Quality of Life Research (2023) 32:1391–1400
https://doi.org/10.1007/s11136-022-03239-3

SPECIAL SECTION: METHODOLOGIES FOR MEANINGFUL CHANGE

Inferring meaningful change in quality of life with posterior predictive distribution: an alternative to standard error of measurement

Yuelin Li[1]

Accepted: 11 August 2022 / Published online: 9 September 2022

Abstract

Objective In the absence of population-based information, distribution-based meaningful change metrics have been previously found to perform similarly. Yet, it is unknown how a Bayesian approach derived from Posterior Predictive Distribution (PPD) of anticipated changes would compare against existing metrics.

Methods PPD defines meaningful change as change scores that exceed the amount expected from the posterior predictive distribution given a previous score. The PPD adjusts for common statistical phenomena that arise in a pre-test–post-test setting, such as regression to the mean and post-test drift. The PPD was compared to Reliable Change Index (RCI) and Gulliksen–Lord–Novick (GLN) methods using published real-world data and simulated hypothetical data, respectively.

Results Real-world data showed that the methods made similar classifications when the measurement reliability was above 0.80. When reliability was low at 0.50 and thus more susceptible to regression to the mean effects, PPD and GLN were able to correct for it but not the RCI. However, PPD was more conservative and sensitive to biased priors. The simulation study showed that the three methods performed similarly overall, but PPD was slightly better in detecting prevalent changes, e.g., at time 2 (against RCI at $p < 0.0001$; against GLN at $p < 0.0001$) and time 3 ($p = 0.024$, $p = 0.002$).

Conclusions When measurement reliability is high, as is frequent in HRQOL development efforts, the three methods performed similarly. At a cost of more conservative cutoffs and complex calculations, the Bayesian PPD nevertheless confers practical advantages when reliability is low. It may be worthy of further research and applications.

Keywords Bayesian regression · Bayesian posterior predictive distribution · meaningful change · Bayesian statistics

Introduction

Much has been written about ways to determine meaningful change in Health-Related Quality of Life (HRQOL) in individual persons [1–14]. The methods are broadly divided into two types, "anchor-based" and "distribution-based" methods [14]. Anchor-based methods typically define meaningful change as score differences that corroborate an important clinical event [15–19], while distribution-based methods use psychometric measurement errors as the metrics [5, 7, 10–12, 14]. One of the earliest distribution-based methods was the Reliable Change Index (RCI) of Jacobson and Truax [7], which was based on standard error of difference (SED) as the unit metric, derived from the Standard Error of Measurement (SEM) [20], where SEM $= s_x\sqrt{1-r}$, solely determined by the standard deviation of the pre-test score and the test–retest reliability. A meaningful change is one that exceeds the inherent measurement error [7]. Many alternative metrics have been proposed along the same line of reasoning [1, 2, 4, 6, 13, 21–23], including the widely accepted Cohen's effect size metric [24], or by asking patients directly as a face-valid approach to meaningful symptom relief [25, 26].

Despite the continuing efforts in finding a general guidance on the "right" way to approach meaningful change [5, 27], the debate is not easily resolved [9], partly because of the variabilities in the specific instruments used, the sample studied, the intervals between repeated assessments, and the specific change metrics examined. Prompted by these issues, Sloan et al. [12] proposed pragmatic rules guided by a unified framework between effect size and a scale's theoretical

✉ Yuelin Li
liy12@mskcc.org

[1] Department of Psychiatry and Behavioral Sciences, Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, 641 Lexington Ave, 7th Floor, New York, NY 10022, USA

range. They concluded that a 'moderate' difference of half a standard deviation (SD) provided a reasonable approximation of a clinically meaningful difference in many QOL measures, which continued to be considered a simple but pragmatic cutoff in subsequent research (see Revicki et al. [27] for a review). Pragmatic rules may be preferred by researchers because of their ease of computation.

The continuing effort to refine meaningful change metrics has been overshadowed by a longstanding controversy since the 1970s by Cronbach and Furby [28]. Recently, Gu et al. [29, 30] retold their assertion that raw pre–post change scores from psychometric measures are generally not very useful because of measurement error compounded by test–retest noise, which can make the change scores discouragingly unreliable. Contrary to such pessimistic beliefs, Gu et al. [29, 30] showed that, although researchers should heed their concerns and proceed with care, change scores are not destined to fail, and that metrics such as the SEM remains useful for evaluating meaningful change at individual persons.

Erosion of reliability can also arise because of statistical artifacts such as regression to the mean. Extreme scores away from the mean at the initial assessment tend to drift back toward the mean when assessed again, such that changes may not necessarily represent a true improvement or deterioration. Regression to the mean always occurs in practice because it only disappears in perfect test–retest correlation, and the lower the correlation the stronger the effect [31]. A head-to-head comparison between the RCI and other metrics, including the Gulliksen–Lord–Novick (GLN) adjustment for mean drift [32], showed no practical difference between specific metrics when reliability exceeds 0.90 [1]. More recently, a bivariate regression approach detected subtle worsening scores from an already low pre-test score [21, 33]. This makes sense because a low pre-test score that fails to drift toward the mean may represent a persistent deficit. There may be nuances between the adjustment in GLN and a bivariate regression that need further exploration, especially when reliability greater than 0.90 may not be available in HRQOL.

A Bayesian bivariate regression, however, may not work as well as classic regression in detecting a subtle persistent deficit, because priors may add to the inherent uncertainty in parameters [34], which would enlarge the meaningful change cutoff (especially if priors are biased). Motivated by these considerations, a Bayesian bivariate regression approach is proposed here. The overall goal of this article is not to make an exhaustive comparison between existing metrics, but to examine how these metrics behave and offer readers an intuition to begin approaching change scores from a Bayesian perspective.

