Making trials matter: providing an empirical basis for the selection of pragmatic design choices in clinical trials

Loudon, Kirstine

Award date: 2015

Awarding institution: University of Dundee

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
PRECIS-2
Making trials matter: providing an empirical basis for the selection of pragmatic design choices in clinical trials

Kirsty Loudon

In support of the degree of Doctor in Philosophy

2015

University of Dundee
Bradford Hill, in the 11th edition of his *Principles of Medical Statistics*, wrote:

... at its best ... a trial shows what can be accomplished with a medicine under careful observation and certain restricted conditions. The same results will not invariably or necessarily be observed when the medicine passes into general use; but the trial has at the least provided knowledge which the physician can adapt to the individual patient [1]

This PhD thesis is dedicated to Dave Sackett (1934 -2015)
It was a huge honour and privilege to continue the work that he started and I will always be greatly touched by his pleasure on hearing that we had “succeeded in bringing PRECIS to the next generation”.

Kirsty Loudon
# Table of Contents

Table of Contents ........................................................................................................................................... 3
Tables and Figures ............................................................................................................................................... 14
Glossary of Terms and Abbreviations .................................................................................................................. 18
Acknowledgments ................................................................................................................................................. 20
Declaration ............................................................................................................................................................ 21
Summary of Contents ............................................................................................................................................ 22
Chapter 1: Introduction ......................................................................................................................................... 24
  History of clinical trials ....................................................................................................................................... 24
  Clinical trials development ................................................................................................................................. 24
  Applicability ...................................................................................................................................................... 25
  Explanatory and pragmatic attitudes .................................................................................................................. 27
  PRagmatic Explanatory Continuum Indicator Summary - PRECIS ..................................................................... 29
  Alternatives to PRECIS ..................................................................................................................................... 30
  Pragmatic trials and internal validity .................................................................................................................. 31
Chapter 2: Systematic review of methodological work on PRECIS ................................................................................. 34
  Background ...................................................................................................................................................... 34
  Aim .................................................................................................................................................................... 34
  Objectives ........................................................................................................................................................ 35
  Methods ............................................................................................................................................................ 35
    Eligibility Criteria .......................................................................................................................................... 35
    Information Sources ...................................................................................................................................... 35
    Search ........................................................................................................................................................... 35
    Study Selection ............................................................................................................................................ 35
    Data collection process ................................................................................................................................. 36
    Data items .................................................................................................................................................... 36
    Synthesis of results ....................................................................................................................................... 36
  Consolidation of the results ............................................................................................................................... 38
    Using PRECIS at trial design stage ................................................................................................................ 39
    Using PRECIS retrospectively ......................................................................................................................... 40
    Using PRECIS with systematic reviews ........................................................................................................ 42
    Use of PRECIS with a non-randomised study ............................................................................................... 45
  Summary of methodological work on PRECIS .................................................................................................... 46
    Strengths and Limitations of the systematic review ...................................................................................... 54
    Strengths and weaknesses of the first PRECIS tool ....................................................................................... 54
  Conclusions and Implication of findings ............................................................................................................. 55
Chapter 3: Critique of potential methods to develop PRECIS-2

Introduction ........................................................................................................ 56
Aims and Objectives .......................................................................................... 56
Aims ..................................................................................................................... 56
Objectives ........................................................................................................ 56
Data collection .................................................................................................. 57
Opinion seeking ............................................................................................... 57
Participants ........................................................................................................ 57
Possible approaches ......................................................................................... 58
Option 1: Interviews ......................................................................................... 58
General ............................................................................................................. 58
Pros .................................................................................................................. 59
Cons .................................................................................................................. 60
Conclusions ..................................................................................................... 61
Option 2: Focus groups .................................................................................... 61
General ............................................................................................................. 61
Pros .................................................................................................................. 62
Cons .................................................................................................................. 63
Conclusions ..................................................................................................... 64
Option 3: Brainstorming .................................................................................. 64
General ............................................................................................................. 64
Pros .................................................................................................................. 65
Cons .................................................................................................................. 65
Conclusion ....................................................................................................... 65
Option 4: Nominal group technique ............................................................... 66
General ............................................................................................................. 66
Pros .................................................................................................................. 67
Cons .................................................................................................................. 67
Conclusions ..................................................................................................... 67
Option 5: Delphi process ................................................................................ 67
General ............................................................................................................. 67
Pros .................................................................................................................. 69
Cons .................................................................................................................. 69
Conclusion ....................................................................................................... 69
Option 6 - User testing ..................................................................................... 70
General ............................................................................................................. 70
Pros .................................................................................................................. 71
Chapter 5: Brainstorming

First brainstorming meeting

Aims

Objectives

Methods

Participants

Pre-meeting preparation

Protocol for the meeting

Facilitators

Analysis

Results

Participant attendance

Initial impressions

How easy is PRECIS to use?

Training Package

Scoring

Alternative views

Cumulative score

Design

No domain information

Missing domains

Eligibility criteria

Multi-centre trials

Flexibility of experimental and comparison interventions and Practitioner expertise
(experimental and comparison)

Follow-up intensity

Participant Compliance with prescribed intervention

Outcomes

Primary Analysis

Weighting

Summary

Other uses for PRECIS

Brainstorming meeting Participant interactions

Conclusions: proposed changes to PRECIS

Conclusion
Chapter 7B: Results of the main study for validity and reliability testing

