The PRECIS-2 tool has good inter-rater reliability but lower 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 was to enable trialists match their design decisions to the purpose of the trial. Some trials are intended to understand how an intervention works (explanatory or efficacy trial), whereas others are intended to inform clinical decisions in routine healthcare settings (pragmatic or effectiveness trial) [1, 2]. The original PRECIS tool (2009) [1,2] was being increasingly cited but methodological work indicated that there were issues around inter-rater variability - with no official rating scale PRECIS users were creating their own [4,5,6,7,8]. There was also discussion around the included domains, with users needing further explanation on understanding the PRECIS domains to use the tool effectively. The PRECIS-2 tool [3] published in 2015, 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 international trialists, on a one to one basis (in person or in Skype) using a wide range of trialists from early career to experienced researchers [3, 13]. This PRECIS-2 tool, like the original, was intended to be used prospectively, at the trial design stage, by a multi-disciplinary team to facilitate discussion and ensure the intervention would be tested as intended and give results to answer the research question in hand.

The nine domains in PRECIS-2: *Eligibility, Recruitment, Setting, Organisation, Flexibility (delivery) and Flexibility (adherence), Follow up, Primary outcome* and *Primary analysis* are scored using a scale of "1" to "5" with "1" very explanatory to "5" more pragmatic to encourage comparison to usual care conditions. A score of "1" indicates a highly explanatory trial under tightly controlled conditions, whereas a "5" would be very pragmatic and test the intervention under conditions close to routine clinical care. Trialists get a visual representation of their planned trial's design on a wheel (Figure 1) – if a trial is highly pragmatic for a domain, that domain would be scored a "5" on the rim of the wheel and if completely explanatory, it would be scored close the hub or centre of the wheel.