This article is organized as follows. The Bayesian approach is described first, then evaluated in two separate studies. Study 1 compares it against RCI and GLN in classifying the published data in Jacobson and Truax. Also examined in study 1 is the model's sensitivity to biased priors. In study 2, the model is evaluated by the simulated hypothetical data prepared by the Psychometric Special Interest Group, International Society on Quality of Life Research (ISOQOL). The syntax code for the model is provided in the supplementary online materials, with annotations, so that readers can apply the examples to verify the results presented herein.

Methods

Bayesian bivariate normal model and posterior predictive distribution

The setting involves a retest score (y_i) from the ith person, given his/her baseline score (x_i), in a bivariate normal model with a *likelihood function*:

$$y_i \sim N\left(\alpha + \beta(x_i - \bar{x}), \sigma_\varepsilon^2\right), \quad i = 1, 2, \ldots, n,$$

where $\bar{x}$ is the mean baseline score, and σ_ε^2 the residual variance. The Bayesian bivariate regression model seeks the posterior distribution of $p\left(\alpha, \beta, \sigma_\varepsilon^2 | x_i, y_i\right)$, and assuming a reference prior that is independently uniform in $\alpha, \beta, \sigma_\varepsilon^2$, it yields the posterior distributions:

$$\hat{\alpha} \sim t_{n-2}\left(\bar{y}, s/\sqrt{n}\right), \quad \hat{\beta} \sim t_{n-2}\left(S_{xy}/S_{xx}, s/\sqrt{S_{xx}}\right), \quad \sigma_\varepsilon^2 \sim S_{ee}\chi_{n-2}^{-2},$$

where $S_{ee} = S_{yy} - S_{xy}^2/S_{xx} = S_{yy}\left(1 - r^2\right)$, $S_{xx} = \sum\left(x_i - \bar{x}\right)^2$, $S_{yy} = \sum\left(y_i - \bar{y}\right)^2$, $S_{xy} = \sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)$, and that $s = \sqrt{S_{ee}/(n-2)}$. The residual standard error, s, represents the inherent uncertainty in predicting a person's post-test score y_i. The regression coefficients follow a Student's t-distribution listed above; and the residual variance follows an inverse Chi-square distribution with $n - 2$ degrees of freedom.

The Posterior Predictive Distribution (PPD) [35–38] defines the potential post-test observation $\tilde{y}_i$ given the pre-test score $\tilde{x}_i = x_0$. Symbolically,

$$\begin{aligned} \tilde{y}_i &\sim t_{n-2}(\mu_i, \sigma^2) \\ \mu_i &= \hat{\alpha} + \hat{\beta}\left(x_{0i} - \bar{x}\right) \\ \sigma &= s\sqrt{1 + n^{-1} + (x_{0i} - \bar{x})^2/S_{xx}}, \end{aligned} \tag{1}$$

where the predicted post-test mean given the pre-test score is $\hat{\alpha} + \hat{\beta}(x_{0i} - \bar{x})$, and the unit metric in the PPD is the scale parameter, $s\sqrt{1 + n^{-1} + \left(x_0 - \bar{x}\right)^2/S_{xx}}$, by taking the residual standard error s and scaling it by the uncertainties in the

intercept $\left(\frac{1}{\sqrt{n}}\right)$ and slope $\left(\sqrt{(x_0-\bar{x})^2/S_{xx}}\right)$, plus an additional inherent uncertainty s in predicting an individual $\tilde{y}_i$ observation [39]. Regression to the mean is largely handled in the slope error because extreme pre-test x_0 values would need a larger post-test score to exceed the PPD.

Clarifications: uniform priors and the predictive distribution

If $\hat{\alpha}, \hat{\beta}, \sigma_\varepsilon^2$ take on an independently uniform prior, then the posterior distributions of the parameters follow the convenient forms above [37, 40]. The notation above is used, rather than a full Bayesian model specification, because it is easier to follow, calculate, and in the familiar classical regression form. The PPD unit metric as it appears in Eq. (1) also corresponds to how conventional SEM-based metrics are calculated. A fuller Bayesian treatment of the classical regression would have a different notation, which can be found elsewhere [34, 36, 41]. Another point to clarify is that the predictive distribution is not uniquely Bayesian. The Frequentist PPD is identical to the Bayesian version under a uniform prior [39]. However, a Bayesian regression involves an explicit prior, uniform or otherwise, and different priors may affect the PPD, which will be investigated below.

Study 1(a): Comparison between RCI and posterior predictive distribution

Study 1 provides preliminary comparisons between Bayesian PPD, RCI, and the GLN, to examine how these methods react to regression to the mean. The cutoffs were calculated using published data in the original 1976 validation paper for the Dyadic Adjustment Score (DAS) [42], a test–retest correlation of 0.80, a pre-test standard deviation of 7.5 ($\sqrt{S_{xx}}$), a post-test standard deviation of 10 ($\sqrt{S_{yy}}$), pre-test and post-test means at 82 and 83.5, respectively, and a validation sample size of 300. Because the true changes were unknown, we examined the agreements between the three methods using the raw DAS scores in Jacobson and Truax [7]. Supplementary Fig. S1 shows how the PPD was calculated (e.g., 95% PPD was approximately 11.98 for a pre-test score of 85, while the RCI was $9.29 = 1.96 \times 4.74$).

Study 1(b): Sensitivity to priors

Study 1(b) examined whether the specific priors used affected the meaningful change metric. Biased priors may cause problems, especially if the sample size is small. Checking for sensitivity to priors is always recommended as part of the best practices in Bayesian workflow [43]. Four priors were examined: (1) priors inferred from the original 1976 DAS validation data, specifically $N(82, 7.5)$ and $N(1.06, 10)$ for the intercept and slope, respectively; (2) weakly informative prior [44], $N(84, 25)$ and $N(0, 2.5)$, proper priors but provide information intentionally set weaker than the DAS prior; (3) informative prior, $t(df = n-2, 83, 9.5)$ and $t(df = n-2, 1, 0.1)$, fairly precise prior for both the intercept and slope; and 4) an arbitrary, highly confident but biased prior, $t(df = n-2, \text{location} = 0, \text{scale} = 0.5)$ to represent stubbornly confident but mistaken belief that the parameter is exactly zero with high precision. The influence of these priors was examined under two different sample sizes ($n = 300$ and $n = 150$) as a crude examination on a stronger influence from priors when sample size is reduced. Details on prior specifications are found in the Supplement. Pertinent here are the convergence of the Bayesian pseudo posterior draws, which was evaluated by the $\hat{R} \leq 1.10$ diagnostic metric [45], and inspected visually (Supplement Fig. S2).