Introduction .................................................................................................................. 228
Results ......................................................................................................................... 228
Participants .................................................................................................................. 228
  Exclusion criteria ....................................................................................................... 232
Validity and reliability deadlines for participants ...................................................... 235
Validity and reliability incentives for participants ..................................................... 235
Recruitment rate for the validity and reliability study ............................................... 235
Time to rate batch of five trials .................................................................................. 236
Results format for PRECIS-2 scores ......................................................................... 236
Duplicate data entry ................................................................................................... 240
Range of scores .......................................................................................................... 240
Inter-rater reliability .................................................................................................... 240
  Reasons given for missing data ................................................................................ 241
Discriminant validity results (Method 1) .................................................................... 243
Discriminant validity results (Method 2) .................................................................... 249
Discussion .................................................................................................................... 250
Rater differences ........................................................................................................ 251
Trial differences .......................................................................................................... 252
Domain descriptions ................................................................................................... 253
Insufficient information .............................................................................................. 253
Possible explanation for domain scoring difficulties ................................................ 254
Comparators and multiple interventions ..................................................................... 256
Sensitivity analyses ...................................................................................................... 257
Usual care .................................................................................................................... 258
Tables or PRECIS-2 wheels ....................................................................................... 258
Discriminant validity results (Method 1) .................................................................... 258
Discriminant validity results (Method 2) .................................................................... 259
Discriminant validity methodological limitations ....................................................... 260
Validity and reliability study strengths ....................................................................... 261
Validity and reliability study limitations ..................................................................... 262
Quality assurance ................................................................. 292
Analysis ............................................................................. 292
Results .................................................................................. 294
Method 1- Cardiovascular trials .................................................. 294
Objective 1: Matched explanatory and pragmatic trials .................... 294
Objective 2, 3 and 4: Measuring risk of bias to assess internal validity, PRECIS-2 to assess external validity in a set of explanatory and pragmatic trials and comparing effect estimates .......................................................... 294
Method 2 - Cochrane hypertension review ....................................... 295
Objective 2, 3 and 4: Measuring risk of bias to assess internal validity, PRECIS-2 to assess external validity and comparing effect estimates .......................................................... 295
Quality assurance of data extracted for Method 1 and 2 ....................... 298
Discussion ............................................................................. 306
Objective 1: Matched explanatory and pragmatic trials .................... 306
OBJECTIVE 2, 3 and 4: Measuring risk of bias to assess internal validity, PRECIS-2 to assess external validity in a set of explanatory and pragmatic trials and comparing effect estimates ............................. 306
Cochrane review ..................................................................... 307
Overall thoughts on internal and external validity ............................ 308
Poor Reporting ....................................................................... 311
Quality assurance of internal and external validity assessments of trials .......................................................... 312
Study Strengths and Weaknesses ................................................. 313
Conclusions ........................................................................... 314
Chapter 9: APT: Applying PRECIS-2 to Primary Care Trials ............... 316
Pilot and two Case studies .......................................................... 316
Background ............................................................................ 316
Aims and Objectives .................................................................. 317
Aims ..................................................................................... 317
Objectives ............................................................................. 317
Methods ................................................................................. 318
Study committee ...................................................................... 319
Participants ............................................................................ 319
Statistical Methodology and Analysis ............................................. 319
Ethics .................................................................................... 319
Data handling and record keeping ................................................ 320
Record Retention and Archiving ................................................... 320
Pilot study – COPERS ............................................................... 321
Aim ....................................................................................... 321
Objectives ............................................................................. 321
APT Discussion .................................................................................................................. 360
APT study strengths and limitations .................................................................................. 362
Conclusions .................................................................................................................... 362
Chapter 10: Discussion and Conclusion ......................................................................... 364
Overall discussion ........................................................................................................... 364
PRECIS-2 domains in context of the wider research world .............................................. 367
Trial efficiency .................................................................................................................. 367
Research and Implementation ......................................................................................... 367
Future work ...................................................................................................................... 373
Using PRECIS-2 at trial design stage .............................................................................. 373
Pre-funding ....................................................................................................................... 373
Post-funding ..................................................................................................................... 374
PRECIS-2 website and pragmatic trial database .............................................................. 375
Use of PRECIS-2 in trial publications ............................................................................. 376
Using PRECIS-2 with systematic reviews ..................................................................... 377
Internal and external validity .......................................................................................... 377
Teaching PRECIS-2 in research courses ......................................................................... 378
Strengths of the thesis methodology .............................................................................. 379
Weaknesses of the thesis methodology .......................................................................... 379
Limitations of the final PRECIS-2 tool .......................................................................... 380
Conclusion ....................................................................................................................... 381
Publications and Presentations ......................................................................................... 382
References ....................................................................................................................... 384
Tables and Figures

Table 1 Improving the reporting of Pragmatic trials: an extension of the CONSORT statement [32] .......................................................................................................................................................... 28

Table 2 PRECIS – Pragmatic-Explanatory Continuum Indicator Summary .............................................. 30

Table 3 Study design of methodological studies on PRECIS ........................................................................ 37

Table 4 Summary of existing work with PRECIS involving 15 groups in chronological order ........... 48

Table 5 Questions for second round of Delphi to improve PRECIS .......................................................... 86

Table 6 Round 1 Delphi reminder schedules for recruitment by e-mail by stage of invitation .... 87

Table 7 Round 2 modified Delphi reminder schedules for recruitment by e-mail by stage of invitation ........................................... 88

Table 8 Demographics of PRECIS participants in Delphi ........................................................................ 90

Figure 1 PRECIS logo for electronic questionnaire .................................................................................. 83

Figure 2 QUORUM flow diagram for methodological articles on PRECIS ............................................. 37

Figure 3 The relative strengths of focus groups, Brainstorming, NGT and Delphi (from Gallagher et al [85]) ................................................................................................................................................. 63

Figure 4 Flow diagram of research methods during PRECIS-2 development ........................................ 74

Figure 5 PRECIS logo for electronic questionnaire .................................................................................. 83

Figure 6 Flow diagram of participants in 3 phases ...................................................................................... 94

Figure 7 Question 2 results for second round of Delphi .......................................................................... 106

Figure 8 Question 2 results for second round of Delphi .......................................................................... 106

Figure 9 The original PRagmatic Explanatory Continuum Indicator Summary from Thorpe et al 2009 [48, 49] .................................................................................................................................................. 127

Figure 10 Question 5 results for second round of Delphi ........................................................................ 107

Figure 11 Flow diagram of participants in 3 phases ...................................................................................... 94

Figure 12 Question 2 results for second round of Delphi ........................................................................ 106

Table 9 Round 1 modified Delphi reminder schedules for recruitment by e-mail by stage of invitation .................................................................................................................................................. 94

Table 10 Round 1 Questions and results for 34 participants ...................................................................... 96

Table 11 Scoring System suggestions for PRECIS-2 ................................................................................. 98

Table 12 Ten additional domain suggestions for PRECIS-2 ................................................................... 99

Table 13 Qualitative feedback from Delphi participants on question 9 on PRECIS (15 comments in no particular order) ................................................................................................................................................. 102

Table 14 Round 1 Qualitative feedback from Delphi participants on question 10 on PRECIS ... 103

Table 15 Round 2 modified Delphi number recruited by e-mail by stage of invitation ......................... 104

Table 16 Round 2 Questions and results of the Delphi .............................................................................. 105

Table 17 Combined results from Round 1 and 2 of modified Delphi ......................................................... 111

Table 18 Participants invited to participate in first brainstorming meeting ............................................. 128

Table 19 Participants who attended the first brainstorming meeting on PRECIS ..................................... 129

Table 20 Ease of using the PRECIS tool .................................................................................................. 130

