. This trend is obvious in PLINY that presented averagely higher treatment estimates than other trials.

An example of how to interpret the estimated treatment coefficients from the included ten statistical methods is presented in Table 4, based on the SF-36 MH score at 6-month follow-up in the LM trial.

Comparison of estimated standardised effect sizes

Overall, the estimated SESs in our datasets were small (i.e. absolute values less than 0.2), except for the SESs of some domains in PLINY (i.e. absolute values between 0.5 and 1.4). CLAD failed to converge on one occasion. When estimating the treatment coefficient of the same response using different methods, SESs with different directions were produced from ten statistical methods. However, there was no case where these statistical methods produced statistically significant estimates with different directions. For SESs with the same direction, different statistical significance and magnitudes of effect size are observed in these analyses.

Figure 3 presents the SES estimated from ten different statistical methods against MLR. For statistical methods that used the untransformed scale of measurement, Tobit has almost identical pattern against MLR, whereas CLAD and Median are more scattered. For methods that used transformed scales, the OL that produced higher estimated coefficients (or odds ratio) showed similar SESs as other methods after standardisation. Conversely, although BLN produced similar estimated coefficients (or odds ratios) as BB, the SESs from BLN was larger than other methods after standardisation. It also shows that the Tobit, BB, OP, and Frac had stronger agreement with MLR across the three included trials, i.e. the difference between each of these four methods against MLR is associated with less bias and narrower 95% CIs than rest of the methods.

Figure 4 shows the SES with its associated 95% CIs of treatment estimates from different statistical methods in two trials (PLINY and LM) that used SF-36 MH score at 6-month follow-up as primary outcome. Two horizontal lines are drawn in the plot, representing the SES having no effect or no difference between two treatment groups (i.e. $y=0$) and clinical significance (i.e. $y=$ minimal clinical important difference/standard deviation $=0.4$ in both trials). When analysing the treatment effect of the same outcome, the use of different statistical methods may draw different results in terms of statistical significance and/or clinical significance. For example, SES from Median is statistically significant in LM, whereas this is not the case for MLR. In PLINY, most methods showed statistically significant estimates except for MLR, CLAD and Median. Effect size plots for other SF-36 domain scores used as secondary outcomes in these three trials are available in Supplementary Material Figure S5.

Figure 5 shows how AIC changed against a different number of possible categorical values (i.e. levels) of domain scores when applying the ten statistical methods in three RCTs. As the comparison of AICs require the methods to model the same response variable [37], these methods were categorised into four groups according to their distributional assumptions and recoding techniques on SF-36 domain scores. When fitting higher levels of SF-36 domain scores, the AICs of Tobit, ordinal, and binomial regression became larger, representing a poorer

Table 3 Characteristics of the three RCTs that used SF-36 domain scores as clinical outcomes

Trial name	Trial population	Primary outcome	Follow-up timepoints (months)	Sample size at baseline			Max sample size analysed (PRO)			Ref
				Total	Control	Treat	Total	Control	Treat	
COPD	Patients with chronic obstructive pulmonary disease	Difference in improvement in endurance shuttle walking test (ESWT) during 18 months follow-up	**2**, 6, 12, 18	239	129	110	174	93	81	[38]
LM[a]	Independently living older people (aged 65 or more)	SF-36v2 MH at 6 months follow-up	**6**, 24	288	143	145	262	126	136	[39]
PLINY[a]	Independently living older people (aged 75 or more)	SF-36v2 MH at 6 months follow-up	**6**	157	79	78	56	30	26	[40]
Total				*684*	*351*	*333*	*492*	*249*	*243*	

[a] Trials using SF-36 domains as primary outcomes. PLINY and LM used SF-36v2 MH at 6 months as primary clinical outcomes

MH mental health, *PRO* patient-reported outcome, *SF-36v2* Short-Form 36 version 2

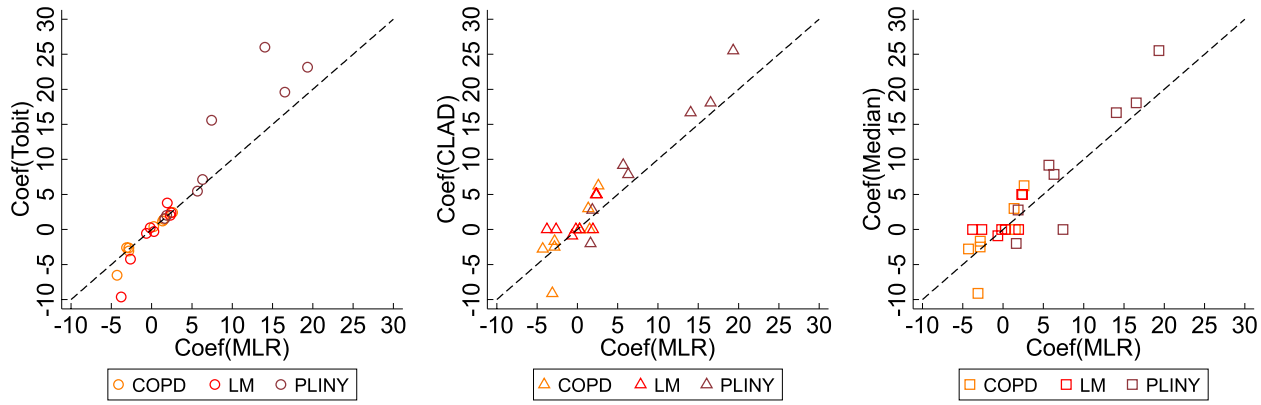

Fig. 1 Estimated treatment coefficients from untransformed scale-based statistical methods against MLR. Coef, treatment coefficient; CLAD, censored absolute least deviation regression; Median, median regression; MLR, multiple linear regression; Tobit, Tobit regression