Study 2: Simulated disease questionnaire 12-item version (SDQ-12)

Study 2 used the simulated data created for this special issue. Volunteers led by Philip Griffiths, PhD., created a hypothetical instrument called Simulated Disease Questionnaire 12-item version (SDQ-12). The simulated dataset contains $n = 2000$ respondents followed longitudinally over four assessment time points. A higher score represents greater disease burden and thus worse outcome. The simulation used seven categories of meaningful change in disease burden in the latent θ parameter values by the following criteria: (1) 'Much Worse', $\Delta\theta \geq 1.5$, indicating a large increase in disease burden; (2) 'Moderately Worse', $1.0 \leq \Delta\theta < 1.5$; (3) 'Minimally Worse', $0.5 \leq \Delta\theta < 1.0$; (4) 'No Change', $-0.5 \leq \Delta\theta < 0.5$; (5) 'Minimally Improved', $-1.0 \leq \Delta\theta < -0.5$; (6) 'Moderately Improved', $-1.5 \leq \Delta\theta < -1.0$; and (7) 'Much Improved', $\Delta\theta < -1.5$. The analyses were blinded to the data simulation parameters. The simulations used ± 1.5 cutoffs for the improved and deteriorated categories, thus study 2 was based on these rather than RCI's ± 1.96 cutoff. A hypothetical validation sample was constructed from information given in the SDQ-12 data. Specifically, the validation dataset had a test–retest reliability of 0.73, a pre-test standard deviation of 133.1, a post-test standard deviation of 141.9, pre-test and post-test means at 18.3 and 18.5, respectively, and a validation sample size of $n = 300$. The RCI was also applied for comparison, using the same reliability of 0.73 and the pre-test standard deviation of 133.1. Model performance was evaluated by the accuracy, sensitivity, specificity, and the F1 score (harmonic mean of sensitivity and specificity) between the methods.

Results

Study 1(a): Simulated posterior predictive distribution

Figure 1 compares the Bayesian PPD and the RCI under four different hypothetical simulation scenarios. Figure 1a, b shows that, when the test–retest reliability was moderately high, the two methods mostly agreed on classifying the changes. The PPD was visibly wider and slightly tilted from the diagonal line due to a small upward drift from 82 at pre-test to the post-test mean of 83.5. The tilt also caused a handful of 'recovered' cases by the RCI be deemed 'unchanged' by the PPD. Figure 1b shows that the validation sample size played a relatively minor role in the PPD intervals when the reliability was high.

However, disagreements arose when the reliability in the validation sample was low. Figure 1c shows that the PPD