Table 21 Possible PRECIS modifications ................................................................................................. 138
Table 22 Participants who accepted invitation to participate in 2nd brainstorming meeting... 147
Figure 10 Number of participants involved in second brainstorming meeting in Toronto ...... 149
Table 23 Version of PRECIS-2 domains that arose from post Theta brainstorming meeting..... 158
Table 24 Participants invited to participate in third brainstorming meeting in Dundee........ 164
Figure 11 PRECIS 2 model post Brainstorming meeting in Toronto March 2013.................. 166
Table 25 Journals that have published methodological work on PRECIS.......................... 172
Table 26 Twelve Participants in User testing who had contributed in the Delphi on PRECIS-2 . 180
Table 27 Seven UK Participants in User testing who had not been involved in the Delphi...... 182
Figure 12 Difficulty in using PRECIS-2 for User testers.................................................. 183
Table 28 Summary of key barriers and facilitators to the use of PRECIS-2.......................... 185
Figure 13 Scoring counts for up to 12 raters using the eight domain pre-publication version of
PRECIS with the ICD, CRASH and BREECH trials .......................................................... 197
Table 29 Trial protocols used for Validity and Reliability study ........................................ 201
Table 30 Ten PRECIS-2 domains for pilot of validity and reliability testing ........................ 207
Figure 14 PRECIS-2 scoring system.................................................................................. 208
Figure 15 Steps 1-4 on how to use PRECIS-2.................................................................. 208
Table 31 Corticosteroid protocol no. 3 (from 15 selected for validity and reliability study)..... 210
Table 32 ASHA protocol no. 1 [154][from 15 selected for validity and reliability testing] ......... 212
Table 33 NESSTAC – protocol no.11 [155][from 15 selected for validity and reliability study]. 213
Figure 16 Wheel for PRECIS-2 following user testing and used in Pilot testing for Validity and
reliability testing ................................................................................................................ 215
Table 34 The nine PRECIS-2 domains descriptions, tested in validity and reliability testing..... 218
Figure 17 Stages of development of PRECIS-2 ................................................................. 219
Table 35 Demographics of PRECIS participants invited to participate in validity and reliability
testing .................................................................................................................................... 229
Table 36 Different groups of 35 participants invited to participate in Validity and Reliability
testing..................................................................................................................................... 232
Figure 18 Flow diagram of participants in validity and reliability testing .......................... 234
Figure 19 Recruitment rate for PRECIS-2 validation ....................................................... 236
Table 37 Example of scoring by one of the raters for the ASHA protocol using the Table (FA) . 238
Figure 20 Example of scoring by one of the raters for the IST-3 protocol [157] using the PRECIS-2
wheel (JS) .............................................................................................................................. 239
Figure 21 Agreement benchmarks for the evaluation of observed k values ....................... 240
Table 38 Overall results for Inter-rater reliability for 9 PRECIS-2 domains including sensitivity
analysis .................................................................................................................................... 242
Table 39 Rater scoring using randomly generated values (1 to 5) of PRECIS-2 domains using
Intraclass coefficient (ICC) for 15 trials, 19 raters .............................................................. 243
Agreement for Pragmatism scores for ST and KL was 80% (12/15) (................................. 243
Table 40 Pragmatism scoring results for KL and ST of 15 trials....................................... 244
Table 41 Discriminant validity measured using Area Under the ROC curves (AUROC) ....... 244
Table 53 Free text feedback for COPERS – APT study pilot .............................................................. 326
Figure 47 PRECIS-2 wheel created by website tool showing ranges and maximum and minimum scores, on COPERS, post APT pilot ........................................................................................................... 330
Figure 48 PRECIS-2 wheel created by website tool showing range of scores (5 raters) for HepFree .................................................................................................................................................. 335
Table 54 Scores for all 5 raters including median scores with maximum and minimum ........ 335
Figure 49 Scores for all 5 raters including median scores with maximum and minimum .......... 335
Table 55 Completed Questionnaire for HepFree – APT study pilot ........................................ 336
Figure 50 Pharmacy/Pharmacist Level PRECIS-2 Wheel for STOP trial ................................. 348
Table 56 Pharmacy/Pharmacist Level PRECIS-2 Scores for STOP trial ................................. 349
Figure 9.8 Patient Level PRECIS-2 Wheel for STOP trial ....................................................... 349
Figure 51 Patient Level PRECIS-2 Wheel for STOP trial ....................................................... 349
Table 57 Patient Level Scores for PRECIS-2 STOP trial ....................................................... 349
Table 58 Completed Questionnaire for STOP – APT study pilot .......................................... 351
Table 59 RE- AIM framework [66] .............................................................................................. 367
Figure 52 Framework for enhancing the value of research for D&I [215] ................................. 369
<table>
<thead>
<tr>
<th>Term</th>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>APT</td>
<td>Applying PRECIS-2</td>
<td>to Primary Care Trials</td>
</tr>
<tr>
<td>BP</td>
<td>Blood Pressure</td>
<td></td>
</tr>
<tr>
<td>CAD</td>
<td>Coronary Artery</td>
<td>Disease</td>
</tr>
<tr>
<td>CHD</td>
<td>Coronary Heart</td>
<td>Disease</td>
</tr>
<tr>
<td>CTMP</td>
<td>Center for Medical</td>
<td>Technology</td>
</tr>
<tr>
<td>CVD</td>
<td>Cardiovascular</td>
<td>Disease</td>
</tr>
<tr>
<td>DPB</td>
<td>Diastolic Blood</td>
<td>Pressure</td>
</tr>
<tr>
<td>DOT</td>
<td>Directly Observed</td>
<td>Treatment</td>
</tr>
<tr>
<td>Dec</td>
<td>December</td>
<td></td>
</tr>
<tr>
<td>D&amp;I</td>
<td>Dissemination and</td>
<td>Implementation</td>
</tr>
<tr>
<td>EES</td>
<td>Everolimus Eluting</td>
<td>Stent</td>
</tr>
<tr>
<td>Explanatory</td>
<td>Efficacy, efficiency – ideal highly controlled</td>
<td></td>
</tr>
<tr>
<td>EVOO</td>
<td>Extra Virgin Olive Oil</td>
<td></td>
</tr>
<tr>
<td>HCTZ</td>
<td>Hydrochlorothiazide</td>
<td></td>
</tr>
<tr>
<td>HCSC</td>
<td>Health Care System</td>
<td>Collaboratory</td>
</tr>
<tr>
<td>HD</td>
<td>High Density Lipoprotein</td>
<td></td>
</tr>
<tr>
<td>HF</td>
<td>Heart Failure</td>
<td></td>
</tr>
<tr>
<td>HSRU</td>
<td>Health Services</td>
<td>Research Unit</td>
</tr>
<tr>
<td>LDL</td>
<td>Low Density lipoprotein</td>
<td></td>
</tr>
<tr>
<td>LDLC</td>
<td>Low Density Lipoprotein Cholesterol</td>
<td></td>
</tr>
<tr>
<td>MeSH</td>
<td>Medical Subject</td>
<td>Heading</td>
</tr>
<tr>
<td>MI</td>
<td>Myocardial Infarction</td>
<td></td>
</tr>
<tr>
<td>MOOC</td>
<td>Massive Open Online Course</td>
<td></td>
</tr>
<tr>
<td>NLM</td>
<td>National Library</td>
<td>of Medicine</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Definition</td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>------------</td>
<td></td>
</tr>
<tr>
<td>NIC:</td>
<td>Nicardipine</td>
<td></td>
</tr>
<tr>
<td>NIH:</td>
<td>National Institute of Health</td>
<td></td>
</tr>
<tr>
<td>NIHR:</td>
<td>National Institute of Health Research</td>
<td></td>
</tr>
<tr>
<td>NS:</td>
<td>Not Statistically Significant</td>
<td></td>
</tr>
<tr>
<td>Participant</td>
<td>An individual who takes part in a clinical trial</td>
<td></td>
</tr>
<tr>
<td>PCI:</td>
<td>percutaneous coronary intervention</td>
<td></td>
</tr>
<tr>
<td>PCORI:</td>
<td>Patient Centered Outcomes Research Institute</td>
<td></td>
</tr>
<tr>
<td>PCTA:</td>
<td>percutaneous transluminal coronary angioplasty</td>
<td></td>
</tr>
<tr>
<td>PCTU:</td>
<td>Pragmatic Clinical Trials Unit</td>
<td></td>
</tr>
<tr>
<td>PES:</td>
<td>Paclitaxel Eluting Stent</td>
<td></td>
</tr>
<tr>
<td>PI:</td>
<td>Principal Investigator</td>
<td></td>
</tr>
<tr>
<td>Pragmatic:</td>
<td>Practical, effectiveness, management, real world</td>
<td></td>
</tr>
<tr>
<td>RCT:</td>
<td>Randomised Controlled Trial</td>
<td></td>
</tr>
<tr>
<td>RD:</td>
<td>Risk Difference</td>
<td></td>
</tr>
<tr>
<td>SNP:</td>
<td>Nitroprusside</td>
<td></td>
</tr>
<tr>
<td>TAXUS:</td>
<td>Paclitaxel releasing stent</td>
<td></td>
</tr>
<tr>
<td>TVR:</td>
<td>Target Vessel Revascularisation</td>
<td></td>
</tr>
</tbody>
</table>
I would like to thank the following individuals and groups who have helped me complete this thesis:

- A special thank you to Professor Shaun Treweek for guidance and sharing my enthusiasm as PRECIS-2 took shape. Professors Merrick Zwarenstein, Frank Sullivan, and Peter Donnan for their wisdom and support.

- Delphi participants, user testers and validity and reliability testers of PRECIS-2 including the pilot testers and Dundee Brainstorming team: Professor Guthrie, Thomas Lamont, Heather Cassie, Fiona Hogarth, Dr Ildiko Gagyor and Alison McDonald for their helpful input.

