Bias Associated with Patient Reported Outcome Measures

Joel J Gagnier

Associate Professor, Departments of Epidemiology and Orthopaedic Surgery, University of Michigan, USA

Bradley C Johnston

Associate Professor, Department of Community Health and Epidemiology, Dalhousie University, CAN

Disclosures

- I have no actual or potential conflict of interest in relation to this presentation

The Problem

- There are a large array of heterogeneous Patient Reported Outcome Measures (PROMs) being used in trials
- Systematic reviews of PROMs indicate that:
  - Most PROMs (>90%) have poor or unknown measurement properties
    - Validity, Reliability, Responsiveness, Interpretability
- This calls into question the data derived from them and decision making based on such findings
  - A large proportion of the results from PROMs in clinical trials are potentially biased and misleading
  - Systematic reviews of this data are also potentially biased
- Misleading results waste resources and potentially harm patients
- To date, there has been no attempt at measuring the bias associated with data from PROMs of varying quality in clinical trials or systematic reviews

Objective

- The objective was to assess the bias in findings associated with PROMs of varying psychometric quality in randomized clinical trials (RCTs).

Methods

5 Step Process:

1. Identify PROMs used in rotator cuff disease (RCD) and assess their psychometric quality
   - From a prior study: Huang S, Grant J, Miller B, Mirza FM, Gagnier JJ. A systematic review of psychometric properties of patient reported outcome instruments for use in patients with rotator cuff disease. AJSM. 2015; Jan 26.
   - Given a score for each and across psychometric properties
2. Identify RCTs in patients with RCD using PROMs identified in 1 above
3. Extract outcomes associated with that PROM only and standardize findings using SD
4. Extract additional data from each study (e.g., Intervention details, Follow-up period, Sample size, Risk of bias)
5. Statistical Analyses
   - Primary analysis: Multilevel Regression and elimination procedure, controlling for grouping variable study ID
     - Response variable = standardized results for a PROM for differences between intervention groups (on change from baseline)
     - Predictor variables = psychometric score, sample size, ROB score, funding source, follow-up (months), lack of evidence
     - Sensitivity analysis for poor reporting or no available evidence
   - Separate analysis for the 8 individual psychometric properties

Results

- Included 72 RCTs and 174 separate outcomes
  - Sample Size: Mean 66.8 (95% CI 62.3 to 71.3)
  - ROB Score: Mean 7/10 (95% CI 6.7 to 7.3)
  - Follow-up: Mean 9.7 months (95% CI 7.6 to 11.7)
  - Psychometric property not assessed: 45/128 (35.2% of items)

Mixed effects linear regression: initial model

Covariate	Beta Coefficient (95% CI)	p-value
Psychometric Summary Score	-0.21 (-0.45 to 0.02)	0.075
Psychometric Property Not Assessed	-0.24 (-1.05 to 0.57)	0.564
Follow-up (months)	0.06 (0.003 to 0.11)	0.040
Sample Size	0.01 (-0.02 to 0.04)	0.376
Risk of Bias Score	-0.03 (-0.47 to 0.41)	0.892

  - N=169 (adjusted for 70 clusters); Model p-value = 0.0006

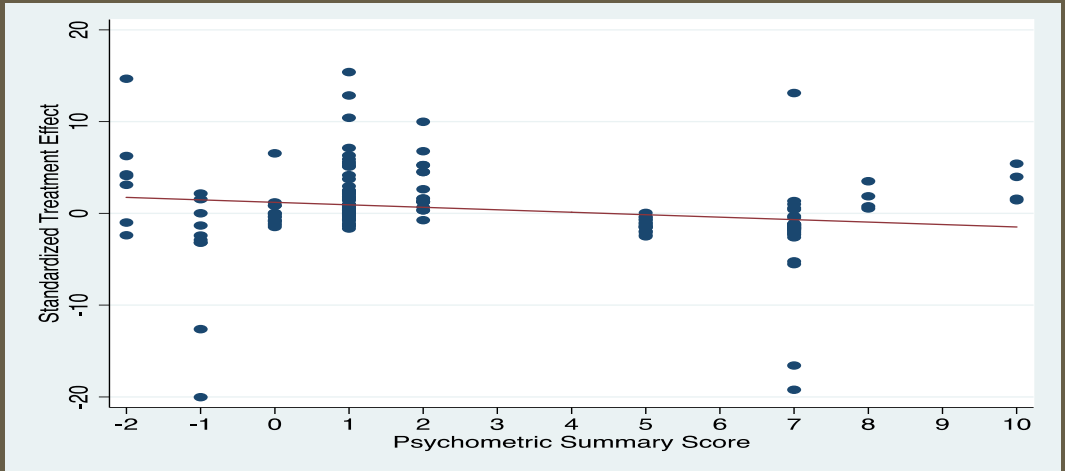
Results

- After step-wise elimination:
  - N=171 (adjusted for 71 clusters); Model p-value = 0.001

Multivariable model

Variable	Beta Coefficient (95% CI)	p-value	N
Psychometric Summary Score	-0.32 (-0.52 to -0.12)	0.002	171
Follow-up (Months)	0.08 (0.02 to 0.13)	0.007	