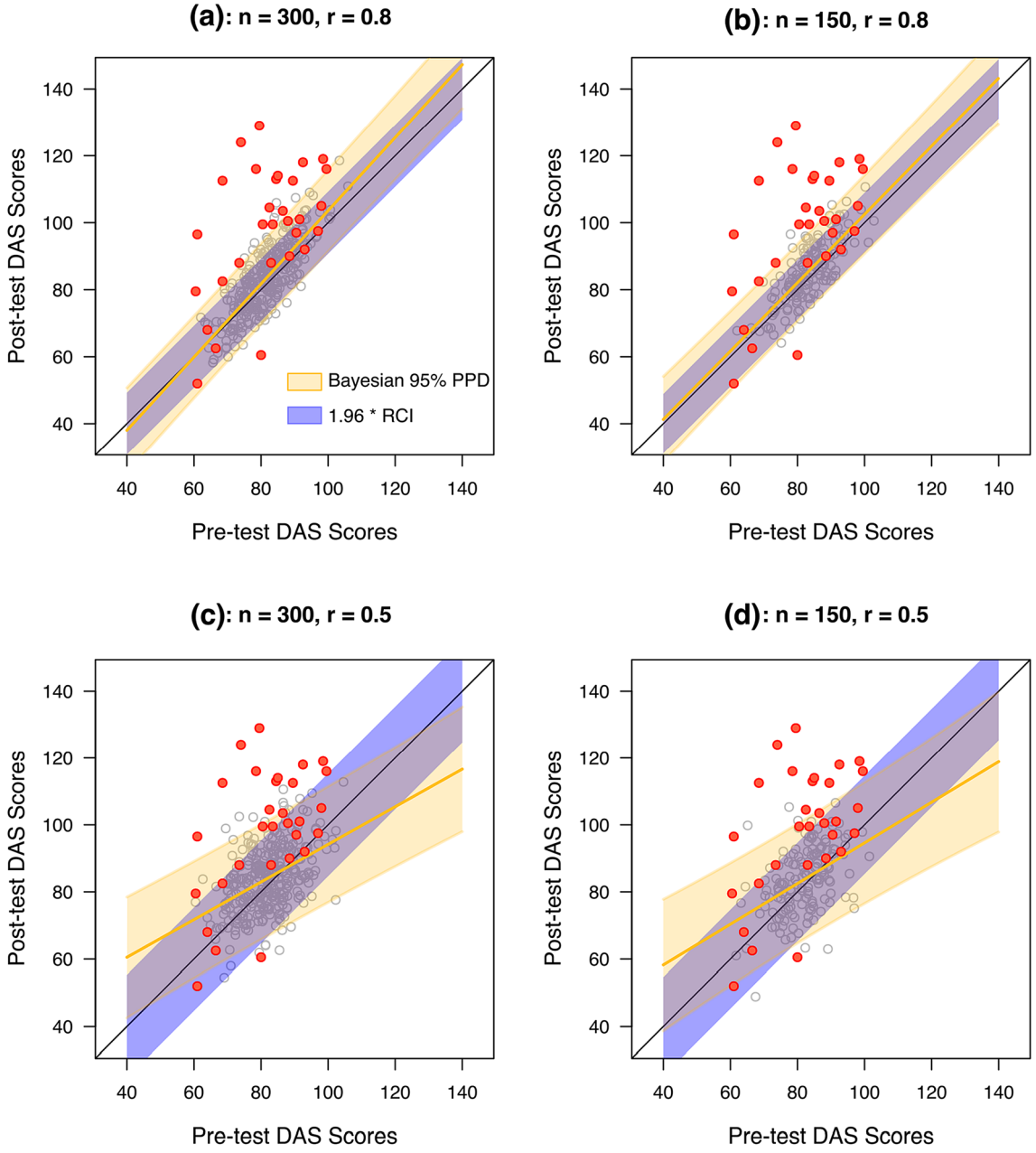

Fig. 1 A visual comparison between Bayesian PPD and the RCI. The blue shaded area represents the RCI. Plotted in light gray circles are the simulated bivariate normal data in the hypothetical validation sample. Also plotted in filled red circles are the original data in Jacobson and Truax.[7] (Color figure online)

had a distinct clockwise tilt toward the post-test mean score, a clear indication of the stronger effect of regression to the mean the lower the correlation [31]. By contrast, the RCI widened but without the tilt. As a result, Fig. 1c shows that a change from a low baseline score of 60.5 to a higher score of 79.5 at post-test was deemed 'improved' by the RCI but 'unchanged' by the PPD. A reduction from 61.0 to 52.0 was deemed 'unchanged' by the RCI but 'worsened' by the PPD. To further characterize the classifications made by PPD, RCI, and GLN, Table 1 summarizes their differences and agreements. All three methods agreed well when the reliability was high ($r = 0.80$), as shown in high kappa (> 0.80) and accuracy metrics above 0.90. Disagreements emerged in low reliability ($r = 0.50$), where discordant classifications were found in the GLN method while the kappa agreement of 0.82 remained high between PPD and RCI.

Study 1(b): Sensitivity to priors

Table 2 summarizes the posterior parameter estimates and classification results under different priors. When the sample size was large ($n = 300$), both the weakly informative and informative priors yielded similar posterior estimates. However, the highly confident but biased prior yielded clearly problematic posterior estimates. The classifications were also similar between RCI's classifications and weakly informative and informative priors. The biased prior deemed all cases 'unchanged' due to the wide PPD interval. Further, when the sample size was reduced ($n = 150$) and reliability was low, the weakly informative and informative priors yielded similar estimates and slightly more conservative classifications than the RCI, and the biased prior yielded completely nonsensical estimates.

Study 2: Results for the simulated SDQ-12 model

Figure 2 shows the changes in total SDQ-12 scores over time. Plotted in color are the known "Much Worse" change (in red) and "Much Improved" change (green). The PPD showed a wider interval, but both methods agreed on most cases. Disagreements were visible primarily in cases with extreme pre-test scores (e.g., improved cases from extreme pre-test low and deteriorated cases from extreme pre-test high), due to regression to the mean at post-test. PPD made no adjustment for post-test drift because the simulation sets the means at 18.2 and 18.5, respectively. Some simulated values contradicted their groups, for example, that several "Much Improved" observations fell above the diagonal line (worse scores), which arose by chance in the simulation process.

Table 1 Comparisons between predictions of meaningful change made by RCI, GLN, and PPD under high ($r = 0.80$) and low ($r = 0.50$) test–retest reliabilities

Comparisons between predictions of meaningful change made by RCI, GLN, and PPD under high (r = 0.80) and low (r=0.50) test-retest reliabilities.										
		PPD vs. RCI [a]			GLN vs. RCI			PPD vs. GLN		
		'worsened'	'no change'	'improved'	'worsened'	'no change'	'improved'	'worsened'	'no change'	'improved'
High reliability (r=0.80)	'worsened'	1	0	0	1	1	0	1	0	0
	'no change'	0	9	1	0	7	0	1	7	2
	'improved'	0	0	19	0	1	20	0	0	19
	Accuracy (CI) [b]	0.97 (0.83, 1.00)			0.93 (0.78, 0.99)			0.90 (0.73, 0.98)		
	Kappa [c]	0.93			0.86			0.79		
	Weighted F1 [d]	0.97			0.94			0.90		
	Sensitivity	0.97			0.93			0.90		
	Specificity	0.99			0.93			0.97		
		'worsened'	'no change'	'improved'	'worsened'	'no change'	'improved'	'worsened'	'no change'	'improved'
Low reliability (r=0.50)	'worsened'	1	1	0	1	1	0	2	0	0
	'no change'	0	12	2	0	9	1	0	10	4
	'improved'	0	0	14	0	3	15	0	0	14
	Accuracy (CI)	0.90 (0.73, 0.98)			0.83 (0.65, 0.94)			0.87 (0.69, 0.96)		
	Kappa	0.82			0.69			0.76		
	Weighted F1	0.91			0.83			0.87		
	Sensitivity	0.90			0.83			0.87		
	Specificity	0.95			0.86			0.93		

[a]The confusion matrices were constructed by cross-tabulating predictions made by the first method (across rows) against the second (across columns)

[b]Accuracy and the CI were the overall accuracy of the model, the fraction of the total samples that were correctly classified (including true positives and true negatives)

[c]Cohen's kappa inter-rater reliability measure

[d]The $F1$ score was a composite metric of precision (accuracy among positive predictions) and recall (accuracy among actual positive cases), which often works well for imbalanced data

Table 2 Summary of model parameter estimates and classification results under RCI and different Bayesian priors