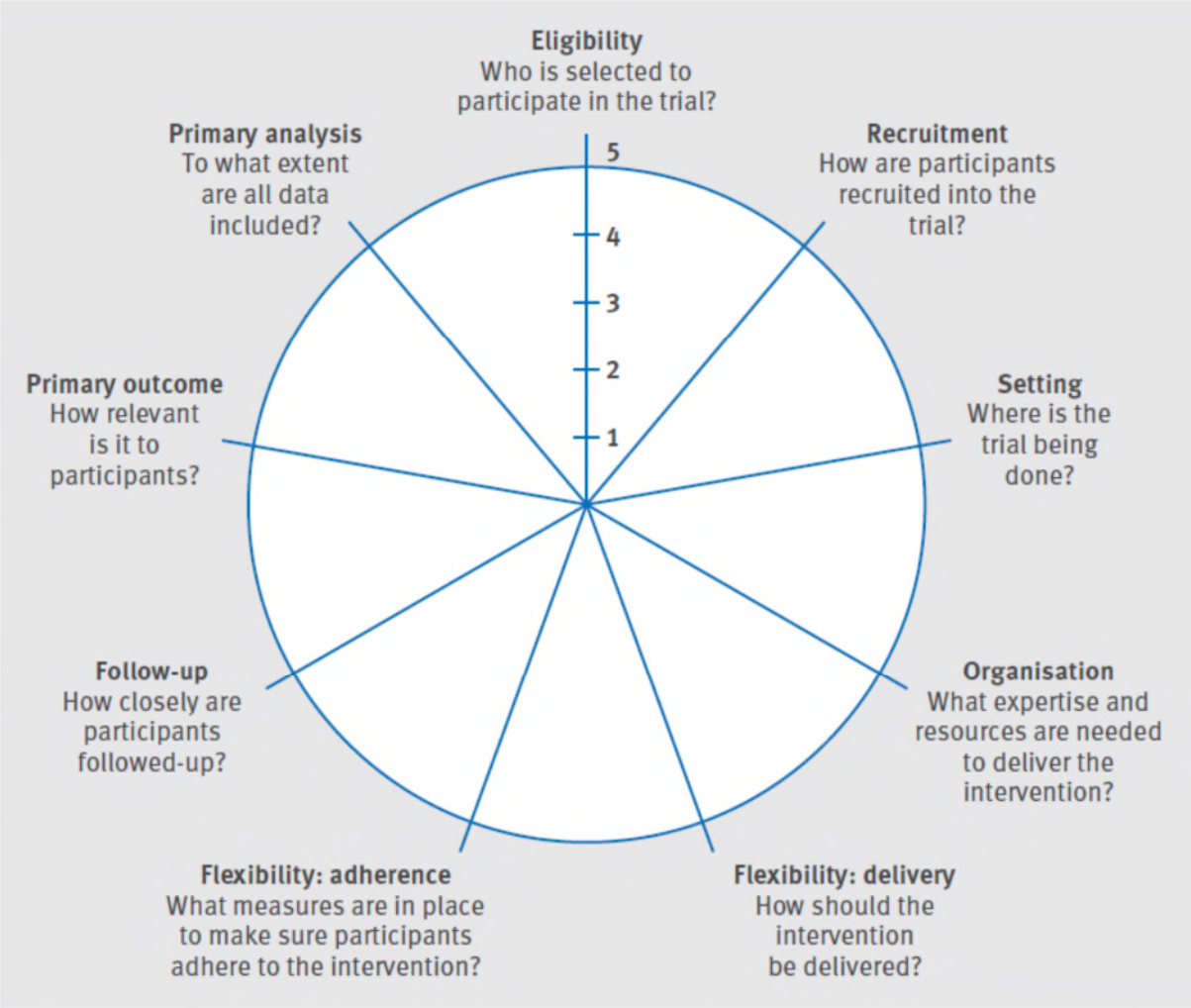

Figure 1 The PRECIS-2 wheel (reprinted with permission from [3]).

The validity and reliability of the original PRECIS tool (2009) was not formally measured [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 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 they looked at 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 [13]). 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. It would, in other words, provide a conservative estimate of validity and reliability.

The aim of the work described here was to validate the PRECIS-2 tool testing the discriminant validity and inter-rater reliability of the modified PRECIS-2 tool when used by experienced trialists and methodologists assessing a diverse mix of trials.

Methods

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.

To guide the validity and reliability testing of PRECIS-2 we used six stages. Firstly we undertook a sample size calculation using the Intra-class coefficient, then we selected the trials that would be used to test out PRECIS-2. This was followed by pilot testing the materials and methods to make it as easy as possible for individual participants to assist with validity and reliability testing. Using purposive sampling trialists we then invited to participate in this project. The inter-rater variability of the nine PRECIS-2 domains was then analysed. Finally statistical analysis indicated the discriminant validity of PRECIS-2 to determine pragmatism. We will now describe in detail this six stage process.

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; we wanted to find out if there was consistency between raters using the PRECIS-2 tool to score the nine domains. We used the Intraclass Correlation Coefficient, ICC to measure PRECIS-2's inter-rater reliability for all nine domains (*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. We measured ICC using SPSS (two-way random effects model).

2. Selection of the trials for raters to assess using the PRECIS-2 tool

We needed a broad spectrum of trials for raters to independently rate using the PRECIS-2 tool. We decided to use trial protocols because, being unrestricted on space, they give more detail on trial design information 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 [9]. 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, automatically excluding all trial protocols longer than 60 pages (approximately 10%). This was to reduce the burden on raters, who would have to read the protocols. KL and ST aimed to initially select a representative sample of 20% of the SPIRIT database of trial protocols which would exemplify the different types of trials. Some were drug trials, others therapy or educational programmes in a variety of settings, countries and published in different journals with different designs, including cluster randomised and factorial analysis design which had not been considered by the original PRECIS team ST and KL independently selected 30 trials each which they thought provided a good spread of trial feature. Combining these gave an aggregate sample of 34 different trials with 93% agreement (28/30) in trials selected. Since we only needed 15 trials for our sample size, ST and KL then agreed through discussion a final selection of 15 trial protocols from the 34 (Supp 1), ensuring that the final sample included a range of countries, number of centres, type of intervention, length of protocol (number of pages) and number of participants. The 15 articles were published between 2008 and 2011.