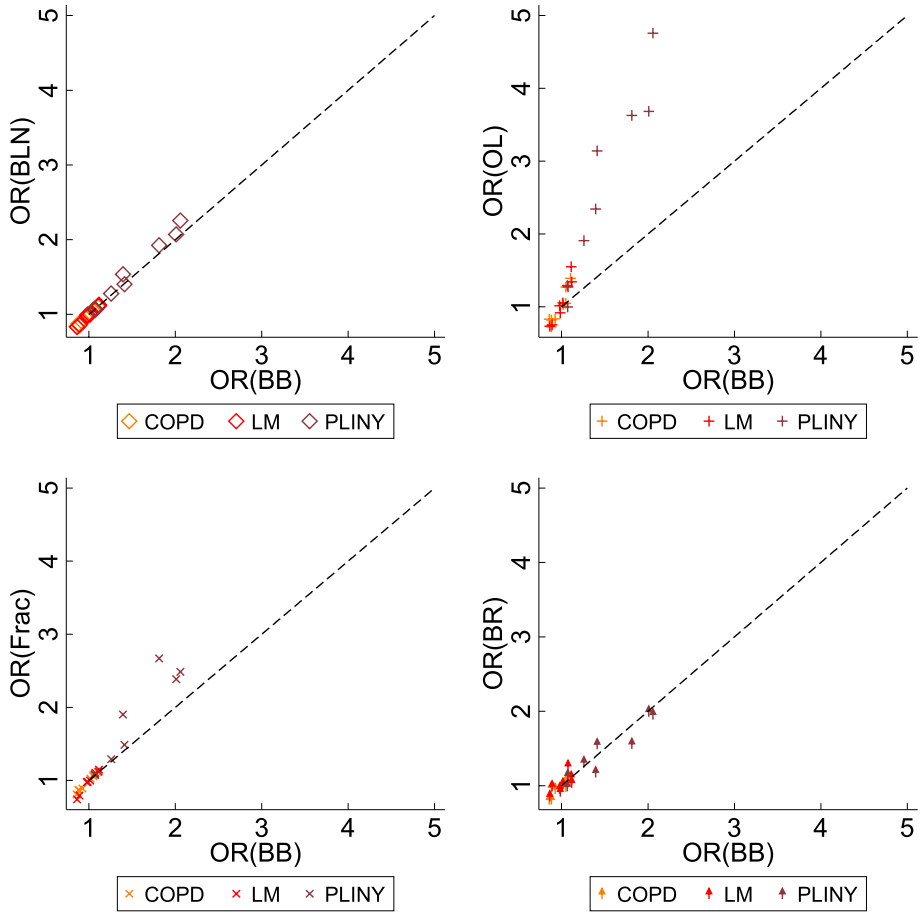

Fig. 2 Estimated ORs from transformed scale-based statistical methods against BB. BB, beta-binomial regression; BLN, binomial-logit-normal regression; BR, beta regression; Frac, fractional logistic regression; OL, ordered logit model; OR, odds ratio

fit, whereas the AICs for the MLR became smaller, representing a better fit. The scatter plot shows that AIC values for BR were not sensitive to the change of possible categorical values of domain scores.

Discussion

This study applied ten statistical methods for the analysis of PROs to three RCTs using SF-36. A total of 240 estimates of treatment coefficients for SF-36 eight domain

Table 4 Interpretation of the treatment coefficient from the ten statistical methods using the SF-36 MH score at 6-month follow-up in the LM trial as an example

(a) Statistical methods that used the untransformed scale

Statistical methods	Coefficient	Interpretation
MLR	2.31	The mean of the MH score at 6 months for the treatment group is 2.31 points higher than the mean for the control group, after adjusting for baseline MH score
Tobit	2.02	The mean of the *uncensored* MH score at 6 months for the treatment group is 2.02 points higher than the mean for the control group, after adjusting for baseline MH score
Median	5.00	The median of the MH score at 6 months for the treatment group is 5.00 points higher than the median for the control group, after adjusting for baseline MH score
CLAD	5.00	The median of the *uncensored* MH score at 6 months for the treatment group is 5.00 points higher than the median for the control group, after adjusting for baseline MH score

(b) Statistical methods that used the transformed scales

Statistical methods	Coefficient	Odds	Interpretation
BB	0.11	1.12	The odds of the MH score being in a given category at 6 months in the treatment group is 1.12 times that of the odds for the control group, after adjusting for baseline MH score
BLN	0.12	1.13	The odds of the MH score being in a given category at 6 months in the treatment group is 1.13 times that of the odds for the control group, after adjusting for baseline MH score
OL	0.29	1.34	The odds of the MH score being in a given category **or less** at 6 months in the treatment group is 1.34 times that of the odds for the control group, after adjusting for baseline MH score
OP	0.16	NA	*(Marginal effects need to be calculated to generate the probability)* The probability of scoring 90.0 at 6 months is 21.2% for the treatment group and 20.1% for the usual care group, after adjusting for baseline MH score
Frac	0.14	1.15	The odds of the MH score being in a given category at 6 months in the treatment group is 1.15 times that of the odds for the control group, after adjusting for baseline MH score
BR	0.04	1.05	The odds of the MH score being in a given category at 6 months in the treatment group is 1.05 times that of the odds for the control group, after adjusting for baseline MH score

BB beta-binomial regression, *BLN* binomial-logit-normal regression, *BR* beta regression, *CLAD* censored absolute least deviation regression, *Frac* fractional logistic regression, *OL* ordered logit model, *OP* ordered probit, *Median* median regression, MH mental health, *MLR* multiple linear regression, *NA* not applicable, *SF-36* Short-Form 36, *Tobit* Tobit regression