- Sensitivity analysis for lack of evidence or poor reporting had no effect on findings
- Treatment Effect by PROM quality



Results

Univariable psychometric property regressions on treatment effect

Variable	Beta Coefficient (95% CI)	p-value	N
Internal Consistency	-0.58 (-1.51 to 0.35)	0.223	51
Reliability	-0.15 (-1.99 to 1.69)	0.874	158
Measurement Error	6.51 (2.93 to 10.09)	<0.001	71
Content Validity[1]			1
Criterion Validity[2]			0
Structural Validity	-0.71 (-3.37 to 1.94)	0.598	60
Hypothesis Testing	-0.83 (-1.22 to -0.44)	<0.001	166
Responsiveness	-1.71 (-4.02 to 0.60)	0.146	130

1. One PRO only
2. No Observations

Multivariable model

Variable	Beta Coefficient (95% CI)	p-value	N
Measurement Error	6.83 (3.09 to 10.57)	<0.001	71
Hypothesis Testing	0.27 (-0.55 to 1.08)	0.522	

Discussion

- PROMs with poor or unknown psychometric properties bias (i.e., inflate) the estimates of treatment effect in RCTs (by about 67% on average)
- To our knowledge, this is the first empirical evidence, that variations in the psychometric quality of PROMs bias treatment effect estimates.
- Researchers and clinicians using data from PROMs must be cautious to explore the quality of measure so as to not make decisions based on biased outcome data
  - Systematic reviewers must be cautious when combining data across PROMs, especially when they have poor properties
- More work needs to be done to explore the influence of individual psychometric properties

Discussion

Strengths

- Screened and included a large number of RCTs
- Dual assessments with high agreement
- Used accepted methods for assessing psychometric evidence and ROB
- We performed multivariable modeling controlling for variety of covariates and a grouping variable
- Findings were robust in the face of sensitivity analysis for lack of evidence
- Looked at individual psychometric properties
- Generalizable across interventions for rotator cuff disease

Limitations

- Limited to RCTs in RCD
- Studies may be underpowered for the outcome used
- Lack of available evidence for many psychometric properties, thus these findings may change as evidence accumulates
- Cumulative psychometric quality and risk of bias scores can be misleading
  - We tried to look at individual properties too
- Bias in treatment effects may be confounded with the varying constructs being measured by each PROM (e.g., shoulder function is variably assessed)
  - Constructs which may themselves variably change across studies
- Potential bias due to excluding non-English studies

Thank-You

jgagnier@umich.edu

Extra slides

Patient Reported Outcome Measures

- Patient Reported Outcome Measures (PROMs)
  - Collect information related to constructs that are reported by the patients themselves, without interpretation by other parties
  - Includes perceptions and opinions on symptoms, functioning, health–related quality of life (HRQoL), and satisfaction, among other areas
- The patient perspective on their health is of primary importance
- PROMs are increasingly used to inform clinical decision-making, patient-centered care, health policy and more recently, reimbursement decisions
- Many organizations are recommending PROMs (including CMS, the National Quality Forum, FDA CDRH, and the National Committee for Quality Assurance)
- PROMs are frequently used outcomes in randomized trials

ROB criteria

- 1. Was the randomization method appropriate?
- 2. Was the allocation sequence concealed from those assigning patients to groups?
- 3. Were the participants blind to the intervention?
- 4. Were the outcome assessors (for the primary outcome) blind to the intervention? Describe how the outcome was measured (be sure there is no detection bias)
- 5. Was the outcome measurement performed in the same manner with similar intensity in the groups being compared? (describe who measured outcomes and how…was it valid?)
- 6. Were similarly trained individuals administering the intervention across groups? Describe who this was and their training if available.
- 7. Were all the withdrawals described? Describe the numbers and reasons for withdrawals in each group.
- 8. Were all originally randomized participants analyzed in the groups they were assigned to (i.e. An intention to treat analysis)?
- 9. Was clustering at the group level accounted for the in analyses if applicable?
- 10. Were the groups similar at baseline? If so were adjustments for differences don

