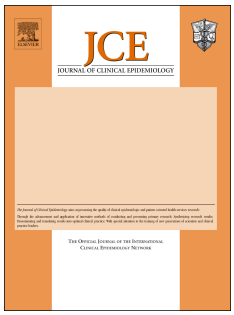
Accepted refereed manuscript of:

Loudon K, Zwarenstein M, Sullivan FM, Donnan PT, Gagyor I, Hobbelen H, Althabe F, Krishnan JA & Treweek S (2017) The PRECIS-2 tool has good interrater reliability and modest discriminant validity, *Journal of Clinical Epidemiology*, 88, pp. 113-121.

DOI: 10.1016/j.jclinepi.2017.06.001

Accepted Manuscript



The PRECIS - 2 tool has good inter - rater reliability and reasonable discriminant validity

Kirsty Loudon, Merrick Zwarenstein, Frank Sullivan, Peter Donnan, Ildikó Gágyor, Hans Hobbelen, Fernando Althabe, Jerry A. Krishnan, Shaun Treweek

PII: S0895-4356(16)30601-1

DOI: 10.1016/j.jclinepi.2017.06.001

Reference: JCE 9415

To appear in: *Journal of Clinical Epidemiology*

Received Date: 26 October 2016

Revised Date: 12 May 2017

Accepted Date: 2 June 2017

Please cite this article as: Loudon K, Zwarenstein M, Sullivan F, Donnan P, Gágyor I, Hobbelen H, Althabe F, Krishnan JA, Treweek S, The PRECIS - 2 tool has good inter - rater reliability and reasonable discriminant validity, *Journal of Clinical Epidemiology* (2017), doi: 10.1016/j.jclinepi.2017.06.001.

The PRECIS - 2 tool has good inter - rater reliability and reasonable discriminant validity

Kirsty Loudon [1*], Merrick Zwarenstein [2], Frank Sullivan [3], Peter Donnan [3], Ildikó Gágyor [4], Hans Hobbelen [5], Fernando Althabe [6], Jerry A. Krishnan [7], Shaun Treweek [8]

1. NMAHP Res earch Unit, Stirling University, Stirling, FK9 4NF, UK

2. Centre for Studies in Family Medicine, Schulich School of Medicine & Dentistry, Western University, Western Centre for Public Health and Family Medicine, London, ON N6A 4K7, Canada

3. Division of Populati on Health Sciences, University of Dundee, Dundee, DD2 4BF, UK

4. Department of General practice, University Medical Center Göttingen , D - 37073 Göttingen , Germany

5. Research Group Healthy Ageing, Allied Health Care and Nursing – Hanze University Groningen, Univer sity of Applied Sciences, The Netherlands

6. Departamento de Investigación en Salud de la Madre y el Niñom, Instituto de Efectividad Clínica y Sanitaria (IECS), Buenos Aires, Argentina

7. Population Health Sciences, University of Illinois at Chicago, Chicago, I L

8. Health Services Research Unit, University of Aberdeen, Foresterhill , Aberdeen, AB25 2ZD , UK

*corresponding author

Abstract

Objective

PRECIS - 2 is a tool that could improve design insight for trialists. Our aim was to validate the PRECIS - 2 tool, unlike its predecessor, testing the discriminant validity and inter-rater reliability.

Study design and Setting

Over 80 international trialists, methodologists, clinicians and policymakers created PRECIS - 2 helping to ensure face and content validity. The inter-rater reliability of PREC IS - 2 was measured using 19 experienced trialists who used PRECIS - 2 to score a diverse sample of 15 RCT protocols. Discriminant validity was tested with two raters to independently determine if the trial protocols were more pragmatic or more explanatory, with scores from the 19 raters for the 15 trials as predictors of pragmatism.

Results

Inter-rater reliability was generally good, with seven out of nine domains having an ICC over 0.65. *Flexibility (Adherence) and Recruitment* had wide confidence intervals but raters found these difficult to rate and wanted more information. Each of the nine PRECIS - 2 domains could be used to differentiate between trials taking more pragmatic or more explanatory approaches with better than chance discrimination for all domains.

Conclusion

We have assessed the validity and reliability of PRECIS - 2. An elaboration paper and website provide guidance to help future users of the tool which is continuing to be tested by trial teams, systematic reviewers and funders.

Six key words:

Randomized controlled trials; Clinical trial methodology; Validity and reliability; Face validity Pragmatic clinical trial; Trial design.

Running title: The PRECIS - 2 tool has good inter-rater reliability and reasonable discriminant validity

[i]The PRECIS-2 tool has good inter-rater reliability and modest discriminant validity

What's new?

- The original PRECIS tool did not have its validity and reliability formally measured.
- The inter-rater reliability of PRECIS-2 was measured using 19 raters (trialists from seven countries) to score a varied sample of 15 RCT protocols.
- Inter-rater reliability was generally good, with seven of nine domains having an Intraclass Correlation Coefficient over 0.65.
- Each of the nine PRECIS-2 domains could be used to differentiate between trials taking more pragmatic or more explanatory approaches with better than chance discrimination for all domains.
- The validity and reliability of PRECIS-2 has been assessed.