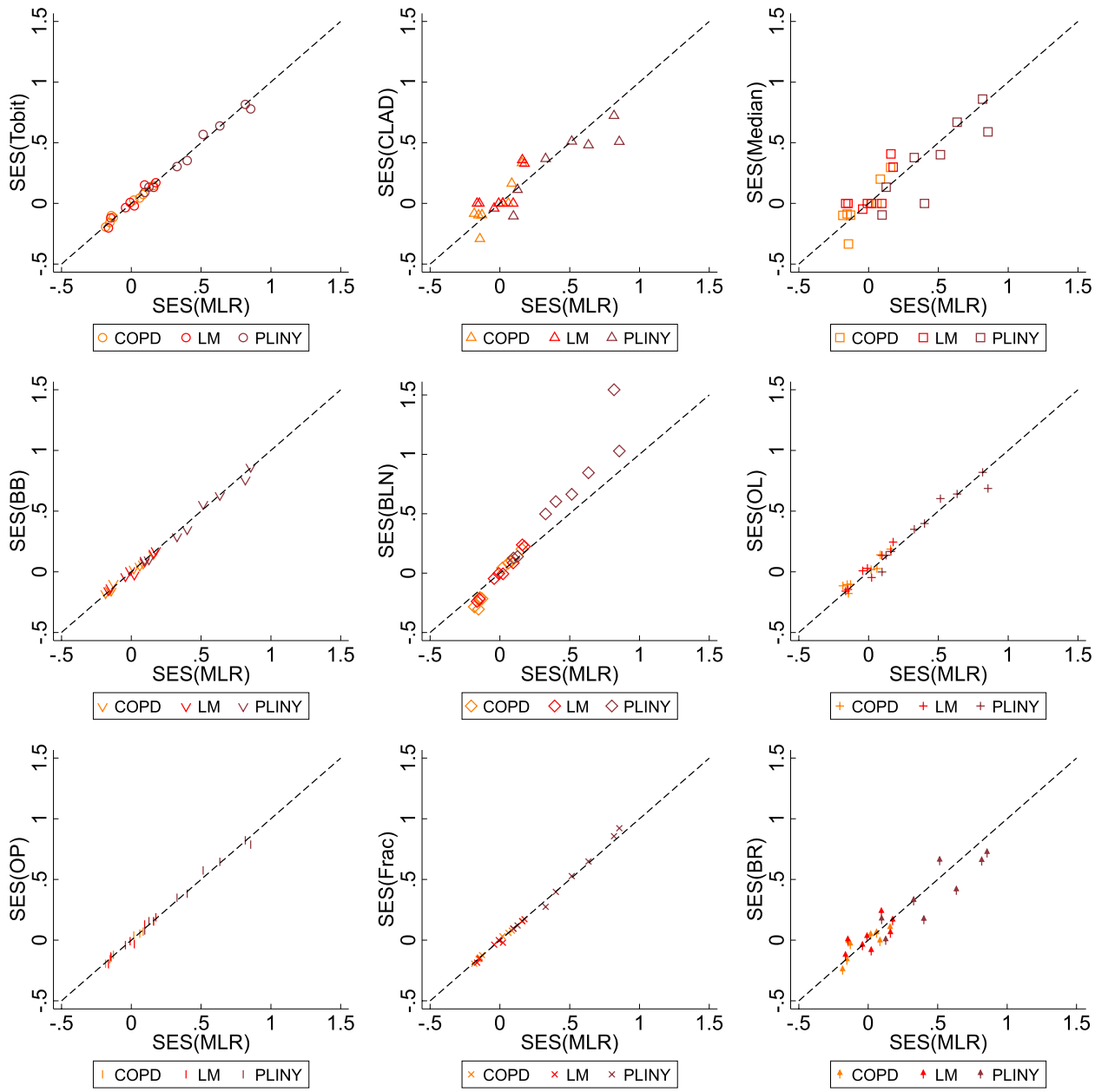

Fig. 3 Estimated SES from different statistical methods against MLR. The x-axis of the scatter plots is the SES estimated by MLR, and the y-axis is the SES estimated by other statistical methods. The black dash line represents the method that produces the same SES as MLR. BB, beta-binomial regression; BLN, binomial-logit-normal regression; BR, beta regression; CLAD, censored least absolute deviation regression; Frac, fractional logistic regression; Median, median regression; MLR, multiple linear regression; OL, ordered logit model, OP, ordered probit model; Tobit, Tobit regression; SES, standardised effect size

scores at a single follow-up timepoint were calculated across three RCTs using MLR, Median, Tobit, CLAD, BB, BLN, OL, OP, Frac, and BR. Their method theory under the GLM framework and interpretation for estimates generated from these methods were explained and presented. The SES was applied to compare the magnitude of estimated treatment coefficients from these methods based on different scales. The AIC statistics were calculated to present the change of model fit against a different number of possible values in SF-36 domain scores.

Our empirical analysis shows that SESs estimated from different methods are generally consistent, using MLR as the reference benchmark, although the estimated treatment coefficients by different methods vary. For