Term			Definition
Domain	Measurement property	Aspect of a measurement property	
Reliability			The degree to which the measurement is free from measurement error
Reliability (extended definition)			The extent to which scores for patients who have not changed are the same for repeated measurement under several conditions: e.g. using different sets of items from the same health related-patient reported outcomes (HR-PRO) (internal consistency); over time (test-retest); by different persons on the same occasion (inter-rater); or by the same persons (i.e. raters or responders) on different occasions (intra-rater)
	Internal consistency		The degree of the interrelatedness among the items
	Reliability		The proportion of the total variance in the measurements which is due to 'true'† differences between patients
	Measurement error		The systematic and random error of a patient's score that is not attributed to true changes in the construct to be measured
Validity			The degree to which an HR-PRO instrument measures the construct(s) it purports to measure
	Content validity		The degree to which the content of an HR-PRO instrument is an adequate reflection of the construct to be measured
		Face validity	The degree to which (the items of) an HR-PRO instrument indeed looks as though they are an adequate reflection of the construct to be measured
	Construct validity		The degree to which the scores of an HR-PRO instrument are consistent with hypotheses *(for instance with regard to internal relationships, relationships to scores of other instruments, or differences between relevant groups)* based on the assumption that the HR-PRO instrument validly measures the construct to be measured
		Structural validity	The degree to which the scores of an HR-PRO instrument are an adequate reflection of the dimensionality of the construct to be measured
		Hypotheses testing	Idem construct validity
		Cross-cultural validity	The degree to which the performance of the items on a translated or culturally adapted HR-PRO instrument are an adequate reflection of the performance of the items of the original version of the HR-PRO instrument
	Criterion validity		The degree to which the scores of an HR-PRO instrument are an adequate reflection of a 'gold standard'
Responsiveness			The ability of an HR-PRO instrument to detect change over time in the construct to be measured
	Responsiveness		Idem responsiveness
Interpretability*			Interpretability is the degree to which one can assign qualitative meaning - that is, clinical or commonly understood connotations – to an instrument's quantitative scores or change in scores.

Included PROMs & Scores

	Internal Consistency	Reliability	Measurement Error	Content Validity	Structural Validity	Hypothesis testing	Criterion Validity	responsiveness
ASES	-	++	-	?	?	--	?	?
Constant	?	++	?	na	na	--	?	+
DASH	?	++	--	na	++	+	na	++
KSS(Korean)	?	na	na	?	na	?	?	?
Linsalata	?	?	na	?	na	?	?	?
OSS	?	-	na	?	na	+	na	?
Penn	?	+	?	?	na	?	na	?
RC QOL	?	?	na	?	na	-	na	?
SAL	na	+	na	na	na	?	na	na
SDQ	?	?	na	na	na	-	na	?
SPADI	+++	++	--	na	+++	-/+	?	+
SST	+++	++	-	?	na	++	?	+
SSV	na	na	na	na	na	-	na	na
UCLA	?	+	?	na	na	+	?	?
WOOS	na	+	na	na	na	?	na	?
WORC	++	++	?	+++	na	++	na	+

+: positive evidence ?: intermediate evidence –: negative evidence na: no information available

Rank		Score
1	WORC	10
2	SST	8
3	SPADI	7
4	DASH	5
5	UCLA	2
6	Constant	1
6	Penn	1
6	SAL	1
6	WOOS	1
10	KSS(Korean)	0
10	Linsalata	0
10	OSS	0
13	RCQOL	-1
13	SSV	-1
13	SDQ	-1
16	ASES	-2

Included PROMs

Name of the instrument	Citation	Year of Publication	No. of items and subdomains	Domains Assessed	Response options
American Shoulder and Elbow Surgeons (ASES) Shoulder Outcome Score	Richards et al[63]	1994	10 items in 6 subdomains	Pain, stability, activities of daily living, range-of-motion (ROM) signs, strength, and instability	One 10-cm VAS and score 0-3 for each item
Constant-Murley Scale (CMS)	Constant and Murley[12]	1986	4 subdomains	Pain, activities of daily living, range, power	Item dependent
Disabilities of the Arm, Shoulder, and Hand (DASH) Score	Hudak et al[32]	1996	30 items in 3 subdomains (extra 2 subdomains are optional)	Symptoms and functional status (physical, social, and psychological) optional: sport and music, and work	Score 1-5 for each item
L'Insalata Shoulder Rating Questionnaire (L'Insalata)	l'insalata et al[37]	1997	6 subdomains	Global assessment, pain, daily activities, recreational and athletic activities, work and satisfaction	One 10-cm VAS and score 1-5 for each item
Penn Shoulder Score (Penn)	Leggin and Iannotti[38]	1999	24 items in 3 subdomains	Pain, satisfaction, and function	Score 0-10 for pain, and score 0-3 for function
Quality-of-life Outcome Measure for Rotator Cuff Disease (RCQOL)	Hollinshead et al[30]	2000	34 items in 5 subdomains	Symptoms and physical complaints, sports and recreation, work-related concerns, lifestyle issues, and social and emotional issues	0-100 for each item
Shoulder Disability Questionnaire (SDQ)	van der Heijden, et al.[79]	2000	16 items in 1 subdomain	Functional limitations	16 items, all yes/no answers for each item
Shoulder Pain and Disability Index (SPADI)	Roach et al[64]	1991	13 items in 2 subdomains	Disability and pain	Score 0-11 for each item
Simple Shoulder Test (SST) Score	Godfrey et al	2007	12 items in 1 subdomain	Assess functional limitations of the shoulder relative to the patient's activity of daily living before or after treatment, and work	All yes/no answers for each item
University of California at Los Angeles (UCLA) Shoulder Score	Ellman et al[19]	1986	5 items in 5 subdomains	Pain, function, active forward flexion, strength of forward flexion and motion, satisfaction	Score 1-10 for each item
The Western Ontario Rotator Cuff (WORC) Index	Kirkley et al	2003	21 items in 5 subdomains	Physical symptoms, sport/recreation, work, lifestyle, and emotions	21 items, score 1-100 (i.e., 100-mm VAS) for each item