Introduction

The aim of the original PRECIS (PRagmatic Explanatory Continuum Indicator Summary) tool [1,2], and of the current PRECIS-2 tool [3] is to enable trialists to match their design decisions to the intended purpose of the trial. Some trials are conducted to understand how an intervention works (explanatory or efficacy trials), whereas others are intended to inform clinical and service delivery decisions in usual healthcare settings in the real world (pragmatic or effectiveness trials) [1,2,3]. Although the original PRECIS tool (2009) [1,2] was increasingly cited it was criticised for the lack of inter-rater variability assessments and the absence of a rating scale for each domain [4,5,6,7,8]. Users also wanted further explanation on the PRECIS domains to use the tool effectively. The PRECIS-2 tool [3] published in 2015, aimed to address these demands. It was the result of

collaboration with over 80 international trialists, clinicians and policymakers from 2011 to 2014 involving a 2-round electronic Delphi, brainstorming meetings in Dundee, UK and Toronto, Canada and user testing of the PRECIS-2 tool with 19 individual trialists ranging from early career to experienced researchers [3,9,10]. This PRECIS-2 tool, like the original, was intended to be used prospectively, at the trial design stage, by a multi-disciplinary team to prompt discussion of each design choice, and thus ensure that the resulting trial design would match the intended question, whether it be pragmatic or explanatory.

The nine domains in PRECIS-2: *Eligibility, Recruitment, Setting, Organisation, Flexibility (delivery) and Flexibility (adherence), Follow up, Primary outcome* and *Primary analysis* are each scored repeatedly as the trial protocol is developed, using a scale from "1" (very explanatory) to "5" (very pragmatic). A score of "1" indicates a highly explanatory design choice for that domain, suggesting that trial domain is intending to test an intervention under idealized, tightly controlled conditions, whereas a "5" would suggest that domain is intending to be very pragmatic and test the intervention under conditions close to routine clinical care. With all nine domains scored, trialists get a visual representation of their trial's design on a wheel (Figure 1) – and can instantly see across all domains, whether the trial is more pragmatic (mostly near the rim of the wheel) or more explanatory (close to the hub or centre of the wheel).