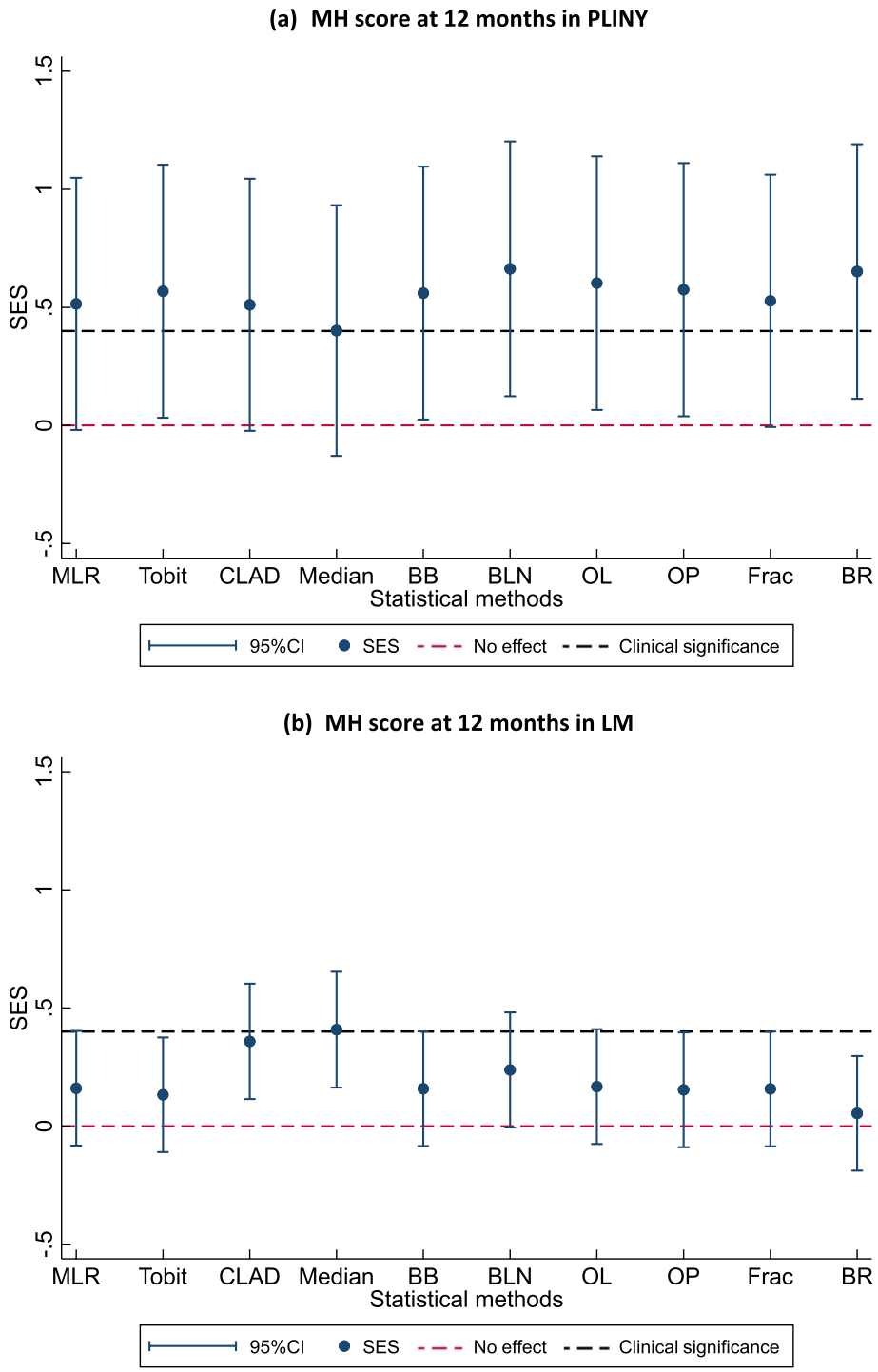

Fig. 4 SES with associated 95% CIs from ten statistical methods BB, beta-binomial regression; BLN, binomial-logit-normal regression; BR, beta regression; CLAD, censored least absolute deviation regression; Frac, fractional logistic regression; Median, median regression; MLR, multiple linear regression; OL, ordered logit model, OP, ordered probit model; Tobit, Tobit regression; SE, standard error; SES, standardised effect size. The bars on two sides of the vertical line for each method represents the 95% confidence intervals (CIs) for SES, which are calculated using $SES \pm 1.96 \times SE(SES)$, where $SE(SES)$ represents the standard error of the SES

example, the magnitude of estimated treatment coefficients by Tobit is larger than MLR, but the estimated SESs by Tobit and MLR are almost identical. This may result from the fact that the Tobit accounts for the censoring and scatters the observed variable beyond the boundaries, the latent variable of which is assumed to have a wider scale than the observed variable that is used for estimation by MLR. However, adjusting the standard deviation of treatment estimates offsets the large magnitude of the estimated coefficients, and thus results in the agreement in SES between Tobit and MLR. The CLAD, the other method applied in this study that can account for boundedness of the outcome, was found to be inefficient compared to other methods, as it took longer to run in Stata and failed to converge on one occasion. For statistical methods that used the untransformed scale, quantile regressions (Median and CLAD) show more variation. This may be because they adopt a different estimation method (i.e. least absolute deviation) in comparison to those that use maximum likelihood estimation. As was found in another study comparing Tobit, Median, and CLAD using Health Utility Index (HUI) [27], Median and CLAD tend to produce estimates with similar patterns and their estimates tend to be shrunk to zero compared to MLR and Tobit.

It is possible for different combinations of treatment coefficient and standard error estimates to produce the same effect size on the standardised scale [17]. An example for the transformed scale-based methods is the OL, which generally produced higher estimates than other methods, resulting in higher estimated odds ratios. However, the standard errors estimated from OL were also higher than other methods, offsetting the high values of estimates when calculating the SES values. Conversely, BLN produced slightly higher SES estimates, whereas the treatment coefficients from it were similar to other methods. This may result from the fact that BLN assumes a Normal distribution for random effects, but this assumption may not be valid for some SF-36 domain scores [14].

In comparison to Frac, the SES from BR biases slightly from the reference benchmark, MLR. This may be caused by the required 'squeezing' procedure in BR, which can reduce the estimation precision [33]. Fractional regression methods are more suitable for the analysis of health utility scores, compared to other included methods. But it is noteworthy that, in scenarios where health utilities index scatter on slightly different scales, e.g. SF-6D scattering between 0.291 and 1 for the United Kingdom value set [8], application of these two fractional regression methods may not be straightforward [33, 42].