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

KL developed the initial materials and procedures and five people comprising four members of the PRECIS-2 development steering group and one visiting primary care trialist from Germany (ST, FS, MZ, IG and KL) piloted these for clarity, typos and ease of use. The training materials included information 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.

We 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 [10]; a behavioural change cluster randomised trial in India using Accredited Social Health Activists **(**ASHA) to improve maternal and neonatal health [11] 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 [12] (Supp. 1).

Our test led to two changes, one that impacted on making it easier for the raters participating in the PRECIS-2 validity and reliability testing and the other that was had not been picked up earlier by the steering group and in User testing with 13 models of the developing PRECIS-2 tool. Firstly, to speed up the PRECIS-2 learning process for raters we produced a much shorter 3 page information sheet. Secondly, the final version of PRECIS-2 for validity and reliability testing was created through pilot testing with the steering group and a visiting academic GP who was a trialists. This resulted in the reduction of 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; 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 the initial 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

As described previously, we chose the Intraclass correlation coefficient (ICC) as it was a relatively simple statistical measure to assess variation and determine if raters were using the information for domain rating of PRECIS-2 to reach similar decisions with regard to domain scores. As there were two variables that could affect the rating, the people 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

Discriminant validity is a type of construct validity. We wanted to determine if PRECIS-2 could discriminate between pragmatic and explanatory trials and therefore had discriminant validity. While we were keen to evaluate discriminant validity by testing whether PRECIS-2 could accurately discriminate trials of varying degree of pragmatism, discriminant validity is binary not a continuum (as is the case with pragmatism), therefore we used a global rating of whether a trial took a design approach that was a) more pragmatic or b) more explanatory.

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 (our implicit gold standard) against the median score of each domains of PRECIS 2 determined by as many as 18 raters, using binary logistic regression.

KL determined if PRECIS-2 could accurately discriminate trials of varying pragmatism as judged by the subjective rating by calculating Area under the curve (AUROC) odds (discriminant validity) - (Receiver Operating Characteristic Curve) (ROC Curve function) [13] 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.

Considering each trial, we used the values of pragmatism as decided by KL and ST (more pragmatic, more explanatory), and the PRECIS-2 domains as predictors of pragmatism: *Eligibility, Recruitment, Setting, Organization, Flexibility of Delivery, Flexibility of Adherence, Follow up, Primary outcome* and *Primary Analysis*. 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 the degree of pragmatism (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 inviting 35 experienced trialists and methodologists to participate in the inter-rater reliability and discriminant validity work, by the 3rd December 2013 we had a response rate of 91% (32/35) and 54% participation rate (19/35) with regards to returned scores. 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, while 12 scored 10 trials. 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.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 completing all 15 was lack of time.