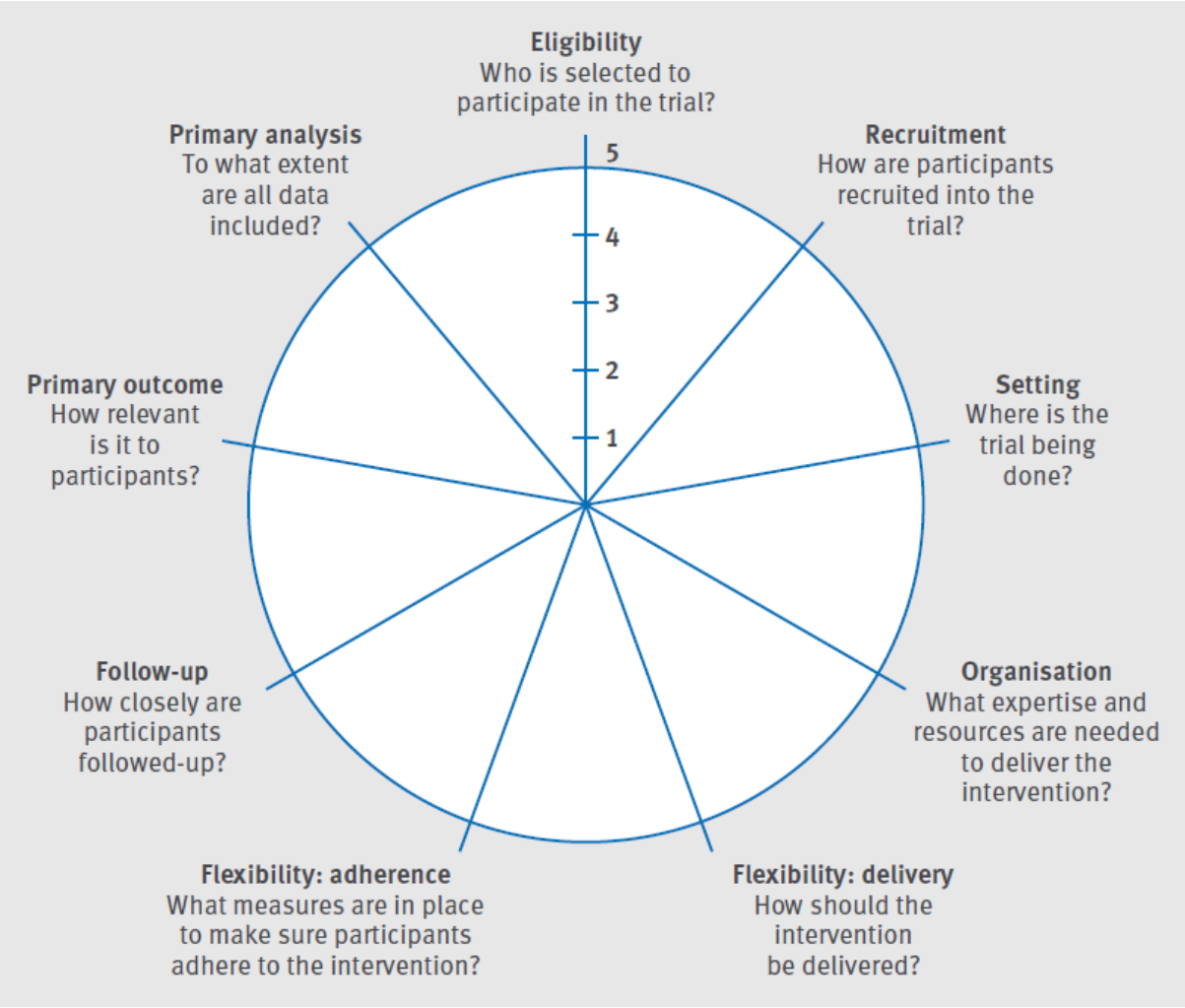


Figure 1 The PRECIS-2 wheel (reprinted with permission from [3]).

The validity and reliability of the original PRECIS tool (2009) was not formally measured [1,2]. Rather it was presented as a concept and readers were encouraged to try it out and further develop PRECIS. Some users did just this, using the original PRECIS tool [2] as it was intended prospectively at the trial design stage, for retrospective assessment in systematic reviews or to assess trials that were past the trial design stage and underway. Many had problems with inter-rater variability when multiple raters scored PRECIS domains for a trial. For example, in a systematic review Koppenhaal et al tested a modified PRECIS tool and recommended using two raters to reduce subjectivity across all domains when assessing 20 trials using two raters [4]; Riddle et al found discussion helped his team of seven raters when they used the PRECIS tool as intended to design a trial; variation in PRECIS-2

scores pre-discussion for each individual were 1.16 for the average Standard Deviation becoming 0.61 post discussion across all domains [5]. Witt in a systematic review of ten trials, had low inter-rater reliability amongst five raters using PRECIS after the first round of scoring but after a consensus meeting this improved and there was usually only one point difference (on a 5-point scale) for all domains [6]. Glasgow *et al* [7] assessed three studies with PRECIS using nine raters and found Intraclass Correlation Coefficient for domains was 0.72. While Sanchez *et al* assessed 113 trials using two raters, calculated weighted agreement scores for PRECIS ranged from 63.9 to 78.5 %, with a median of 73.9% [8]. All of these studies have been small, with few raters and/or trials being rated; the most raters used was seven but Riddle *et al* scored only one trial (their own) [5]. This work highlights that there were issues around inter-rater variability which we were keen to address in formally testing the validity and reliability of the PRECIS-2 tool.

For PRECIS-2, we achieved face validity by consulting a large number of participants and potential users of the tool in creating and developing the updated tool. The modified tool, PRECIS-2, kept the simple format but addressed weaknesses through a scoring system, domains changes and additional guidance. We felt, however, that it would be useful to assess the reliability and other aspects of validity of the new tool at the point it was developed, so that its strengths and limitations would be known early in its life. Moreover, having a validated tool might encourage more trialists to consider using the tool with their own trials. Within the timeframe of KL's PhD, it was not practical to prospectively quantitatively assess the use the PRECIS-2 tool by trial teams at the design stage of their trials (though some largely qualitative work was done [10]. To give an indication of validity and reliability we therefore decided to use the tool retrospectively on trials that had already been published, recognising that this would be a harder test for PRECIS-2 i.e. to be used by raters unfamiliar with the trials they were assessing and without discussion between raters. It would, in other words, provide a conservative estimate of validity and reliability.

The aim of the work described here is to validate the PRECIS-2 tool. To ensure PRECIS-2 could be used to design different trials by different raters on a spectrum of pragmatism, from very explanatory to very pragmatic, we tested the face validity, inter-rater reliability and discriminant validity (ability of the domains to determine pragmatism) of PRECIS-2. We believed that it was important that the participants reflect trialists who are experienced and could be future users of the PRECIS-2 tool. It was also important that the sample of trial protocols that they assessed was varied to allow the tool to be used for all trial protocol designs.

Methods

Firstly (1) we undertook a sample size calculation using the Intra-class coefficient, then (2) we selected the trials that would be used to test out PRECIS-2. This was followed by (3) pilot testing the materials and methods to make it as easy as possible for individual participants to assist with validity and reliability testing. (4) A purposive sample of trialists were then invited to participate. In this project. (5) The inter-rater variability of the nine PRECIS-2 domains was then analysed. Finally (6) statistical analysis indicated the discriminant validity of PRECIS-2 to determine pragmatism.

1. Sample size

The key requirement for assessing the reliability of PRECIS-2 was to ensure we had sufficient raters involved in testing the PRECIS-2 tool. The Intraclass Correlation Coefficient, ICC acts as a measure of inter-rater reliability (*see* 5. Statistical analysis). We were expecting an ICC near 0.7; Land and Koch view the range of 0.61 to 0.80 as “substantial agreement”. Assuming the ICC was in the region of 0.7, then 15 raters looking at 10, 15 or 20 trials would give precisions of +/- 0.20, +/- 0.17 and +/- 0.14, respectively. We aimed to give our 15 raters between 10 and 15 trials to rate.