Hypothetical $n=300$, $r=0.80$	DAS psychometrics		Weakly informative prior[a]		Informative prior[b]		Informative and biased prior[c]	
	Coef	95% RCI	Coef	95% PPI[d]	Coef	95% PPI	Coef	95% PPI
Parameters								
Intercept	81.58	72.10, 91.06	83.53	82.84, 84.20	83.52	82.82, 84.22	0.93	−0.06, 1.96
Slope	1.0	NA[e]	1.09	1.00, 1.19	1.08	0.99, 1.16	0.42	−0.39, 1.17
Classifications								
'Unchanged'	9		11		11		30	
'Improved'	20		18		18		0	
'Deteriorated'	1		1		1		0	
Hypothetical $n=150$, $r=0.50$	Coef	95% RCI	Coef	95% PPI	Coef	95% PPI	Coef	95% PPI
Parameters								
Intercept	81.58	66.88, 96.28	83.30	81.84, 84.65	83.30	81.90, 84.64	0.49	−0.47, 1.48
Slope	1	NA	0.92	0.74, 1.11	0.96	0.83, 1.09	0.21	−0.62, 1.05
Classifications								
'Unchanged'	13		18		17		30	
'Improved'	16		11		12		0	
'Deteriorated'	1		1		1		0	

[a]Normal(location = 84, scale = 25) for the intercept and Normal(location = 0, scale = 2.5) for the slope

[b]Student's $t(df=n-2$, location = 83, scale = 9.5) for the intercept and $t(df=n-2$, location = 1, scale = 0.1) for the slope

[c]Student's $t(df=n-2$, location = 0, scale = 0.5) for both the intercept and the slope

[d]95% posterior predictive interval

[e]The RCI has the slope always fixed at 1.0, thus the margin of error does not apply

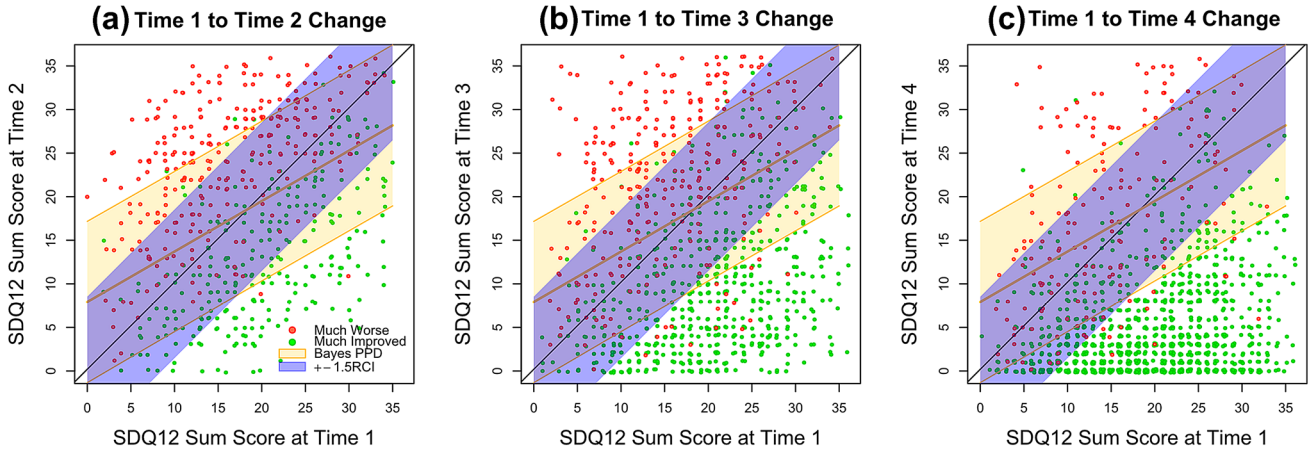

Fig. 2 Scatter plots of the changes in total SDQ-12 scores over time, with the blue shaded area representing the ±1.5 reliable change index. The corresponding ±1.5 posterior predictive intervals are plotted in light yellow. Points in red are the cases with a known "Much Worse" change and points in green are "Much Improved" cases. Small amounts of random perturbations are added to the points to provide better visual clarity. The main difference between the PPD and RCI is the Bayesian tilt to accommodate regression to the mean, which makes the PPD less likely to identify the "Much Worse" cases. (Color figure online)

Table 3 summarizes the performance of the three methods in identifying the simulated changes over time. The leftmost column shows the distribution of true change categories over time. The most prevalent category is the "Middle" category at time 2 (1429 cases), again the "Middle" category at time 3 (1057 cases), and the "Much Improved" category at

Table 3 Comparison between the Bayesian PPD, RCI, and GLN methods in identifying true changes in the simulated symptom data

Comparison between the Bayesian PPD, RCI, and GLN methods in identifying true changes in the simulated symptom data.									
	Bayesian PPD			RCI			GLN		
	"Improved"	Middle	"Worse"	"Improved"	Middle	"Worse"	"Improved"	Middle	"Worse"
True Category at Time 2 [a]									
"Much Improved" (n=278)	99	176	3	110	161	7	115	173	2
Middle (n=1429)	146	1166	117	174	1078	177	156	1070	154
"Much Worse" (n=293)	2	191	100	2	159	132	7	186	137
Accuracy (CI)	0.68 (0.66, 0.70)			0.66 (0.64, 0.68)			0.66 (0.64, 0.68)		
Kappa	0.24			0.26			0.27		
Weighted F1 score	0.67			0.66			0.67		
Sensitivity	0.68			0.66			0.66		
Specificity	0.52			0.57			0.58		
Time 3									
"Much Improved" (n=634)	375	252	7	383	239	12	399	260	18
Middle (n=1057)	231	756	70	242	707	108	224	688	161
"Much Worse" (n=309)	16	188	105	19	154	136	11	109	130
Accuracy (CI)	0.62 (0.60, 0.64)			0.61 (0.59, 0.63)			0.61 (0.59, 0.63)		
Kappa	0.33			0.34			0.34		
Weighted F1 score	0.61			0.61			0.61		
Sensitivity	0.62			0.61			0.61		
Specificity	0.69			0.71			0.71		
Time 4									
"Much Improved" (n=1141)	916	223	2	859	278	4	944	332	23
Middle (n=685)	312	349	24	292	354	39	193	318	102
"Much Worse" (n=174)	20	116	38	17	103	54	4	35	49
Accuracy (CI)	0.65 (0.63, 0.67)			0.63 (0.61, 0.65)			0.66 (0.63, 0.68)		
Kappa	0.33			0.32			0..34		
Weighted F1 score	0.64			0.63			0.64		
Sensitivity	0.65			0.63			0.66		
Specificity	0.69			0.69			0.69		