- Everyone who participated in the Toronto THETA meeting 11th March 2013 for their palpable enthusiasm and good humour for my PhD and remaining glued to their seats for a face to face full day meeting, without me needing to lock the door.

- Leandro Tramma, Health Informatics Centre (HIC) for great collaboration in creating the PRECIS-2 website.

- Gordon Forbes for discussions on PRECIS-2, testing PRECIS-2 website and as PI on APT.

- Dr Sally Hopewell, my friend and mentor who offered helpful and constructive advice on my thesis pre-submission.

- QSI and all my friends for all their love and support and providing much needed distraction over the last few years.

- My wonderful parents and sons, Angus and Halvar for their forbearance - my apologies for putting my sons through their mother’s PhD during their teenage years - I think we have all matured during this period.

- The Medical Research Council and the University of Dundee for their stipend and the Chief Scientist Office for their grant to enable me to carry out this work.
Declaration

I declare that I am the author of this thesis and that all of the work it describes has been carried out by me. Professor Shaun Treweek undertook duplicate work to check data extraction and scoring using the PRECIS-2 tool and the Cochrane Risk of Bias tool. In addition, Professors Shaun Treweek, Merrick Zwarenstein, Frank Sullivan and Peter Donnan participated in Brainstorming meetings in Toronto and Dundee. Peter Donnan also checked statistical calculations for the Validity and Reliability work for the PRCIS-2 tool. I declare that I have consulted all the references cited in this thesis and that this work has not previously been accepted for a higher degree.

Kirsty Loudon

Date 22/5/2015
Summary of Contents

**Aim**
PRECIS (PRagmatic Explanatory Continuum Indicator Summaries 2009) is a tool with a simple wheel format that trialists can use when designing their trials to improve the applicability of results but users highlighted problems. The aim of the study was to produce an improved and validated version of PRECIS, called PRECIS-2 and test this tool out with trial teams designing primary care trials.

**Methods**
Brainstorming and a 2-round Delphi survey of authors who cited PRECIS plus user-testing of candidate PRECIS-2 models was followed by validity and reliability testing of the most promising PRECIS-2 candidate using a sample of 15 trials rated by 19 different trialists. The validated PRECIS-2 tool was then used to consider the risk of bias (internal validity) and estimates of treatment effect of a matched set of explanatory (ideal conditions) and pragmatic (real world) trials. The PRECIS-2 website was also created with a database of pragmatic trials and a toolkit for trial groups. This was tested out at the Pragmatic Clinical Trials Unit (PCTU) in London with trial teams designing primary care trials.

**Results**
Forty-two people responded to the Delphi and highlighted scoring, domain choice, and tool format as issues. An expert panel of 14 in Toronto provided the basis for a PRECIS-2 model that was then user tested by 19 other methodologists and trialists. After 13 iterations, a PRECIS-2 model with 9 domains (i.e. Eligibility, Recruitment, Setting, Organisation, Flexibility Delivery, Flexibility Adherence, Follow up, Primary Outcome, Primary Analysis) was tested for validity and reliability. Inter-rater reliability was generally good, with eight of nine domains having an ICC over 0.65. Discriminant validity was reasonable for all domains, though with wide confidence intervals. Matching trials taking pragmatic (‘real world’) and explanatory (‘ideal world’) approaches was challenging but we found no indication that a pragmatic approach compromises internal validity. We were unable to extract sufficient
information for a planned analysis of estimates of treatment effect. At the PCTU, the tool highlighted differences in opinion with trial team members and demonstrated convergence of opinion following discussion. There was acknowledgment that scoring of PRECIS-2 domains assisted trials teams in considering the intended audience and creation of trials relevant to practice. Useful feedback was obtained to improve the PRECIS-2 tool software for users.

Conclusions
PRECIS was improved by the addition of scoring and additional domains after consultation with over 80 international trialists. We have a validated PRECIS-2, in the visually appealing wheel format with 9 spokes, which is being made available through an increasingly accessed website. Work at the PCTU improved the usability of the PRECIS-2 website and demonstrated that the tool increases transparency in trial design and assists trialists in considering applicability of trial results. More matching work on the impact of design approaches on effect size is needed, and further data to support the risk of bias results would be valuable.
Chapter 1: Introduction

History of clinical trials

Trials have been around a long time. From the first recorded trial in the Old Testament (Book of Daniel 1:3-16 New English Bible) when Daniel and three others fared much better on “only vegetables to eat and water to drink” in comparison to the other healthy Israelite exiles who were given a daily allowance of fine food and wine from the Royal Table of King Nebuchadnezzar 11, to Ambroise Pare’s gunshot wound trial in 1575 [2], to James Lind’s scurvy trial on sailors on HMS Salisbury in 1747 [3, 4] culminating in the first randomised controlled trial of the Medical Research Council (MRC) in 1948 to test streptomycin for pulmonary tuberculosis [5], a clinical trial is simply a way of testing whether or not a drug, device or technique works. Sixty seven years ago, the MRC trial became an important milestone in controlling selection bias in clinical trials, using allocation concealment for randomisation of patients to create unbiased comparison groups [6]. Bradford Hill has been credited with this very important methodological contribution to the development of therapeutic trials [6, 7].

Clinical trials development

In the UK it is estimated that nearly £5 billion is spent on research and development each year but to develop a new drug it currently takes on average more than ten years and £1 billion (www.gov.uk/government/news/major-investment-in-life-sciences). With new technology it is hoped this development time will be dramatically reduced, all good for patients and clinicians and with the global market for clinical trials expected to be worth £29 billion in 2015 this is also very good for the UK economy with investment in this area.

The Cooksey report suggested there were two gaps in effective knowledge translation, the production and the uptake of evidence [8]. Barriers to translation and uptake of evidence include the creation of relevant research and adequate consideration of the intervention characteristic which will be the focus
in this thesis but also include the knowledge, attitudes and beliefs of the stakeholders, their readiness for change and engagement in adapting the intervention to make it “fit” local conditions and needs [9-11]. Other issues include financial/organisational instability, limited time and resources, challenges implementing interventions with quality which may be separate or include the challenge of ensuring the fidelity of the intervention and adapting it [9, 12]. However, with problems of “adoption and diffusion” being addressed in the NHS one could argue that the first gap, the design of trials is the basis of translating evidence into practice [13, 14], if the right questions are asked. As evidence based medicine is critical to improving patient care thus trialists need to be more aware of the end user and design more relevant trials that are robust using transparent methodology. Thus, trialists need to make it easier for trial results to be used at an earlier stage, considering cost implications and speed of replication in clinical settings.

It is widely accepted that randomised controlled trials are the best design to evaluate clinical and organisational interventions [15-17]. But words that still hold true nearly fifty years later by Schwartz and Lellouch are not always heeded or fully understood “...most therapeutic trials are inadequately formulated, and this from the earliest stages of their conception. Their inadequacy is basic, in that the trials may be aimed at the solution of one or other of two radically different kinds of problem; the resulting ambiguity affects the definition of the treatments, the assessment of the results, the choice of subjects and the way in which the treatments are compared.” [18]

**Applicability**

Many authors have highlighted the need for trials with greater applicability [19-21]; in other words, trials that pay attention to external validity as well as internal validity. Lack of consideration of external validity is the most frequent criticism by clinicians of RCTs, systematic reviews and guidelines [20].
Several studies have emphasised that healthcare interventions are seldom given under circumstances similar to those used in such trials [22-26] from exclusion of the elderly [24] to people with renal disease routinely being excluded from cardiovascular trials [23]. Travers [26] conducted a postal survey of 3500 randomly selected individuals in New Zealand. He found that a median of 4% of participants with current asthma met the eligibility criteria for 17 included RCTs for consensus guidelines and a median of 6% of participants with current asthma on treatment met the eligibility criteria. In other words 94% of people with asthma taking medication would not have been eligible to be included in an RCT. In this particular case, we can see that the results of asthma trials may not be applicable to the majority of asthma patients and their practitioners trying to determine the best care, yet these trials are the basis for recommendations in clinical guidelines on asthma. A similar study also by Travers [22] on chronic obstructive pulmonary disease (COPD) found that a median of 5% of participants fulfilled the eligibility criteria for 18 RCTs on COPD, again indicating that the trial results had limited external validity to the majority of patients taking these medication. The results of these trials, however, were the basis for the Global Initiative for Chronic Obstructive Lung Disease (GOLD) guidelines. Another study compared the eligibility of Scottish patients with Type 2 Diabetes to the participants included in seven RCTs on glycaemic control; the trial participants were not representative of the population receiving treatment with 3.5 to 50.7% of people with Type 2 Diabetes Mellitus (T2DM) fulfilling eligibility criteria. These trials were designed with limited applicability and make it hard for guidelines developers informing best practice for physicians treating patients with T2DM in Scotland.