2. Selection of the trials for assessment

We needed a broad spectrum of trials for raters to independently rate using the PRECIS-2 tool. We decided to use trial protocols because they give more detailed information on trial design than the final trial publications. We were given permission to access a database of trial protocol examples

assembled from public websites, journals, trial investigators, and industry sponsors by An-Wen Chan and Jennifer M Tetzlaff for SPIRIT – Standard Protocol Items: Recommendations for Interventional Trials [11]. The SPIRIT guidance for protocol reporting was published in 2013 (http://www.equator-network.org/reporting-guidelines/spirit-2013-statement-defining-standard-protocol-items-for-clinical-trials/) in response to poor reporting.

ST and KL independently screened the 150 SPIRIT protocols, excluding all trial protocols longer than 60 pages (approximately 10%) to reduce the burden on raters, who would have to read them. KL and ST each selected 20% of the SPIRIT database of trial protocols to include the different types of trials in terms of interventions (drug trials, therapy or educational programmes), settings (primary care, hospital and community), a range of countries, publications in different journals and, including cluster randomised and factorial designs. They agreed through discussion on a final selection of 15 trial protocols (Supp 1), published between 2008 and 2011.

3. Internal testing of the materials and procedure to guide future use

KL developed the training materials and procedures for reviewers on how to use PRECIS-2, the nine PRECIS-2 domains and descriptions and the scoring system, with examples of trials rated by PRECIS-2. These were pilot tested with four members of the PRECIS-2 development steering group and one independent primary care trialist (ST, FS, MZ, IG and KL) for clarity and ease of use.

The pilot testers used these materials to assess three contrasting trials, purposively selected to test the PRECIS-2 tool: a single centre factorial trial in the USA -the physiotherapy versus corticosteroid trial [12]; a behavioural change cluster randomised trial in India using Accredited Social Health Activists **(**ASHA) to improve maternal and neonatal health [13] and a multi-centre trial in North of

England and Scotland comparing surgical intervention with conventional medical treatment in children with recurrent sore throat (NESSTAC) trial [14] (Supp. 1).

Our test led to two changes. Firstly, to speed up the PRECIS-2 learning process for raters we produced a much shorter 3 page information sheet. Secondly, we reduced the number of PRECIS-2 tool domains from ten to nine and removed "*Organisation – comparison*". As the intention of the tool is primarily to help design trials which are useful for decision-making in usual care, our approach to PRECIS-2 was to simplify it by always drawing the comparison with existing patterns of usual care or standard of care.

4. Selection and invitation of participants

Thirty five personalised invitations (Figure 2) were sent on September 24th, 2013, to six different groups of potential raters, including: researchers who had been involved in an early stage of PRECIS-2 development (a Delphi); early user-testers who had given feedback on initial versions of PRECIS-2; individuals who had participated in brainstorming meetings; methodologists in the Cochrane Methodology Review Group, the CONSORT group, the Scottish Clinical Trials Units and the EU-funded DECIDE (Developing and Evaluating Communication Strategies to Support Informed Decisions and Practice based on Evidence) group; researchers who had worked with the original PRECIS tool; and editors of medical/trial journals. Sampling was purposive: for this retrospective assessment of trial protocols using PRECIS-2 we wanted experienced trialists and methodologists who would be able to commit the not inconsiderable time required to do the assessments.

Nineteen researchers agreed to take part, and were sent a concise PRECIS-2 training package comprising a 3-page explanation sheet, a PRECIS-2 wheel and table that could be used for scoring trial protocols. Raters were sent 5 to 15 protocols each, in batches of five, with a new batch sent on receipt of the returned previous batch; the number of protocols depended on rater preference. As this was a significant burden on time, raters were initially offered £100 as a notional payment for

about four hours work but as many raters waived payment, this enabled us to increase the financial incentive to £200 to complete the assessment of 15 trial protocols using PRECIS-2.

5. Statistical analysis of the inter-rater reliability of the nine PRECIS-2 domains

The Intraclass correlation coefficient (ICC) is a relatively simple statistical measure to assess variation and determine if raters were reaching similar decisions with regard to domain scores. As there were two variables that could affect the rating, the raters and the trial protocols, we chose the two-way random effects model where both people (PRECIS-2 raters) and measures effects (trial protocols for scoring) are treated as random variables i.e. as a random sample of all potential raters and trial protocols. To determine the effect of missing data we undertook sensitivity analysis, imputing missing data in two ways: 1) using a score of "3" - equally pragmatic/explanatory and 2) undertaking multiple sensitivity analysis (10 imputations) in which randomly-generated values of "1" to "5" were inserted if there were missing values.

6. Statistical analysis of the discriminant validity of PRECIS-2 to determine pragmatism

While interrater reliability determines the reliability of the tool in different hands, discriminant validity examines the ability of the tool itself to discriminate between pragmatic and explanatory trials. We were keen to evaluate discriminant validity by testing whether PRECIS-2 could accurately discriminate trials of varying degree of pragmatism (an ordinal variable) but discriminant validity is a binary concept, and requires a gold standard against which the performance of the instrument is compared.