[a]The confusion matrices were constructed by cross-tabulating predictions made by the methods (across columns) against the simulated true category (across rows)

time 4 (1141 cases). The PPD was consistently better than RCI and GLN in identifying the most prevalent category, at time 2 (1166/1429 vs. 1078/1429, $p<0.0001$; 1166/1429 vs. 1017/1429, $p<0.0001$) and time 3 (756/1057 vs. 707/1057, $p=0.024$; 756/1057 vs. 688/1057, $p=0.002$). The size of the "Much Improved" category increased steadily over time, from the 278 cases at time 2 to 634 cases at time 3, and again to 1141 cases at time 4. By time 4, the PPD was better than RCI in detecting the "Much Improved" category (916/1141 vs. 859/1141, $p=0.0048$) but was about the same as GLN (916/1141 vs. 944/1141, $p=0.145$). PPD's ability in identifying the most prevalent category over time came at a cost of identifying fewer cases of the least frequent category, e.g., the "Much Worse" category at time 3 (105/309 vs. 136/309 using RCI, $p<0.013$) and marginally at time 4 (38/174 vs. 54/174, $p=0.068$).

Overall, no discernible difference was apparent between the three methods, with the overall accuracies across all three methods hovering around 0.60 and 0.70, with comparable other performance metrics across the post-test time points. Over time 3 and time 4, more people fell into the "much improved" category. The Bayesian PPD and the GLN were slightly better than the RCI in picking up this change, at 916, 944, and 859 out of 1141 cases for PPD, GNL and RCI, respectively.

Discussion

This study proposes a Posterior Predictive Change model (PPD) to define meaningful change in HRQOL research, where meaningful change is an observed score that exceeds the amount expected from the posterior predictive error given a previous score. If we summarize the overall results in a broad stroke, then the main finding is that PPD, RCI, and GLN perform similarly when the measurement tool has a high reliability. This is consistent with previous head-to-head comparisons between alternative SEM-guided metrics [1]. The gist is that there are no compelling reasons in the current study settings to favor one method over another, and the ease of calculation may support simple rules such as RCI, GLN, or the half-SD rule over the more complex Bayesian PPD if the measurement tool is known to be highly reliable. These pragmatic rules may also include Sloan

et al.'s 'worms, ducks, and elephants' rule [12] and various other rules based on standardized scores [27]. However, subtle but important differences set Bayesian PPD apart from SEM-guided metrics.

The Bayesian PPD is better than the RCI and GLN in identifying the most prevalent changes in the simulated large sample of 2000 individuals. Of particular interest is that, in study 2, the Bayesian PPD is especially better at identifying "Much Improved" cases at time 4, because by then most cases have a considerably improved score. This is of course unknowable in real-world HRQOL assessments. But if the classifiers are imperfect (clearly shown in sensitivity near 0.60), the PPD may confer a small but nontrivial advantage in identifying the most common change pattern. These findings appear to agree with literature findings cited above. PPD may be particularly useful if the scientific goal is to accurately identify the most common change patterns, but it inevitably misses some uncommon cases.

Another subtle difference is that PPD discounts changes associated with extreme pre-test scores. More extreme pre-test scores away from the mean need to change to a greater extent before they exceed the PPD cutoff. For instance, Fig. 1a shows that a high baseline symptom burden of 35 must drop by 20 points to be deemed a meaningful improvement, while a score near the pre-test mean of 20 only needs to drop by 12 points. This is unique to PPD, due in part to its inclusion of the pre-test score in Eq. (1) in calculating the change threshold, and not necessarily because it adjusts for regression to the mean, because the GLN cutoffs remain the same irrespective of the pre-test score. PPD in essence treats extreme pre-test scores more like aberrations than the norm, and chance alone may pull them closer to the norm at post-test, thus the wider meaningful change margin. This appears to be consistent with previous Bayesian methods of meaningful change, proposed by Hsu [46, 47], where extreme pre-test scores should be evaluated with caution. However, Hsu's method requires knowledge of the norms in the dysfunctional and functional populations, which are often unavailable in HRQOL research. In practice, the discount of extreme pre-test scores may be worth consideration because scores near the ceiling or floor of a HRQOL scale can only move in one direction.

Another relevant point is that a PPD cutoff does not have to be fixed at 95%. In study 2, a 93% confidence is used because a ± 1.5 standard deviation is used to create the data used in this special issue. For a Bayesian analyst, the posterior confidence is an integral part of the statistical thinking. There may be only an 80% posterior confidence that a patient has improved. In a different case the posterior confidence may be lower still. What is "meaningful" to one person's change in QOL may be different from another person's appraisal.

The PPD can leverage a Bayesian notion that "today's posterior is tomorrow's prior" [48]. For instance, a psychometric instrument may be used by several research teams. The PPD allows the posteriors of one study to be considered as the prior specifications of a new study. For example, one study might originally use somewhat uncertain priors to calculate the PPD cutoff ($N(\tilde{y} = 84, s = 25)$ for the intercept and $N(0, 2.5)$ for the slope). When combined with data in their study, the researchers update their posteriors (say $N(\tilde{y} = 83, s = 12)$ and $N(-0.5, 1.5)$). Note that the posteriors have lower margins of error—we now know more precisely what the cutoff would be. Knowledge accumulation can go further in Bayesian inference. The next research team, heeding the "today's posterior is tomorrow's prior" concept, may consider these posteriors for their priors in a new study. Thus, information pooled across studies may help improve the precision of a meaningful change model as data accrue. Over time, as an instrument is applied in many research studies, a model like the PPD can accumulate knowledge more effectively than the fixed-parameter RCI and its derivatives. Sensitivity to priors should always be checked, and the stubbornly confident but biased prior in Study 1(b) shows the peril of omitting a sensitivity analysis in real-world data. Study 1(b) was a pared-down version of a fuller sensitivity analysis using sample sizes more commonly observed in HRQOL research. A fuller sensitivity analysis requires more detailed investigations on how priors exert an increasingly stronger influence as sample size becomes progressively small. The current findings do not reflect a general rule about sensitivity to priors. The syntax code in the Supplement may help a reader to begin a more detailed analysis.