Some trials, however, do consider the applicability of trial results and the end user. For example, Anderson [27] in a feasibility trial for a lifestyle intervention programme (ACTWELL) targeted post-menopausal women being screened for breast cancer. This group of women have an increased risk of breast cancer, particularly if overweight, with greatest risk amongst women from the most deprived areas. This trial had few exclusion criteria (BMI < 20 kg/m², contraindications for physical activity and contraindications for weight loss programme) and tested out an intervention that could be slotted into
the existing framework for delivering healthcare to the population being targeted. This pilot study succeeded in including participants from areas of high social deprivation, who are usually hard to reach, in opportunistic recruitment with minimal contact from a personal trainer giving promising results. Trials like this, which seek to produce relevant results and take into consideration external validity at the design stage, should ensure greater applicability of trial results.

Explanatory and pragmatic attitudes

Schwartz and Lellouch in 1967 proposed there were two attitudes to randomised controlled trials: pragmatic and explanatory [18, 28]. Pragmatic trials are performed under normal conditions with the intention of providing results that are more applicable to clinical practice and decision making. The alternative, taking a more explanatory approach, leads to tightly controlled trials under ideal conditions that aim to provide understanding of how treatments work [28]. Explanatory trials have an important role but healthcare interventions are seldom given under circumstances similar to those used in such trials [23, 24]. There is seldom a purely explanatory or pragmatic trial; it is not a dichotomy but a continuum. The problem, as Schwartz and Lellouch noted, is that trialists often design trials that are more explanatory than they should be - given their purpose. These trials may be free of bias but are of no or limited relevance to the general population of people with the targeted condition. Applicability is vital to those who have to make healthcare decisions including patients, clinicians and policymakers. Many authors have highlighted the need for trials with greater applicability [19, 20, 29] - in other words, trials that pay attention to external validity as well as internal validity. Lack of consideration of external validity is the most frequent criticism by clinicians of RCTs, systematic reviews, and guidelines [20]. This leads to research waste, as described in the Lancet series of articles on increasing value, reducing waste [30].

Pragmatic trial design decisions, however, are not without controversy. In 2011, pragmatic trial design hit the research headlines over the discussion of the Advisory Committee to the US Food and Drug
Administration (FDA) to approve rivaroxaban to minimise stroke risk for patients with atrial fibrillation [31]. This decision to recommend rivaroxaban, over the more well-known alternative anticoagulant treatment warfarin, was not unanimous - debate arose over the pragmatic study design used in the trial and the applicability of the trial results. The majority (9 vs. 2, one abstention) thought the pragmatic design of the ROCKET AF trial was good but some questioned the rigour of the design and the compliance rates seen in the trial. The debate highlighted the impact design decisions have on clinicians’ and others’ confidence in trial results and, in particular, the need to select patients who are “truly reflective of the type of patients physicians would see in everyday practice” [31].

But how does a trialist know that his or her trial has a more pragmatic or explanatory design? Campbell published a table (Table 1) to explain the differences between pragmatic and explanatory trials as part of the CONSORT statement to improve reporting of Pragmatic trials [32], others have done the same with slightly different definitions [33]. The common factor is that pragmatic trials (effectiveness) are real world, usual care for typical patients in typical settings and explanatory trials (efficacy) are specialised care for specific patients in ideal (very specific) settings. Pragmatic trials have been called Management trials or Practical trials but there are differences in the definitions for these trials and Explanatory trials are also called Efficiency or mechanistic trials [34, 35].

Table 1 Improving the reporting of Pragmatic trials: an extension of the CONSORT statement [32]

<table>
<thead>
<tr>
<th>Question</th>
<th>Efficacy—can the intervention work?</th>
<th>Effectiveness—does the intervention work when used in normal practice?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Setting</td>
<td>Well resourced, “ideal” setting</td>
<td>Normal practice</td>
</tr>
<tr>
<td>Participants</td>
<td>Highly selected. Poorly adherent participants and those with conditions which might dilute the effect are often excluded</td>
<td>Little or no selection beyond the clinical indication of interest</td>
</tr>
<tr>
<td>Intervention</td>
<td>Strictly enforced and adherence is monitored closely</td>
<td>Applied flexibly as it would be in normal practice</td>
</tr>
<tr>
<td>Outcomes</td>
<td>Often short term surrogates or process measures</td>
<td>Directly relevant to participants, funders, communities, and healthcare practitioners</td>
</tr>
<tr>
<td>Relevance to practice</td>
<td>Indirect—little effort made to match design of trial to decision making needs of those in usual setting in which intervention will be implemented</td>
<td>Direct—trial is designed to meet needs of those making decisions about treatment options in setting in which intervention will be implemented</td>
</tr>
</tbody>
</table>
PRagmatic Explanatory Continuum Indicator Summary - PRECIS

In 2009 a tool called the Pragmatic-Explanatory Continuum Indicator Summary – PRECIS - was published to help trialists to think more carefully about the impact their design decisions would have on applicability and external validity [36]. PRECIS was developed by more than 25 international trialists and methodologists, including ST and MZ (PhD supervisors), working between 2005 and 2008 [36], linked to simultaneous work on the CONSORT extension for pragmatic trials [32]. The aim was to produce a tool that helped trialists match their design decisions to the purpose of the trial, be that informing a clinical decision or increasing knowledge of how an intervention works. PRECIS provides a simple graphical summary - a 10-spoked ‘wheel’ with key design decisions (see Figure 1). These ten domains help trialists ensure their design decisions are consistent with the purpose of the trial. Trials that take an explanatory approach produce wheels nearer the hub; those with a pragmatic approach are closer to the rim.

There was, and is, demand for this sort of tool (PRECIS had been cited over 300 times by the end of 2014) and has been selected as one of eight *Useful papers* for trialists on the UK National Institute for Health Research Randomised Trials Methods website ([http://www.nets.nihr.ac.uk/resources/trials-coordination](http://www.nets.nihr.ac.uk/resources/trials-coordination)).
**Alternatives to PRECIS**

There was another tool designed by Gartlehner [37], published in 2006, three years prior to the PRECIS publications in 2009, that also considered the explanatory and pragmatic nature of a trial but it was created for assessing trials in systematic reviews. This tool has not received so many citations (75 times by the end of 2014) and it aims to distinguish effectiveness (pragmatic) trials from efficacy (explanatory) trials. It has seven items: populations in primary care; less stringent eligibility criteria; health outcomes; long study duration with clinically relevant treatment modalities, assessment of adverse events, adequate sample size to assess a minimally important difference from a patient perspective, Intention To Treat analysis. Although it has been cited, few people have used the tool as
it was intended: to assess the applicability of results from randomised controlled trials. Instead it has been criticised for not being simple enough to use and being confusing for users.