Generally, when increasing the number of possible categorical values, the AIC for statistical methods with logit or probit link increased; it decreased, however, for MLR. Interestingly, Tobit, an extension of MLR designed to adapt for censored outcomes, generated lower AIC values (i.e. better model fit) when analysing outcomes with a small number of possible categorical values. This shows an adverse trend compared to MLR and requires further investigation.

To achieve the aim of comparing the estimates by different statistical methods that are based on different scales, we have adapted the SES as the population summary measure of the treatment effect in the estimand framework, which is not as frequently seen as other population summary measures of the treatment effect such as means, risks, or odds ratios [43]. In practice, the concept of SES has been applied in various scenarios in trials using PROs and their related studies. This includes summary studies such as meta-analysis in literature reviews that compare PROs that are based on different scales, sample size calculation in trial designs that use PRO as primary outcomes, and trials with PROs that used the effect size as the measurement of treatment effectiveness [44–48].

For linear models, the effect size is a ratio of estimated coefficient over standard deviation of the estimate. The standardisation procedure is completed by the Z statistics, adjusting for sample size. Therefore, when estimating the same treatment effect using different methods, the SES assesses the statistical power of these methods for a given sample size. For instance, if the data is ordinal and the model assumption of OL is satisfied, the OL is likely to have higher power than other statistical methods, and thus be the most appropriate method for analysing the data. In theory, the most appropriate method is more likely to capture the 'truth' than other methods. However, as the domain scores of SF-36 have different categories and distribution patterns, it is therefore difficult to assign them to a certain type of distribution and to decide what statistical methods to use for analysis.

(See figure on next page.)

Fig. 5 Scatter plot of AIC for different statistical methods against the number of possible observable values of SF-36 domains in three RCT datasets. Note that median and CLAD regression do not have AIC scores, and thus are not compared in this figure. These methods are classified according to their distributional assumptions on SF-36 domain scores. AIC, Akaike information criterion; MLR, multiple linear regression; Tobit, Tobit regression; OL, ordered logit model, OP, ordered probit model, BB, beta-binomial regression; BLN, binomial-logit-normal regression; frac, fractional logistic regression; BR, beta regression

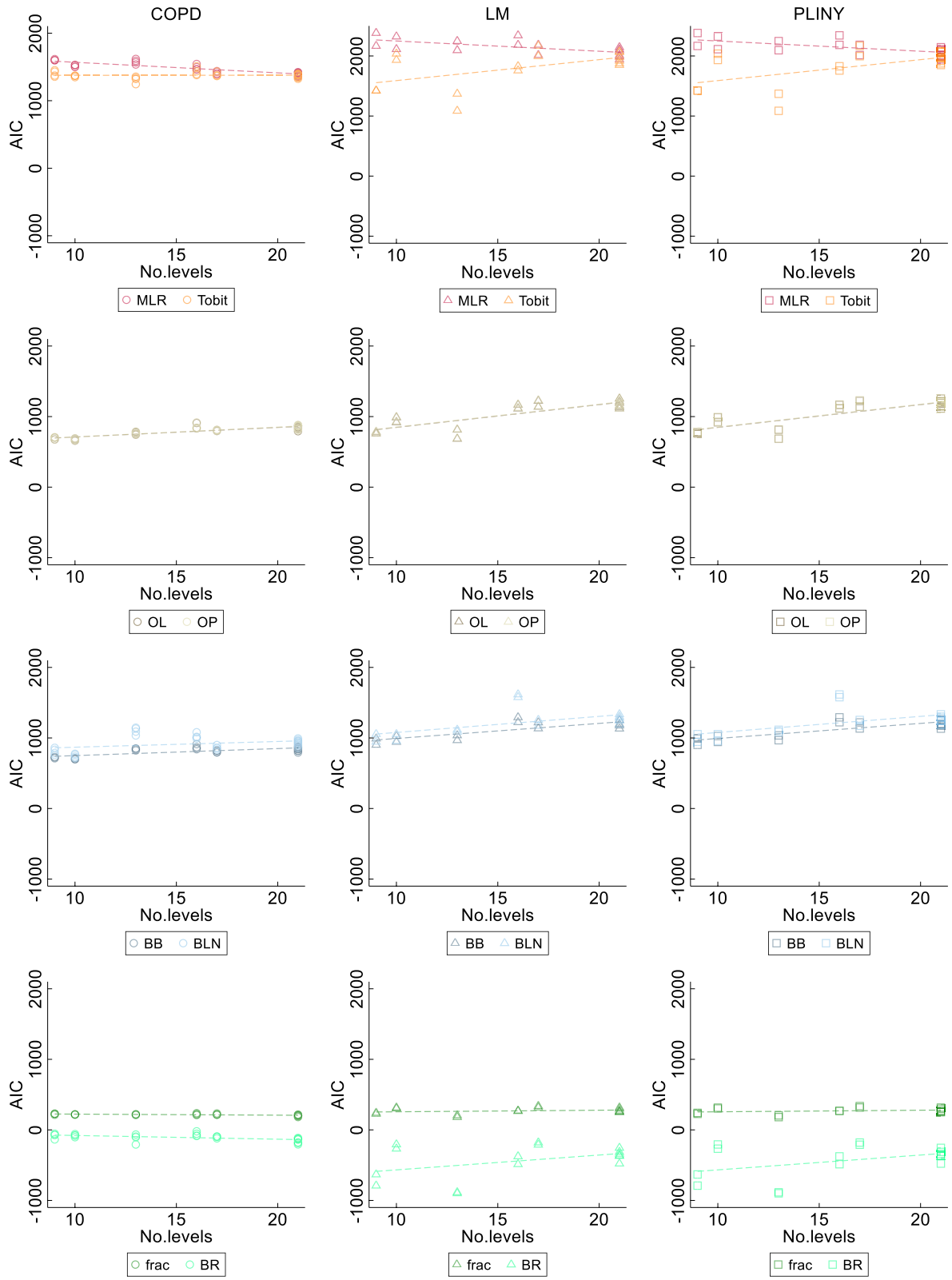

Fig. 5 (See legend on previous page.)