The PPD can also include additional covariates to cope with the issues that complicate a consensus on the "right" rule for meaningful change, such that the inclusion of age, frailty and comorbidities may help enhance accuracy. RCI and its derivatives can do this, but it would probably involve different cutoffs for different age groups, an impractical application. However, knowledge accumulation in this context is not without its concerns. We show that a highly confident but biased prior can easily break a model such that its effects cannot be rectified even with a reasonably large sample.

This study also shows several drawbacks in the Bayesian approach. The PPD is computationally more involved, which may seem intimidating to a beginner. Indeed, why would someone spend the time and effort calculating the posterior predictive intervals when they seem to only make small differences? However, the Bayesian steps are not much more complicated than running a regression, and the supplementary materials provide the needed syntax code. Furthermore, the PPD offers the potential for knowledge

accumulation, which may likely make a substantive difference over time. Future research may focus on a more comprehensive examination, e.g., when regression to the mean and drift are controlled and manipulated in greater detail in the simulations to examine its performance more thoroughly. Further work would also need to be applied to real data, possibly corroborated by clinical anchors to better demonstrate its real-life applicability.

These limitations notwithstanding, the notion of Bayesian PPD as applied in meaningful change is novel. Previous distribution-based methods are all based on the psychometric measurement error or some other pragmatic but fixed metric, which fail to incorporate the inherent uncertainty in regression to the mean and possibly practice effect. The proposed Bayesian PPD approach reminds HRQOL researchers to take a broader statistical perspective to incorporate these uncertainties to go beyond the traditional boundaries between psychometrics and Bayesian statistics.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11136-022-03239-3.

Authors' contributions YL, PhD. is the sole author of this manuscript and is responsible to all aspects of data preparation and analyses.

Funding This work was in part funded by the National Institute of Health Grant P30 CA008748 to Memorial Sloan Kettering Cancer Center.

Availability of data and materials Special thanks go to Philip C. Griffiths, PhD. for leading the efforts in creating the dataset on the hypothetical Simulated Disease Questionnaire 12-item version (SDQ-12), as well as other volunteers in the Psychometrics Special Interest Group, International Society for Quality Of Life Research (ISOQOL). The dataset is publicly accessible at: https://drive.google.com/drive/folders/1pt800y4xMGjzu7NyXoJgkFqTdggfSO3q (accessed March 1, 2021).

Code availability The supplementary online materials contain syntax codes written in the statistical language R to calculate the metrics used in this manuscript.

Declarations

Conflict of interest The authors declare that there is no conflict of interest.

Consent for publication I consent to participate in the process of peer review for consideration for publication in Quality Of Life Research.

Consent to participate I consent to participate in the process of peer review for consideration for publication in Quality Of Life Research.

Ethical approval The data used in this manuscript include two datasets: (1) a previously published, publicly available data (Jacobson and Truax [7], their Table 2); and (2) simulated data as described above, which was created specifically for this special issue.

References

1. Atkins, D. C., Bedics, J. D., McGlinchey, J. B., & Beauchaine, T. P. (2005). Assessing clinical significance: Does it matter which method we use? *Journal of Consulting and Clinical Psychology, 73*(5), 982–989.
2. Bauer, S., Lambert, M. J., & Nielsen, S. L. (2004). Clinical significance methods: A comparison of statistical techniques. *Journal of Personality Assessment, 82*(1), 60–70.
3. Busch, A. M., Wagener, T. L., Gregor, K. L., Ring, K. T., & Borrelli, B. (2011). Utilizing reliable and clinically significant change criteria to assess for the development of depression during smoking cessation treatment: The importance of tracking idiographic change. *Addictive Behaviors, 36*(12), 1228–1232.
4. Chelune, G. J., Naugle, R. I., Luders, H., Sedlak, J., & Awad, I. A. (1993). Individual change after epilepsy surgery: Practice effects and base-rate information. *Neuropsychology, 7*, 41–52.
5. Hill, S. W. (2019). Components and methods of evaluating reliable change in cognitive function. In *Neurosurgical neuropsychology: The practical application of neuropsychology in the neurosurgical practice* (pp. 39–61). Elsevier Academic Press.
6. Hsu, L. M. (1989). Reliable changes in psychotherapy: Taking into account regression toward the mean. *Behavioral Assessment, 11*, 459–467.
7. Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59*(1), 12–19.
8. Lunnen, K. M., & Ogles, B. M. (1998). A multiperspective, multivariable evaluation of reliable change. *Journal of Consulting and Clinical Psychology, 66*(2), 400–410.
9. Massen, G. H. (2001). The unreliable change of reliable change indices. *Behaviour Research and Therapy, 39*, 495–498.
10. Sloan, J. A., Cella, D., & Hays, R. D. (2005). Clinical significance of patient-reported questionnaire data: Another step toward consensus. *Journal of Clinical Epidemiology, 58*(12), 1217–1219.
11. Sloan, J. A., Frost, M. H., Berzon, R., et al. (2006). The clinical significance of quality of life assessments in oncology: A summary for clinicians. *Supportive Care in Cancer, 14*(10), 988–998.
12. Sloan, J. A., Vargas-Chanes, D., & Kamath, C. (2003). Detecting worms, ducks and elephants: A simple approach for defining clinically relevant effects in quality-of-life measures. *Journal of Cancer Integrative Medicine, 1*, 41–47.
13. Speer, D. C., & Greenbaum, P. E. (2002). "Five methods for computing significant individual client change and improvement rates: Support for and individual growth curve approach." Correction to Speer and Greenbaum (1995). *Journal of Consulting and Clinical Psychology, 70*, 6.
14. Yost, K. J., & Eton, D. T. (2005). Combining distribution- and anchor-based approaches to determine minimally important differences: The FACIT experience. *Evaluation and the Health Professions, 28*(2), 172–191.
15. Diaz-Arribas, M. J., Fernandez-Serrano, M., Royuela, A., et al. (2017). Minimal clinically important difference in quality of life for patients with low back pain. *Spine (Phila Pa 1976), 42*(24), 1908–1916.
16. Kon, S. S. C., Canavan, J. L., Jones, S. E., et al. (2014). Minimum clinically important difference for the COPD assessment test: A prospective analysis. *The Lancet Respiratory Medicine, 2*(3), 195–203.
17. Musoro, J. Z., Coens, C., Singer, S., et al. (2020). Minimally important differences for interpreting European Organisation for Research and Treatment of Cancer Quality of Life