There have also been concerns concerning face validity as it is unclear how the seven items were selected as being important for determining pragmatism [38]. Finally, there are also problems with composite scoring if the seven different items in the Gartlehner tool are added up as similar scores will have different compositions [38]. This was also a problem raised by groups that used PRECIS and composite scoring of the domains [39, 40] One group, however, that has undertaken methodological work led by Zettler [41] analysed 151 RCTs using the Gartlehner tool. Using three raters for N (non-pharmacological) and P (pharmacological) trials, they determined that the tool needs work, and that the items need to be improved to improve rater reliability but the tool highlighted the “importance of judging the aim, relevance, and external validity of a given study within the decision-making process.” [41]. As the Gartlehner tool focus is on systematic reviews to consider issues of pragmatic trial design and applicability, and it emphasises trials are either explanatory or pragmatic and there is not a continuum from explanatory to pragmatic, we did not consider improving this tool.

**Pragmatic trials and internal validity**

Pragmatic trials have been proposed as a way of achieving wide applicability and providing research answers to routine care problems [19-21]. Designing pragmatic trials, however, is not without controversy; compromised internal validity is often put forward as an argument against pragmatic designs [42-44]. There does not, however, appear to be any empirical evidence supporting (or refuting) this belief that trials that take a more pragmatic approach sacrifice internal validity. Indeed, the CONSORT Statement [32] has been extended to pragmatic trials [32] and makes the same demands regarding reporting of information relevant for assessing internal validity as the full CONSORT Statement [45]; it does not suggest that internal validity can be compromised in a trial taking a pragmatic approach. Internal validity can be assessed using the Cochrane Risk of Bias tool and it comprises of: Sequence generation for randomisation, allocation concealment, blinding, selective
outcome reporting, incomplete outcome data and other sources of bias [46]. The only aspect that could be different in pragmatic and explanatory trials is blinding - “if blinding was not done, or was not possible, explain why” so even here the expectation is that to ensure internal validity, blinding should be carried out wherever possible. However, in routine care blinding is rarely undertaken.

Rothwell states that internal validity is independent of “all aspects of the design and performance that impact on the external usefulness of the result of a trial”, which he calls the external validity [45]. He also suggests internal validity underpins the rules for randomised trial design and is thus a given whether pragmatic or explanatory design, whereas external validity and applicability requires clinical judgements from medical practitioners who have clinical expertise [45]. Similarly explanatory trials almost by definition do not explicitly consider external validity, which would be reasonable if we were certain that applying the results to patients in the real world would not cause harm. This however is not always the case e.g. Merck Vioxx testing for the treatment of rheumatoid arthritis pain where everyone was excluded with heart disease http://kenan.ethics.duke.edu/wp-content/uploads/2012/08/Vioxx_Case2015.pdf. Trialists need to ensure internal validity in all trials, “real world setting” or “ideal setting”. It is important that we do not confuse confounding with the reality of practice.

Our hypothesis is that pragmatic trials do not compromise internal validity (we should not mistake realism for bias) and that explanatory trials overestimate effect size compared to more pragmatic designs and that recommendations regarding trial design, especially for regulatory trials, need to be modified. This thesis intends to explore this hypothesis and address this methodological gap by providing empirical evidence to support or refute pragmatic design choices in trials. As the PRECIS tool was developed to help trialists make design choices that support applicability, the PRECIS was the starting point to answer this question. First, a systematic review was undertaken to gather all available evidence and synthesise themes to indicate changes that would improve the tool and develop PRECIS-2. Then the new, modified PRECIS-2 tool was used to compare the internal validity of, and effect
estimates from, a set of matched (by intervention) trials, some taking an explanatory approach, others a more pragmatic approach. It was hoped that this work would assist further understanding of the implications of trial design decisions and, therefore, the relevance and impact of clinical trials. The intention was to answer one way (or the other) whether or not pragmatic trials sacrifice internal validity for external validity (applicability).
Chapter 2: Systematic review of methodological work on PRECIS

Background

Many guideline producers need trials to provide data that are relevant to the clinical decisions made by patients, health professionals and policymakers. Many trials, however, are less applicable than they could be due to the actual participants selected, setting different to usual care, outcomes that are not so relevant for practice, only including expert clinicians, or because of some other departure from usual practice conditions. These result in a great deal of resources being put into trials that are not as useful as they should be.

The aim of the original PRECIS tool was to produce a tool that helped trialists match their design decisions to the purpose of the trial, be that informing a clinical decision or increasing knowledge of how an intervention works. Resources are scarce and improving and validating PRECIS provides a tool for focussing the trial team’s discussion on their research question and the trial’s purpose with the worthy outcome of reducing research waste [30]. The original PRECIS tool has been well cited but there are indications that although useful, modifications are necessary to improve the use of the tool to design clinical trials. A systematic review of the literature was therefore undertaken to determine the extent of the methodological work that had been undertaken in order to inform development of a new improved version of the tool - PRECIS-2 that could be validated.

Aim

To undertake a comprehensive assessment of the methodological research on the PRECIS tool.
Objectives
The aim of this systematic review was achieved through assessing who had undertaken the research; what type of research they had used the PRECIS tool for and why; how the tool had been modified and finally the methodology for using the PRECIS tool.

Methods
The PRISMA guidance was used to structure this systematic review [47].

Eligibility Criteria
Studies that were included in the systematic review of the literature for assessment were all citations of the two jointly published articles on the original PRECIS tool in the Journal of Clinical Epidemiology and the Canadian Medical Association Journal. Included studies were published methodological articles using the original PRECIS tool.

Information Sources
The included studies were found from citations of the original PRECIS article in joint publications of the Canadian Medical Association Journal [48] and the Journal of Clinical Epidemiology [49] (2009) in any language. Colleagues were also a source of information for published and grey literature.

Search
The knowledge database Web of Science www.webofscience.com was used to keep track of citations on a regular basis from October 2011. The first formal search which was the basis for the published study protocol was undertaken 1st September 2012 and checked again (no new methodological citations were found) prior to submitting the protocol manuscript 1st January 2013. (This described the work of developing PRECIS-2 and testing the internal validity of pragmatic trials.) The formal date for the literature search update for the thesis was December 2014.

Study Selection
All articles that had cited the original tool were screened through reading the full article to detect methodological articles as this was not always possible through scanning titles and abstracts. Initially they were categorised into article type: discussion or opinion piece, trial protocol, trial results or
methodological piece. As the number of citations increased articles were simply screened for methodological work (Figure 1 PRECIS – Pragmatic-Explanatory Continuum Indicator Summary)

**Data collection process**
Electronic or hard copies of all the citations were screened. If there was uncertainty in any aspect of the use of the original PRECIS tool e.g. number of raters or if consensus scoring had been employed then the authors were contacted for further information.

**Data items**
The following data was extracted by KL from the methodological publications on PRECIS:

- who had undertaken the methodological work (disciplines);
- rationale for the work;
- what type of study PRECIS being used in (i.e. designing a randomised controlled trial (RCT); retrospective assessment of RCT design; designing a non-randomised study; or a systematic review);
- number of raters involved;
- scoring system – the scale;
- if consensus scoring had followed independent scoring;
- domains added or removed;
- additional information that would be helpful in modifying the original PRECIS.