Violation of model assumptions may suggest that an inappropriate statistical method is used, which may lead to inaccurate standard errors, misleading statistical significance, unreliable treatment effect estimates, and even alter the results of decision-making [49]. For example, patients may fail to receive an effective treatment because this treatment is falsely shown not to be clinically effective based on inaccurate estimates; vice-versa, patients can receive a treatment which may potentially harm their health when unreliable evidence supports the use of this treatment. Therefore, applying appropriate statistical methods for the analysis of PROs in trials is crucial to reduce biases of estimates, to accurately evaluate clinical effectiveness and to support healthcare decision-making.

It is important to understand different statistical methods before selecting and applying them [50]. Our comparison of SES estimates with its associated 95% CIs suggested that the choice of statistical methods for data analysis might result in different conclusions drawn from the hypothesis tests in terms of the clinical and statistical significance. Other factors that need to be considered for the selection of a method for the analysis of PROs include the aim of the study, the statistical features of the PRO scores, the complexity in the interpretation of a model coefficient, the computing time to run a method, and the software programs and packages availability to apply a method [20, 51].

This study has the following limitations: First, we included three trials that focused on different disease areas and populations, which can be seen as a source of heterogeneity. However, this study does not intend to compare the size of treatment estimates across different trials but to compare whether different statistical methods can produce similar estimates using the SES approach. Therefore, the results of this study should not be influenced by the magnitude of effect sizes and heterogeneity in trials.

Second, this study focused on domain scores in SF-36v2, and extrapolation to other versions of SF-36 and other types of PROs may require further validation. However, the SF-36 is a widely used generic PRO and it may be plausible to extrapolate the results to other PROs that share similar data features (discrete, bounded, and skewed) of SF-36, such as the Beck Depression Inventory (BDI), Hospital Anxiety and Depression Scale (HADS), European Organization for the Research and Treatment of Cancer Quality of Life Questionnaire (EORTC QLQ-C30), and potentially preference-based PROs such as SF-6D, and EuroQol-5 Dimensions (EQ-5D).

Third, as the aim of this study is to compare different statistical methods but not to identify the best model, we kept the analysis models in simple and similar forms, i.e. we only adjusted for the treatment group and the corresponding baseline score. Other potential effects such as time and clustering were not considered. However, the majority of the statistical methods included in this study are under the GLM framework and can be extended for longitudinal analysis by using generalized linear mixed models with coefficients estimated by maximum likelihood estimation or GLM with coefficient estimated by generalized estimating equations [43].

Fourth, our empirical analysis is based on the real case data such that the 'truth' of the treatment effect is unknown [52, 53], so we are not able to evaluate which statistical methods have less bias than other methods using results from this empirical analysis. Using empirical data from three RCTs in this study can show how robust the methods could be when applied to real case data. However, it still needs further investigation on how close the estimates produced by these methods are to the predefined 'truth', and which method remains robust when analysing different domain scores of the SF-36 and when model assumptions are violated. Our future research will focus on computational simulation to evaluate these statistical methods in terms of the accuracy of estimations and model robustness in different scenarios.

Conclusion

In this study, estimates of treatment effect from ten statistical methods are generated and compared for the analysis of PROs in RCTs. It shows the possibility to use the SES summary measure to unify and therefore allow the comparison of estimates from various statistical methods on different scales. It is worth highlighting that these methods did not produce identical SES values, indicating the selection of inappropriate statistical methods may result in wrong estimates and conclusions when analysing clinical trial data. It is, therefore, crucial to comprehensively understand and carefully select appropriate statistical methods, especially for analysing SF-36 type data that tend to be discrete, bounded, and skewed. Future research involves using simulation methods to compare the accuracy and robustness of these methods to analyse PROs in various scenarios.

Abbreviations

BB	beta-binomial regression
BDI	Beck Depression Inventory
BLN	binomial-logit-normal regression
BP	bodily pain
BR	beta regression
CLAD	censored absolute least deviation regression
Coef	treatment coefficient
EORTC QLQ-C30	European Organization for the Research and Treatment of Cancer Quality of Life Questionnaire
EQ-5D	EuroQol-5 Dimensions
Frac	fractional logistic regression
GH	general health
GLM	generalized linear model
HADS	Hospital Anxiety and Depression Scale

HUI	Health Utility Index
LAD	least absolute deviation
Median	median regression
MH	mental health
MLE	maximum likelihood estimation
MLR	multiple linear regression
OL	ordered logit model
OLS	ordinary least squared
OP	ordered probit model
OR	odds ratio
PF	physical functioning
PRO	patient-reported outcome
RCT	randomised controlled trials
RE	role limitation – emotional
RP	role limitation – physical
SES	standardised effect size
SF	social functioning
SF-36	Short-Form 36
SF-6D	Short-Form 6-Dimension
Tobit	Tobit regression
VT	vitality

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12955-025-02373-z.

Additional file 1.

Acknowledgements
Not applicable.

Authors' contributions
All authors contributed to the study conception and design. Summary of statistical methods, data analysis, and interpretation were performed by YQ, SW, and RJ. The first draft of the manuscript was written by YQ. This work was supervised and supported with technical help by SW, RJ, and LF. All authors critically revised the manuscript and approved the final manuscript.