Ideally, to asses discriminant validity we would have asked participants to give subjective global ratings of pragmatism of the trial after reading the trial protocol, then participants would use PRECIS-2 to rate the nine domains of PRECIS-2. However due to the already significant burden on raters this was not possible. Therefore we decided to determine discriminant validity to compare our own (ST, KL) subjective global (more pragmatic vs. more explanatory) ratings of each of the 15 trial

protocols. Two raters (KL, ST) independently used binary scores of more pragmatic = “1”, more explanatory =”0” to rate the overall pragmatism for the 15 trials. This was done by making a judgement of degree of pragmatism based on reading the trial publication, with KL and ST then reaching consensus through discussion. We used this as an implicit gold standard to compare with the median score of each domain of PRECIS 2), determined by as many as 18 raters, analysed using binary logistic regression, calculating Area under the curve (AUROC) odds (discriminant validity) - (Receiver Operating Characteristic Curve) (ROC Curve function) [10]. We used a Hosmer-Lemeshow goodness-of-fit statistic to assess calibration of the model. We saved the predicted probabilities for each domain and then using the ROC Curve function (Receiver Operating Characteristic Curve) calculated AUROC – Area under the curve. This showed us the sensitivity/specificity of the different PRECIS-2 domain variables for different cut-offs. So, using the test variable as the predicted probability (PRE_1, PRE_2 etc.) and the State variables as Pragmatism (more pragmatic, more explanatory) we calculated how good each domain in the PRECIS-2 tool is at predicting whether the trial was more, or less pragmatic (using our subjective, consensus based gold standard) (1) displaying a ROC curve with diagonal reference line and Standard error and confidence interval. SPSS (version 15.5) was used for this analysis.

Results

Participants

After 10 weeks we had received a response from 91% (32/35), with 54% (19/35) of ratings returned. The 19 raters came from seven countries – USA (8), UK (3), Canada (3), The Netherlands (2), Argentina (1), Australia (1), and Germany (1). Of these, seven of the raters scored 15, and 12 scored 10 trials (Supp 2). Six out of nineteen of the raters had assisted with the Delphi round, brainstorming or user testing and four of the nineteen raters had assisted with methodological testing of the original PRECIS tool. The remaining nine raters had not previously been involved in development of PRECIS or PRECIS-2.

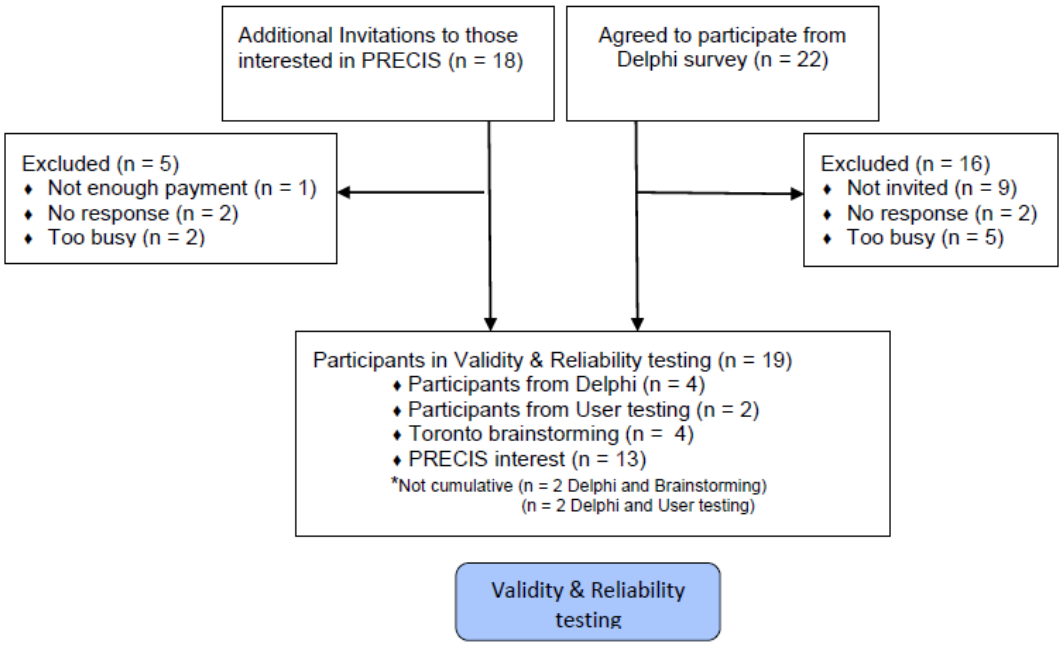


Figure 2 Flow diagram of participants in validity and reliability testing

Results of the statistical analysis of the inter-rater reliability of the nine PRECIS-2 domains

For seven out of nine domains inter-rater reliability was good or modest with ICC over 0.65 and tight confidence intervals. Two domains *Flexibility - adherence* - ICC 0.57 and Recruitment - ICC 0.60 has lower inter-rater reliability with wide confidence intervals. These results are based on the best estimate of the ICC and confidence intervals for 10 trials and 12 raters as this was closest to our sample size calculations (Table1 and 2). To assess the effect of missing data we imputed values of "3" as equally pragmatic/explanatory to indicate the uncertainty in scoring (Table 1). We also randomly imputed values of "1" to "5" (Table 2). In Table 1 we looked at the ICC for five trials scored by 18 raters, 10 trials scored by 12 raters and 15 trials scored by seven raters, generally there was not much difference for the ICC for the different batches of trials. The ICC scores for the different domains in Table 1 with imputed values of "3" compared well to the ICC scores for the PRECIS-2 domains in Table 2 for a complete set of 15 trials scored by 19 raters which included random values "1" to "5" for missing data (up to 38%).