**Synthesis of results**
A narrative synthesis was used to combine the methodological studies using the original PRECIS tool. Similarities in data and differences were highlighted with the aim of utilising the information to guide the development of a new improved version of the tool PRECIS-2. The systematic review was the initial stage in this process.
Figure 2 QUORUM flow diagram for methodological articles on PRECIS
Consolidation of the results
Most of the references referred to the PRECIS publication to indicate the importance of applicability and in their discussion describing why they considered their trial to be pragmatic: “the trial was designed according to the PRECIS tool in order to simulate “usual-care” conditions” [50]. Other studies were opinion pieces citing PRECIS to encourage trialists to consider using this new tool for creating more relevant trials: “Place for studying interventions already shown to be explanatory in a pragmatic way” [51]. Some studies, however, had used PRECIS in methodological work: to design trials, retrospective use of PRECIS on trial protocols, to assess trials in systematic reviews and finally to assess a non-randomised study (Table 2).

Table 2 Study design of methodological studies on PRECIS

<table>
<thead>
<tr>
<th>Methodological study design using PRECIS</th>
<th>Study authors (year of publication)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trial design</td>
<td>Riddle (2010), Witt (2013)</td>
</tr>
<tr>
<td>Non-randomised study</td>
<td>Elder (2014)</td>
</tr>
</tbody>
</table>

Following the publication of the protocol for the work described in this thesis (April 2013) [52] which included the previous review of work by eight groups [52], seven additional research groups have also published methodological work on PRECIS. Each of the methodological studies will now be considered in turn in the different uses of: trial design, retrospective use to demonstrate trial decisions, systematic reviews and in a non-randomised study.
Using PRECIS at trial design stage

The first study to publish using PRECIS in trial design was by Riddle and colleagues ([53] planning total knee surgery. The trial investigators came from a wide range of specialities (physical therapy, biostatistics, rehabilitation medicine, psychiatric and behavioural sciences, medicine, rheumatology). Riddle [53] found the PRECIS tool useful for discussing trial design and achieving consensus during a one day face to face meeting. To enable constructive discussion Riddle [53] used a PRECIS wheel with spokes that were 0-4cm in length with the zero point at the hub (explanatory) and 4cm at the rim of the wheel (pragmatic). They used the tool as it was intended - to design the protocol for a clinical trial; pre-meeting, ideal score and post-consensus meetings when all the raters were present but blinded to each other’s score. Raters marked a point on the wheel which was then measured. This must have been a little time consuming but allowed the participants to quantify disagreement between raters which could then be discussed. Variation in scoring for the 10 PRECIS categories, as reflected in average standard deviation moved from initial assessment of 0.83, to 1.16 for ideal trial for individual raters, to 0.61 for the final assessment.

In 2013 Witt published a protocol for an RCT [54] to compare traditional Ayurvedic medicine against conventional treatment for osteoarthritis of the knee. Seven researchers, including physicians and statisticians independently used PRECIS with a 1-5 scale to “guide the design of the trial” and assess where the trial was on the maximum efficacy “1” to maximum effectiveness “5” continuum. The raters then discussed ratings to reach consensus. The PRECIS tool, however, was adapted by removing primary analysis and including patients’ compliance and practitioners’ adherence domains for both the Ayurvedic intervention and conventional care control to create a wheel with eleven domains. The trial team created a trial that was generally midway between explanatory and pragmatic but more pragmatic on the Flexibility of delivery of the Ayurveda and more explanatory on the Follow up intensity, Practitioner adherence and Patient compliance domains. Witt believes the PRECIS wheel indicates the trial design choices to “exclude bias” and produce a personalised treatment for patients.
Using PRECIS retrospectively

In the study by Tosh et al [55], three doctors who had research backgrounds and specialised in psychiatry, used PRECIS to evaluate mental health protocols. Tosh’s [55] group decided they wanted to quantify how pragmatic 10 protocols were, and decided to score PRECIS using 1 (most explanatory) to 5 (most pragmatic) with zero for domains that did not contain any information. They used 0-30 for an explanatory trial, 31-39 for a trial balanced between pragmatic and explanatory, and 40-50 for a pragmatic trial. Calling their modified PRECIS, the Pragmascope tool, reliability was reasonable for the raters (weighted Kappa = 0.72).

Glasgow et al [56] invited nine raters who were medical doctors or had doctorates and were experienced researchers to use PRECIS; six were involved in the trials they were rating, Practice-based Opportunities for Weight Reduction (POWER) and three independent raters who had nothing to do with the trials being scored. Glasgow et al [56] used a score from 0 (most pragmatic) to 4 (most explanatory) but as Selby [40] noted “several telephone calls were required to develop consensus on the meaning of each score” and he hypothesised that “perhaps the scale was not sensitive enough to detect a difference”. There were problems with inter-rater reliability and raters determining that their own trial was more pragmatic than independent raters – Glasgow questioned if this was due to bias or more trial knowledge.

In his first paper using PRECIS, Bratton [57] scored three trials retrospectively on tuberculosis (TB) treatment with another biostatistician who had a particular interest in TB. They combined the spokes on expertise and inserted a domain for “blinding” and found the tool helpful to pinpoint weaknesses in the trial design as well as assisting trialists to consider the applicability of their trials. Bratton [58] then used PRECIS again in an ongoing trial of an autoimmune disease non-inferiority trial, the degree of pragmatism was considered by a trial manager, two clinicians with expertise in the trial area of pemphigoid and two statisticians. Bratton [58] advocated the procedure described by Riddle [53] and
so his group also used a blank PRECIS wheel approximately 15cm in diameter. While they recommended using PRECIS to achieve consensus they had a few concerns that there was a great deal of variability scoring domains depending on expertise of raters and interpretation of guidance from the original Thorpe paper [48, 59].

Selby et al [40] for a trial protocol on smoking cessation had a team of six raters with a range of expertise (one academic family physician with an interest in smoking cessation; one cardiac rehabilitation physician with expertise in pharmacoeconomics; one addiction medicine physician and clinical scientist with a focus on tobacco dependence; one pharmacist with expertise in pharmacoeconomics; one pharmacologist with clinical research and medical affairs experience in the pharmaceutical industry; and one consultant physician with pharmacoeconomics and policy advice experience in Quebec). Selby et al [60] highlighted how difficult it was to use a Visual Analogue Scale for raters working online using PRECIS to rate the trial protocol of a smoking cessation trial (which had started). As the printers they were using were all different, so too hard to use a VAS for comparison purposes, they chose to use a 20 point scale, where 1 represented ‘entirely explanatory’ and 20 represented ‘entirely pragmatic’ making it easier to score using e-mail. They scored the protocol twice using a modified Delphi process. Selby advocates that using a system of scoring PRECIS anonymously allows different views to be expressed and assists in consensus decision making from multi-disciplinary raters. This ensures the best design to answer the study question. Rater scores varied less after the second round than the first indicating convergence in opinions following discussion.

As highlighted by Selby [40] composite PRECIS scores are not helpful in assessing how pragmatic or explanatory a trial is, as very dissimilar trials can have the same score which is only revealed by closer examination of the PRECIS domains and the different aspects of the trial design. This was backed up by Glasgow [56], he found that using the PRECIS tool, trials with similar overall scores could have completely different scores for different criteria, for example the eligibility criterion.
Krist and two colleagues [61] used PRECIS to demonstrate how pragmatic their protocol was for the My Own Health Record (MOHR) cluster randomised trial at the end of the design stage, instead of at the beginning of the process, as suggested by Thorpe et al. Three raters used a 0-4 scale to agree on a score to indicate trial design mostly very pragmatic, there was no independent scoring. Krist and colleagues scored their trial with the maximum score of pragmatism - “4” for Intervention flexibility and Control flexibility and Practitioner expertise and “3” for the other seven domains. They concluded using PRECIS that this trial was more pragmatic than published trials which also addressed behavioural changes [56]. The raters included two medical doctors involved in health research and a social anthropologist/social worker.