Questionnaire Core 30 scores in patients with head and neck cancer. *Head and Neck, 42*(11), 3141–3152.
18. de Vet, H. C. W., Ostelo, R. W. J. G., Terwee, C. B., et al. (2006). Minimally important change determined by a visual method integratingan anchor-based and a distribution-based approach. *Quality of Life Research, 16*(1), 131.
19. Hays, R. D., & Peipert, J. D. (2021). Between-group minimally important change versus individual treatment responders. *Quality of Life Research, 30*, 2765–2772.
20. Harvill, L. M. (1991). Standard error of measurement. *Educational Measurement: Issues and Practice, 10*(2), 33–41.
21. Frerichs, R. J., & Tuokko, H. A. (2005). A comparison of methods for measuring cognitive change in older adults. *Archives of Clinical Neuropsychology, 20*(3), 321–333.
22. Speer, D. C., & Greenbaum, P. E. (1995). Five methods for computing significant individual client change and improvement rates: Support for an individual growth curve approach. *Journal of Consulting and Clinical Psychology, 63*, 6.
23. Crosby, R. D., Kolotkin, R. L., & Williams, G. R. (2003). Defining clinically meaningful change in health-related quality of life. *Journal of Clinical Epidemiology, 56*(5), 395–407.
24. Feinstein, A. R. (1999). Indexes of contrast and quantitative significance for comparisons of two groups. *Statistics in Medicine, 18*(19), 2557–2581.
25. Sloan, J. A., Loprinzi, C. L., Novotny, P. J., Barton, D. L., Lavasseur, B. I., & Windschitl, H. (2001). Methodologic lessons learned from hot flash studies. *Journal of Clinical Oncology, 19*(23), 4280–4290.
26. Hammack, J. E., Michalak, J. C., Loprinzi, C. L., et al. (2002). Phase III evaluation of nortriptyline for alleviation of symptoms of cis-platinum-induced peripheral neuropathy. *Pain (Amsterdam), 98*(1), 195–203.
27. Revicki, D., Hays, R. D., Cella, D., & Sloan, J. (2008). Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *Journal of Clinical Epidemiology, 61*(2), 102–109.
28. Cronbach, L. J. (1970). How we should measure change—Or should we. *Psychological Bulletin, 74*(1), 68–80.
29. Gu, Z. G., Emons, W. H. M., & Sijtsma, K. (2018). Review of issues about classical change scores: A multilevel modeling perspective on some enduring beliefs. *Psychometrika, 83*(3), 674–695.
30. Gu, Z. G., Emons, W. H. M., & Sijtsma, K. (2021). Estimating difference-score reliability in pretest–posttest settings. *Journal of Educational and Behavioral Statistics, 2021*, 1.
31. Bland, J. M., & Altman, D. G. (1994). Regression towards the mean. *BMJ, 308*(6942), 1499.
32. Hu, L. H., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Model, 6*, 1–55.
33. Hinton-Bayre, A. D. (2016). Clarifying discrepancies in responsiveness between reliable change indices. *Archives of Clinical Neuropsychology, 31*(7), 754–768.
34. McElreath, R. (2016). *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC Press.
35. Christensen, R., Johnson, W., Branscum, A., & Hanson, T. E. (2011). *Bayesian ideas and data analysis*. CRC Press.
36. Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., Rubin, D. B. (2020). *Bayesian data analysis* (3rd edn.). http://www.stat.columbia.edu/~gelman/book/.
37. Lee, P. M. (2012). *Bayesian statistics: An introduction* (4th ed.). Wiley.
38. Novick, M. R., & Jackson, P. H. (1974). *Statistical methods for educational and psychological research*. McGraw-Hill.
39. Wonnacott, T. H., & Wonnacott, R. J. (1987). *Regression, a second course in statistics*. R.E. Krieger Pub. Co.
40. Gelman, A. (2004). *Bayesian data analysis* (2nd ed.). Chapman & Hall/CRC.
41. Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.
42. Spanier, G. B. (1976). Measuring dyadic adjustment: New scales for assessing the quality of marriage and similar dyads. *Journal of Marriage and the Family, 38*, 11.
43. Depaoli, S., & van de Schoot, R. (2017). Improving transparency and replication in Bayesian statistics: The WAMBS-Checklist. *Psychological Methods, 22*(2), 240–261.
44. Gelman, A., Jakulin, A., Grazia Pittau, M., & Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics, 2*(4), 1360–1383.
45. Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*, 457–472.
46. Hsu, L. M. (1995). Regression toward the mean associated with measurement error and the identification of improvement and deterioration in psychotherapy. *Journal of Consulting and Clinical Psychology, 63*(1), 141–144.
47. Hsu, L. M. (1996). On the identification of clinically significant client changes: Reinterpretation of Jacobson's cut scores. *Journal of Psychopathology and Behavioral Assessment, 18*(4), 371–385.
48. Lindley, D. V. (1972). *Bayesian statistics: A review*. Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611970654