Table 1 Overall results for Inter-rater reliability for 9 PRECIS-2 domains including sensitivity analysis

Domain	Number of trials, raters	No. Imputed values = 3 [1] (%)	Intraclass Correlation	95% Confidence Interval	
				Lower Bound	Upper Bound
Eligibility	15, 7	1 (0.74)	0.88***	0.75	0.95
	10, 12	2 (1.67)	0.89***	0.76	0.97
	5, 18	4 (4.4)	0.94***	0.81	0.99
Recruitment	15, 7	1 (0.74)	0.59**	0.18	0.84
	10, 12	2 (1.67)	0.60*	0.10	0.88
	5, 18	4 (4.4)	0.83***	0.50	0.98
Setting	15, 7	1 (0.95)	0.80***	0.60	0.92
	10, 12	2 (1.67)	0.80***	0.56	0.94

Domain	Number of trials, raters	No. Imputed values = 3 [1] (%)	Intraclass Correlation	95% Confidence Interval	
	5, 18	5 (4.4)	0.92***	0.76	0.99
Organisation	15, 7	2 (1.99)	0.72***	0.44	0.89
	10, 12	9 (7.5)	0.83***	0.61	0.95
	5, 18	6 (5.55)	0.75**	0.25	0.97
Flex Delivery	15, 7	3 (3.33)	0.74***	0.47	0.90
	10, 12	6 (6.67)	0.85***	0.67	0.96
	5, 18	6 (6.67)	0.92***	0.75	0.99
Flex Adherence	15, 5	0	0.50*	-0.06	0.81
	15, 7	8 (7.62)	0.24ns	-0.54	0.70
	10, 12	17 (15.74)	0.57*	0.04	0.88
	5, 18	18 (15.56)	0.72**	0.16	0.97
Follow-up	15, 7	1 (1.11)	0.60**	0.18	0.84
	10, 12	8 (8.89)	0.80***	0.55	0.94
	5, 18	6 (6.67)	0.85***	0.55	0.98
Primary outcome	15, 7	0	0.44ns	-0.13	0.78
	10, 12	1 (1.11)	0.66**	0.24	0.900
	5, 18	3 (3.33)	0.84***	0.54	0.98
Primary analysis	15, 7	0	0.67***	0.32	0.87
	10, 12	3 (3.33)	0.73***	0.39	0.92
	5, 18	5 (5.55)	0.83***	0.50	0.98

† PRECIS-2 score = 3 equally pragmatic/explanatory chosen as score to indicate uncertainty in scoring.

‡ Trials were scored in batches of 5, there were overall 19 raters but one rater asked for 10 trials but only scored the 2nd batch of trials so did not score the 1st five trials that 18 other raters scored.

Table 2 Rater scoring using randomly generated values (1 to 5) to impute missing data for PRECIS-2 domains using responses for 15 trials by 19 raters

Domain	% missing data*	Intraclass Correlation	95% Confidence Interval		Significance
			Lower Bound	Upper Bound	
Eligibility	33	0.84	0.69	0.94	***
Recruitment	33	0.58	0.20	0.84	**
Setting	34	0.79	0.60	0.92	***
Organisation	36	0.72	0.46	0.89	***
Flex deliv. prov.	35	0.80	0.62	0.92	***
Flex adherence	38	0.54	0.12	0.82	*
Follow-up	34	0.71	0.44	0.88	***
Primary Outcome	34	0.68	0.38	0.87	***
Primary Analysis	34	0.67	0.37	0.84	***

*approx. 93% of missing data is due to trials not being scored at all by raters

Discriminant validity results

Agreement between ST and KL on their implicit gold standard was 80% (12/15) pre-discussion, Cohen's kappa 0.59 indicating moderate agreement. The three trials where there was disagreement in assigning a trial as being pragmatic or explanatory i.e. "1" instead of "0" were resolved following discussion.