Funding
Part of this work was completed during YQ's PhD study which was jointly sponsored by the University of Sheffield and the China Scholarship Council (grant number 201908890049). SJW, RMJ, and LF received funding across various projects from the National Institute for Health and Care Research (NIHR). SJW was an NIHR Senior Investigator supported by the NIHR (NF-SI-0617-10012) for this research project. The views expressed in this publication are those of the authors and not necessarily those of the China Scholarship Council, NIHR, NHS or the UK Department of Health and Social Care. These organisations had no role in the study design; in the collection, analysis and interpretation of the data; in the writing of the report; or in the decision to submit the paper for publication.

Data availability
The datasets generated and/or analysed during the current study are not publicly available but may be available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate
This study uses non-personal robustly anonymised, existing data from which the original participants cannot be identified. The data has been previously obtained from RCTs run with Sheffield Centre of Health and Related Research (SCHARR), University of Sheffield, where ethics approvals were received from the appropriate NHS Research Ethics Committee and individual informed consent was obtained to take part in the original trial. The ethics approval for this secondary analysis was obtained from the University Research Ethics Committee, University of Sheffield (Reference Number 036168). We confirm that all methods were performed in accordance with the relevant guidelines and regulations.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Centre for Health Economics, University of York, York, UK. [2]Division of Population Health, School of Medicine & Population Health, University of Sheffield, Sheffield, UK.

Received: 22 January 2025 Accepted: 13 April 2025
Published online: 29 April 2025

References

1. Ware JE, Kosinski M, Gandek B. The SF-36 health survey: manual and interpretation guide. Boston: The Health Institute, New England Medical Center; 1993.
2. Ware JE, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care. 1992;30:473–83.
3. McHorney CA, Ware JE, Raczek AE. The MOS 36-item short-form health survey (Sf-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. Med Care. 1993;31:247–63.
4. McHorney CA, Ware JE, Rachel Lu JF, Sherbourne CD. The MOS 36-item short-form health survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. Med Care. 1994;32:40–66.
5. Hays RD, Sherbourne CD, Mazel RM. The rand 36-item health survey 1.0. Health Econ. 1993;2:217–27.
6. Ware JE. SF-36 Health Survey update. Spine (Phila Pa 1976). 2000;25:3130–9.
7. Jenkinson C, Stewart-Brown S, Petersen S, Paice C. Assessment of the SF-36 version 2 in the United Kingdom. J Epidemiol Commun Health. 1978;1999(53):46–50.
8. Brazier J, Roberts J, Deverill M. The estimation of a preference-based measure of health from the SF-36. J Health Econ. 2002;21:271–92.
9. Laucis NC, Hays RD, Bhattacharyya T. Scoring the SF-36 in Orthopaedics: A Brief Guide. J Bone Joint Surg. 2015;97:1628–34.
10. Maruish ME. User's manual for the SF-36v2 Health Survey. 3rd ed. Lincoln: QualityMetric Inc.; 2011.
11. Walters SJ, Campbell MJ. The use of bootstrap methods for analysing health-related quality of life outcomes (particularly the SF-36). Health Qual Life Outcomes. 2004;2:1–9.
12. Pe M, Dorme L, Coens C, Basch E, Calvert M, Campbell A, et al. Statistical analysis of patient-reported outcome data in randomised controlled trials of locally advanced and metastatic breast cancer: a systematic review. Lancet Oncol. 2018;19:e459–69.
13. Arostegui I, Núñez-Antón V, Quintana JM. Analysis of the short form-36 (SF-36): the beta-binomial distribution approach. Stat Med. 2007;26:1318–42.
14. Arostegui I, Núñez-Antón V, Quintana JM. Statistical approaches to analyse patient-reported outcomes as response variables: An application to health-related quality of life. Stat Methods Med Res. 2012;21:189–214.
15. Pullenayegum EM, Tarride J-E, Xie F, O'Reilly D. Calculating Utility Decrements Associated With an Adverse Event: Marginal Tobit and CLAD Coefficients Should Be Used With Caution. Med Decis Making. 2011;31:790–9.
16. Smithson M, Verkuilen J. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. Psychol Methods. 2006;11:54–71.
17. Cook JA, Hislop J, Adewuyi TE, Harrild K, Altman DG, Ramsay CR, et al. Assessing methods to specify the target difference for a randomised controlled trial: DELTA (Difference ELicitation in TriAls) review. Health Technol Assess. 2014;18:1–175.
18. Rothwell JC, Julious SA, Cooper CL. A study of target effect sizes in randomised controlled trials published in the Health Technology Assessment journal. Trials. 2018;19:1–3.