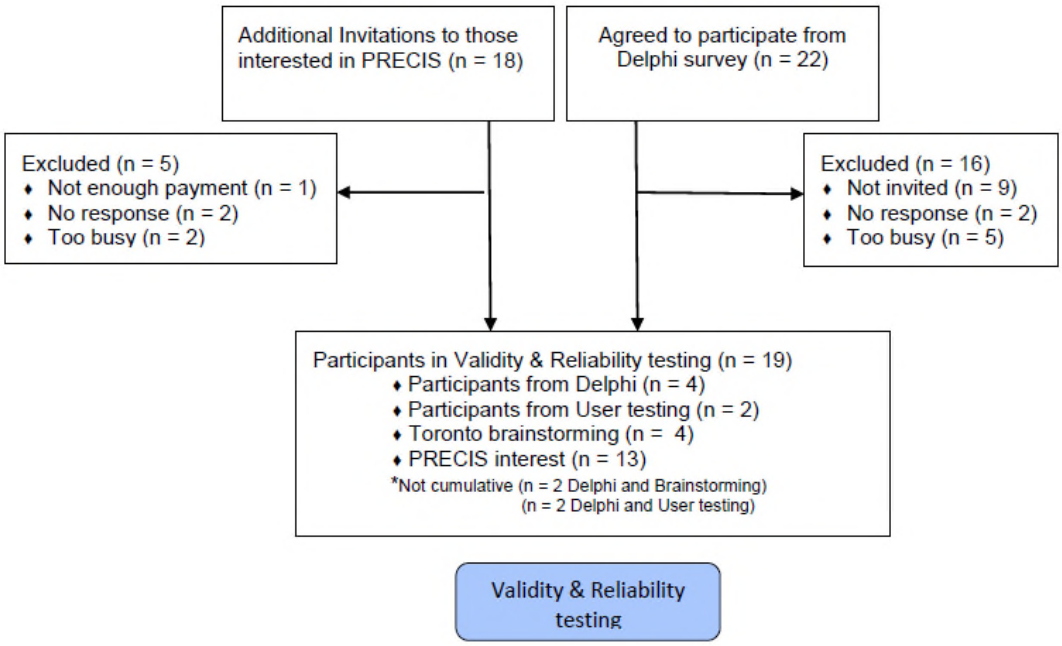

Figure 2 Flow diagram of participants in validity and reliability testing

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

Overall results for Inter-rater reliability for the nine PRECIS-2 domains including sensitivity analysis indicated that all of the domains except for one (*Flexibility - adherence)* were reasonable or good 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 used sensitivity analysis in two ways imputing values of "3" as equally pragmatic/explanatory to indicate the uncertainty in scoring (Table 1) and randomly imputing 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 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 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
	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

Domain	Number of trials, raters	No. Imputed values = 3 [1] (%)	Intraclass Correlation	95% Confidence Interval	
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 fill 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

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 2). 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 determination 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.

Agreement for global design approach scores comparing ST and KL 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.

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 reasonable discriminant validity indicating better than chance relationships for seven domains with subjective global ratings of pragmatism. The two domains which were not statistically better discriminants than chance were *Flexibility Adherence* and *Recruitment* but both were often poorly reported in the trial protocols. This lead to uncertainty in both the global degree of pragmatism assessments generated by ST and KL, as well as the assessments made by raters.

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. However it was not logistically possible to work with large enough trial design teams, during their design process. We have thus constructed an artificial situation which would be expected to underestimate the ICC and validity, since designers would be familiar with intricate

details of each domain, than the assessors we recruited for trials they did not design. It is encouraging that inter-rater reliability was still good, and discriminant validity reasonable, even when PRECIS-2 was used by researchers unconnected with the trial being scored.

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 due to lack of expertise in the area (e.g. physiotherapist) "*No entry, obviously no content knowledge on this one. Too far afield of my content to judge." Recruitment, Organisation* due to "*inadequate information*", *Setting* "*unclear to judge*", *Flexibility (adherence) "although mentioned in most study protocols in protocol publications often not enough information is given to judge on this*" and *Primary Analysis* due again to lack of information in trial protocol publications. Many of the imputed values are due to "*lack of time"* and 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 was less than might have been anticipated.

Despite this 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.

Our decision to exclude protocols over 60 pages was perhaps more likely to exclude more explanatory trials than pragmatic ones. Including very long protocols, however, would have made it even harder to get raters to score protocols. With hindsight it might have reduced any potential bias

towards rating trials as pragmatic if we had added additional explanatory trials to our selection to give raters a broader selection of trial protocols on the explanatory to pragmatic continuum.

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 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. This highlights a need to adhere to the SPIRIT statement [9] to improve reporting of information on design and methods and in conjunction the CONSORT statement for pragmatic trials [14] 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.

The discriminant validity work was limited by not asking each of the 19 raters involved in the project to assess whether a trial was pragmatic or explanatory. This was to reduce workload but may have an impact on the results. 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. 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 reasonable, even when tested retrospectively using individuals unconnected with the trials being scored. The discriminant validity is better than chance

for all nine domains, though it varies from very good for *Primary outcome,* which was the most discriminatory in indicating how pragmatic or explanatory a trial was, to 0.59 for *Organisation*. 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 [15]. We believe it could also be helpful in reducing research waste [16] 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] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.