The AUROC values for determining whether a trial is more pragmatic or more explanatory for the nine PRECIS-2 domains are displayed in Table 3, these are a numerical summary of the ROC curves (Supp 3). These values have been placed in order of discriminative ability. A score of 1 would be the ideal score and indicate that a PRECIS-2 domain was perfect at discriminating between more pragmatic and more explanatory trials. Random performance, with no discriminant ability beyond chance would be 0.5. For the ROC curves ideally we would want the whole curve to be above the diagonal line. The results for all PRECIS-2 domains are greater than 0.5 although some are not significantly different from chance. *Primary outcome* is the single variable that is most likely to discriminate how pragmatic a trial is based on this data – AUROC 0.75. Then in order of discriminating between a more pragmatic and more explanatory approach: *Follow-up* 0.73*, Primary analysis* 0.72*, Flexibility (delivery)* 0.71*, Eligibility* 0.62*, Recruitment* 0.62*, Flexibility adherence* 0.60*, Setting* 0.59*, Organisation* 0.57*.*

Table 3 Discriminant validity measured using Area Under the ROC curves (AUROC)

Domains	**AUROC**	**95% Confidence intervals**
Primary Outcome	0.75	0.49-1.00
Follow-up	0.73	0.48-0.99
Primary analysis	0.72	0.45-1.00
Flexibility delivery	0.71	0.44-0.99
Eligibility	0.62	0.33-0.92
Recruitment	0.62	0.32-0.92
Flexibility adherence	0.60	0.30-0.89
Setting	0.59	0.26-0.92
Organisation	0.57	0.27-0.87

Discussion

Our reliability and validity work found that PRECIS-2 has generally good inter-rater reliability across the nine domains with 7/9 ICCs over 0.65 and modest discriminant validity with better than chance discriminant validity for 7/9 domains in comparison with our subjective global ratings of pragmatism. The two domains which were not statistically better discriminants than chance were *Flexibility Adherence* and *Recruitment,* and this is likely because both were poorly described in the trial protocols.

It is important to note that PRECIS-2 was developed to help designers of trials to match their design choices to their intended degree of pragmatism, not for retrospective assessment of trials designed by others and that the poor description of certain domains in the protocols is therefore not relevant to the main use of this tool. Trial design teams would be much more familiar with the intricate details of each domain for a trial they were currently designing, than would be our assessors, who were rating trials they did not design. Since it was not logistically possible to work with large enough trial design teams, during their design process, to evaluate performance of the PRECIS-2 tool we constructed an artificial situation which would be expected to underestimate both inter-rater reliability and discriminant validity. It is encouraging that inter-rater reliability was still good, and discriminant validity modest, even when PRECIS-2 was used by researchers unconnected with the trial being scored.

Sensitivity analysis indicated there was no obvious difference in the scores between individuals with regard to country, research area, or profession who completed 10 or 15 trials. The main reasons for not assessing all 15 trial protocols using PRECIS-2 was lack of time.

Strengths and Limitations

This is the first validation and estimation of reliability of the PRECIS 2 tool, which was never done for the original PRECIS tool [1,2]. We involved raters who reflect the target group of experienced trialists who could be future users of the PRECIS-2 tool. The sample of trial protocols that they assessed was

varied, indicating the tool can be used for diverse trial designs. Our assumption that raters have limited time turned out to be correct and we were unable to get all 15 raters to complete all 15 trials. Also, some raters did not score particular domains giving various reasons, for instance *Eligibility*, *Organisation*, *Flexibility (delivery)* were not scored on medical and surgical trials due to lack of expertise in the area (e. g. physiotherapist) who reported: "No entry, obviously no content knowledge on this one. Too far afield of my content to judge." Other examples of explanations for missing ratings included: for *Recruitment, Organisation*, "*inadequate information*", for *Setting* "*unclear to judge*", for *Flexibility (adherence)* "*although mentioned in most study protocols in protocol publications often not enough information is given to judge on this*". Many of the imputed values are needed due to "*lack of time"* or whole trials not being scored by individual raters. Comparing the values for the different batches of trials and indeed for a complete set there is little difference in the values for the ICC thus the impact of the missing data on our assessment of interrater reliability was not serious.

We are confident in the ability of PRECIS-2 to pick out trials taking different design approaches, and that different raters looking at the same trials come to similar conclusions. We have used the feedback from participants in validity and reliability testing of PRECIS-2 to add additional information to the guidance for users on PRECIS-2 [19] and the PRECIS-2 website http://www.precis-2.org/.

Asking raters to retrospectively score trial protocols is perhaps a rather artificial way of using the PRECIS-2 tool when we suggest that it is used at the design stage by the team designing the trial. While a prospective study is conceivable, the time that would be needed to do it was prohibitive. The results presented here could be considered as a worst-case test of the tool given that there was sometime inadequate reporting of trial information that was relevant to assessing the PRECIS-2 domains. This was also one of the reasons for a high percentage of missing data (in addition to the being unable to get a fully completed assessment of 15 trial protocols by 15 raters). This highlights a need to adhere to the SPIRIT statement [11] to improve reporting of information on design and

methods and in conjunction the CONSORT statement for pragmatic trials [15] and also the full CONSORT statement for randomised trials (http://www.consort-statement.org) as this data is useful to understand a trial's design and assess applicability of trials.

Discriminant validity could only be tested using a global, dichotomous 'more pragmatic' or 'more explanatory' assessment based on the independent judgement of ST and KL rather than on the judgements of more raters. This was to reduce workload but may have an impact on the results. Clearly, judgements based on the opinions of more raters would have been preferable but we were very wary of the burden we were already placing on raters.

Conclusion

The validity and reliability of the PRECIS-2 tool is modest, even when tested retrospectively using individuals unconnected with the trials being scored. PRECIS-2 is a relatively simple, visual tool that can be used to focus the trial team's discussion on the match between their design decisions and the needs of those for whom the results are intended, and this perhaps helps to explain why PRECIS-2 is already proving useful in pragmatic trial design [16]. We believe it could also be helpful in reducing research waste [17] by helping trialists to consider the consequences of their design decisions on the usefulness of the trial results to their intended users.