19. Cohen J. Statistical Power Analysis for the Behavioral Sciences. Statistical Power Analysis for the Behavioral Sciences: Routledge; 2013.
20. Coens C, Pe M, Dueck AC, Sloan J, Basch E, Calvert M, et al. International standards for the analysis of quality-of-life and patient-reported outcome endpoints in cancer randomised controlled trials: recommendations of the SISAQOL Consortium. Lancet Oncol. 2020;21:e83-96.
21. Qian Y, Walters SJ, Jacques R, Flight L. Comprehensive review of statistical methods for analysing patient-reported outcomes (PROs) used as primary outcomes in randomised controlled trials (RCTs) published by the UK's Health Technology Assessment (HTA) journal (1997–2020). BMJ Open. 2021;11:e051673.
22. Nelder JA, Wedderburn RWM. Generalized Linear Models. J R Stat Soc Ser A. 1972;135:370.
23. Vickers AJ, Altman DG. Statistics notes: analysing controlled trials with baseline and follow up measurements. BMJ. 2001;323:1123–4.
24. Lumley T, Diehr P, Emerson S, Chen L. The importance of the normality assumption in large public health data sets. Annu Rev Public Health. 2002;23:151–69.
25. Wilhelm MO. Practical considerations for choosing between tobit and SCLS or CLAD estimators for censored regression models with an application to charitable giving*. Oxf Bull Econ Stat. 2008;70:559–82.
26. Austin PC, Escobar M, Kopec JA. The use of the Tobit model for analyzing measures of health status. Qual Life Res. 2000;9:901–10.
27. Austin PC. A comparison of methods for analyzing health-related quality-of-life measures. Value Health. 2002;5:329–37.
28. Sullivan PW. Are Utilities Bounded at 1.0? Implications for Statistical Analysis and Scale Development. Med Decis Making. 2011;31:787–9.
29. Liang Y, Sun D, He C, Schootman M. Modeling bounded outcome scores using the binomial-logit-normal distribution. Chil J Stat. 2014;5:3–14.
30. Papke LE. Econometric methods for fractional response variables with an application to 401 (k) plan participation rates. J Appl Economet. 1996;11:619–32.
31. Arostegui I, Núñez-Antón V, Quintana JM. On the recoding of continuous and bounded indexes to a binomial form: an application to quality-of-life scores. J Appl Stat. 2013;40:563–82.
32. Ferrari SLP, Cribari-Neto F. Beta regression for modelling rates and proportions. J Appl Stat. 2004;31:799–815.
33. Hunger M, Baumert J, Holle R. Analysis of SF-6D Index Data: Is Beta Regression Appropriate? Value Health. 2011;14:759–67.
34. Little RJ, Lewis RJ. Estimands, Estimators, and Estimates. JAMA. 2021;326(10):967–8.
35. Lawrance R, Degtyarev E, Griffiths P, Trask P, Lau H, D'Alessio D, et al. What is an estimand & how does it relate to quantifying the effect of treatment on patient-reported quality of life outcomes in clinical trials? J Patient Rep Outcomes. 2020;4:1–8.
36. Hedges LV. Distribution theory for glass's estimator of effect size and related estimators. J Educ Stat. 1981;6:107–28.
37. Akaike H. A new look at the statistical model identification. IEEE Trans Automat Contr. 1974;19:716–23.
38. Qian Y, Walters SJ, Jacques R, Flight L. Comprehensive review of statistical methods for analysing patient-reported outcomes (PROs) used as primary outcomes in randomised controlled trials (RCTs) published by the UK's Health Technology Assessment (HTA) journal (1997–2020). BMJ Open. 2021. p. 51673.
39. Waterhouse J, Walters S, Oluboyede Y, Lawson R. A randomised 2 × 2 trial of community versus hospital pulmonary rehabilitation, followed by telephone or conventional follow-up. Health Technol Assess. 2010;14:1–140.
40. Mountain G, Windle G, Hind D, Walters S, Keertharuth A, Chatters R, et al. A preventative lifestyle intervention for older adults (lifestyle matters): a randomised controlled trial. Age Ageing. 2017;46:627–34.
41. Mountain GA, Hind D, Gossage-Worrall R, Walters SJ, Duncan R, Newbould L, et al. 'Putting Life in Years' (PLINY) telephone friendship groups research study: pilot randomised controlled trial. Trials. 2014;15:141.
42. Kharroubi SA. Analysis of SF-6D health state utility scores: is beta regression appropriate? Healthcare. 2020;8:525.
43. Walters SJ, Campbell MJ, Lall R. Design and analysis of trials with quality of life as an outcome: a practical guide. J Biopharm Stat. 2001;11:155–76.
44. Clare L, Kudlicka A, Oyebode JR, Jones RW, Bayer A, Leroi I, et al. Goal-oriented cognitive rehabilitation for early-stage alzheimer's and related dementias: The GREAT RCT. Health Technol Assess. 2019;23:1–244.
45. Brealey S, Northgraves M, Kottam L, Keding A, Corbacho B, Goodchild L, et al. Surgical treatments compared with early structured physiotherapy in secondary care for adults with primary frozen shoulder: The UK frost three-arm RCT. Health Technol Assess. 2020;24:1–161.
46. Vanderhout S, Fergusson DA, Cook JA, Taljaard M. Patient-reported outcomes and target effect sizes in pragmatic randomized trials in ClinicalTrials.gov: A cross-sectional analysis. PLoS Med. 2022;19:e1003896.
47. Bell ML, Fiero MH, Dhillon HM, Bray VJ, Vardy JL, Kabourakis M, et al. Statistical controversies in cancer research: using standardized effect size graphs to enhance interpretability of cancer-related clinical trials with patient-reported outcomes. Ann Oncol. 2017;28:1730–3.
48. Parsons N, Griffin XL, Stengel D, Carey Smith R, Perry DC, Costa ML. Standardised effect sizes in clinical research. Bone Joint J. 2014;96-B:853–4.
49. Forstmeier W, Wagenmakers EJ, Parker TH. Detecting and avoiding likely false-positive findings – a practical guide. Biol Rev. 2017;92:1941–68.
50. Calvert M, Kyte D, Mercieca-Bebber R, Slade A, Chan A-W, King MT, et al. Guidelines for inclusion of patient-reported outcomes in clinical trial protocols. JAMA. 2018;319:483.
51. Khan I, Bashir Z, Forster M. Interpreting small treatment differences from quality of life data in cancer trials: an alternative measure of treatment benefit and effect size for the EORTC-QLQ-C30. Health Qual Life Outcomes. 2015;13:1–2.
52. Boulesteix A-L, Groenwold RH, Abrahamowicz M, Binder H, Briel M, Hornung R, et al. Introduction to statistical simulations in health research. BMJ Open. 2020;10:39921.
53. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. Stat Med. 2019;38:2074–102.

Publisher's Note