Acknowledgments

This work was supported by the Chief Scientist Office (CSO) of Scotland grant CZH/4/773, the UK Medical Research Council and the University of Dundee work through the provision of a stipend for KL and from the Health Services Research Unit at the University of Aberdeen, which is core funded by the CSO of the Scottish Government Health Directories. We are grateful to all the participants who assisted in this study: F Althabe, A-W Chan, D Altman, D Bratton, E Brass, M Campbell, G Forbes, B Gaglio, R Glasgow, H Hobbelen, S Hopewell, J Krishnan, D Riddle, J Segal, D Steinfort, P Tugwell, SN Van der Veer, VA. Welch, C Witt.

References

[1] Thorpe KE, Zwarenstein M, Oxman AD, Treweek S, Furberg CD, Altman DG, et al. A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers. CMAJ 2009;180:E47-57.

[2] Thorpe KE, Zwarenstein M, Oxman AD, Treweek S, Furberg CD, Altman DG, et al. A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers. J Clin Epidemiol. 2009;62:464-75.

[3] Loudon K, Treweek S, Sullivan F, Donnan P, Thorpe K, Zwarenstein M. The PRECIS-2 tool: Designing trials that are fit for purpose. BMJ. 2015;350.

[4] Koppenaal T, Linmans J, Knottnerus JA, Spigt M. Pragmatic vs. explanatory: an adaptation of the PRECIS tool helps to judge the applicability of systematic reviews for daily practice. J Clin Epidemiol. 2011;64:1095-101.

[5] Riddle DL, Johnson RE, Jensen MP, Keefe FJ, Kroenke K, Bair MJ, et al. The Pragmatic-Explanatory Continuum Indicator Summary (PRECIS) instrument was useful for refining a randomized trial design: experiences from an investigative team. J Clin Epidemiol. 2010;63:1271-5.

[6] Witt CM, Manheimer E, Hammerschlag R, Ludtke R, Lao LX, Tunis SR, et al. How Well Do Randomized Trials Inform Decision Making: Systematic Review Using Comparative Effectiveness Research Measures on Acupuncture for Back Pain. PLoS One. 2012;7.

[7] Glasgow RE, Gaglio B, Bennett G, Jerome GJ, Yeh HC, Sarwer DB, et al. Applying the PRECIS Criteria to Describe Three Effectiveness Trials of Weight Loss in Obese Patients with Comorbid Conditions. Health Serv Res. 2011.

[8] Sanchez MA, Rabin BA, Gaglio B, Henton M, Elzarrad MK, Purcell P, et al. A systematic review of eHealth cancer prevention and control interventions: new technology, same methods and designs? Transl Behav Med. 2013;3:392-401.

[9] Loudon K, Zwarenstein M, Sullivan F, et al. Making clinical trials more relevant: improving and validating the PRECIS tool for matching trial design decisions to trial purpose. *Trials* 2013;14:115. doi: 10.1186/1745-6215-14-115 [published Online First: 2013/06/21]

[10] Loudon K. Making trials matter: providing an empirical basis for the selection of pragmatic design choices in clinical trials [Ph.D]. Dundee: University of Dundee; 2015.

[11] Chan AW, Tetzlaff JM, Altman DG, Dickersin K, Moher D. SPIRIT 2013: new guidance for content of clinical trial protocols. Lancet. 2013;381:91-2.

[12] Rhon DI, Boyles RE, Cleland JA, Brown DL. A manual physical therapy approach versus subacromial corticosteroid injection for treatment of shoulder impingement syndrome: a protocol for a randomised clinical trial. BMJ Open. 2011;1:e000137.

[13] Tripathy P, Nair N, Mahapatra R, Rath S, Gope RK, Bajpai A, et al. Community mobilisation with women's groups facilitated by Accredited Social Health Activists (ASHAs) to improve maternal and newborn health in underserved areas of Jharkhand and Orissa: study protocol for a cluster-randomised controlled trial. Trials. 2011;12:182.

[14] Bond J, Wilson J, Eccles M, Vanoli A, Steen N, Clarke R, et al. Protocol for north of England and Scotland study of tonsillectomy and adeno-tonsillectomy in children (NESSTAC). A pragmatic randomised controlled trial comparing surgical intervention with conventional medical treatment in children with recurrent sore throats. BMC Ear Nose Throat Disord. 2006;6:13.

[15] Zwarenstein M, Treweek S, Gagnier JJ, Altman DG, Tunis S, Haynes B, et al. Improving the reporting of pragmatic trials: an extension of the CONSORT statement. BMJ. 2008;337:a2390.

[16] Ford I, Norrie J. Pragmatic Trials. N Engl J Med. 2016;375:454-63.

[17] Moher D, Glasziou P, Chalmers I, Nasser M, Bossuyt PM, Korevaar DA, et al. Increasing value and reducing waste in biomedical research: who's listening? Lancet. 2016;387:1573-86.