Damschroder, a health researcher, and colleagues used PRECIS at the end of the trial design period immediately prior to protocol publication (as mentioned previously not as recommended by Thorpe and colleagues [62]. One rater (LD – first author) scored the trial using a scale of “1” restrictive to “5” non-restrictive with a three armed intervention for weight loss to treat obesity prior to discussion with eight colleagues. The authors noted that the ratings were subjective and based on published guidance. The authors demonstrated that the trial was very pragmatic with seven domains as highly pragmatic (score ≥ 4).

**Using PRECIS with systematic reviews**

Koppenaal et al [39] adapted PRECIS - PRECIS-Review tool (PR-tool) - to rate trials in a systematic review to determine which intervention would improve lifestyle most in general practice. As the original PRECIS tool produces a wheel like figure for each trial it is hard to assess how applicable to routine care a systematic review is. So, Koppenaal et al [39] proposed scoring the domains and each trial using a scoring system of 1-5 as well as 0-100%; they found that a scale of 1-5 was less subjective compared with a 0-10 scale. By scoring all the trials in the systematic review they were able to determine which trial was most pragmatic and overall how pragmatic and thus how applicable the review was. They were also able to determine how heterogeneous the individual trials were. They did have concerns
though about arbitrary scoring and that more pragmatic scoring occurred when there was inadequate information. Koppenaal et al [39] proposed that two researchers should score RCTs to improve validity. His team also highlighted problems with context that the PR-tool did not pick up and that weighting of domains may be dependent on the situation. Overall he concluded that the PR-tool – PRECIS with a scoring system - was very helpful to users of RCTs and systematic reviews.

Witt et al [63] used PRECIS to evaluate acupuncture trials for a systematic review. Her team of PhD and MD raters had more than 10 years of experience in clinical research and had worked on aspects of research methodology, with experience in acupuncture. Witt et al [63] also chose to use a score, 1-5, to allow score comparisons of inter-rater correlations and ensuring results could be presented as figures not just diagrams. Witt’s groups observed that the first round of scorings were highly heterogeneous but that inter-rater reliability improved following discussion of domains.

Harden who has a doctorate in human nutrition [64] worked with three raters who were also nutritionists to use PRECIS to rate pragmatism for 17 RCTs in a systematic review to determine ease of translation of a complex intervention to promote physical activity promotion. This group used a 0-4 scale with “0” completely pragmatic and “4” very explanatory which was the opposite of all of the other authors (though similar to the original Spokes presentation by Sackett on the preliminary work on PRECIS which the group used for training [65], in addition to the paper by Thorpe et al [48]. First nine trials were rated by three raters until there was 90% agreement then the remaining eight trials were randomly assigned to two authors. This group also used RE-AIM Framework (reach, effectiveness, adoption, implementation, maintenance) [66] as well as PRECIS to consider internal as well as external validity and ease of the intervention being adopted into usual care.

Sanchez used PRECIS for a systematic review on eHealth cancer prevention and control interventions (e.g. web based, online or text messages) of 113 RCTs using a total of seven raters with two raters
jointly assessing each trial using a 1-5 scale [67]. There was no independent scoring. Training involved
pilot of four trials to develop consensus. The average rating of domains for the included trials was 2.7-
3.6 so rather pragmatic. Percentage agreement scores for PRECIS domains ranged from 63.9 to 78.5%
with a median of 73.9%. Like Harden who also used PRECIS for a systematic review, Sanchez used RE-
AIM [66] to assess generalisability of results and translation into practice [9]. Re-AIM has subsequently
been reviewed alongside its use with PRECIS by Gaglio et al., including Glasgow who originally published
the framework [68]. They felt both tools were useful in designing studies to ensure more pragmatic,
and to assist translating into practice but training for both tools was important to ensure consistency
in domain interpretation, it was important to have subject expertise to work on project design and
groups may need to make some modifications in domains to assist with the design of different studies
[68].

In the fifth study using PRECIS for a systematic review, Yoong and a team of health researchers,
explored pragmatic and explanatory study design on outcomes of systematic reviews of public health
interventions [69]. Although the authors appear to only be aware of two other studies using PRECIS as
a tool in systematic reviews [39, 70] and suggested this was the first using PRECIS in a public health
systematic review, the publication by Harden [64] using PRECIS to analyse trials in a systematic review
of a complex intervention in physical activity promotion is public health and the systematic review of
electronic health strategies to prevent and control cancer also has a public health focus [67]. In this
study, two raters independently scored 55 RCTs with a scale of “0” (completely explanatory) to “4”
(completely pragmatic), before joint discussion to reach consensus. Overall kappa agreement ranged
from 0.23 (follow up) up to 0.75 (Experimental intervention practitioner expertise) they had difficulty
coding the flexibility of the complex interventions that were delivered by various practitioners. They
also found due to reporting issues, lack of information in six out of ten domains was a problem to score
and tried two methods – one imputing “2” to enable average score to be calculated, the second
method involved simply calculating the average score from the average of the available score. The
authors used three categories to describe trial pragmatism: explanatory if average scores for the 10 domains were between 0 and 1.7; combined explanatory/pragmatic if the average score was 1.7 and 2.2 and mostly pragmatic if the average score was 2.2 up to 4.0. Using method 1 meant that six studies (10.9%) had different classifications of pragmatism so Yoong emphasised the important of sensitivity analyses. They also looked at the effect sizes for Body mass index (BMI) of explanatory and pragmatic trials included in the review on obesity prevention, and found that overall the smallest intervention effects were in pragmatic trials (not significant) and largest intervention effect in explanatory trials. However, there was an overlap in effect sizes which was non-significant for the different study designs and the different age groups. The authors concluded that the impact of different domains on intervention effect size needed to be investigated further. The authors made recommendations that there was clearer domain guidance for using PRECIS (quoting the protocol on improving and developing PRECIS [52, 71] which was part of this thesis) and advised systematic reviewers to include PRECIS when writing the review protocol to ensure optimal data collection, they also advised that primary authors should be contacted to reduce problems with missing information.

**Use of PRECIS with a non-randomised study**

The final study of 15 methodological studies that have been published since the publication of the original PRECIS tool used PRECIS in a pilot study which was not randomized nor a clinical trial and was atypical because of its examination of two different interventions for complementary therapies for chronic lower back pain, without the intent of comparing [72]. Elder and colleagues used PRECIS to determine how pragmatic the study Kentucky Pain and Research Outcomes Study (KUPROS) was using “experienced researchers”. Each domain had a rating of 0-4 but they used cumulative scores for 12 PRECIS domains. So had a scale of 0 to 48; a higher score indicates a more pragmatic approach. The study modified the PRECIS wheel to 12 domains (as opposed to 10) inserting the “flexibility of the experimental/comparison intervention” and “practitioner expertise (experimental/comparison)” components for the PRECIS substituted massage therapy (MT) and progressive muscle relaxation (PMR) for “comparison” and “experimental,” respectively. To actually score the trial 14 raters with
clickers were trained in PRECIS then the trial was presented and raters independently rated each
domain of the study using “clickers”, there was no consensus discussion following individual ratings
but only six of the raters results were included as they described as experienced. The KUPROS study
had an overall score of 30.17 which the authors’ state is 63% of pragmatic-ness. Elder recommends
PRECIS is used a priori for design purposes as using the tool highlighted that that their study was not
as pragmatic as they had intended it to be thus enabling the team to discuss particular areas. Elder,
also suggested using PRECIS post hoc to assist designers justify decision making [72].

**Summary of methodological work on PRECIS**

To summarise, a diverse group of people have used PRECIS: clinicians, health researchers, nutritionists,
physiotherapists, pharmacist, social anthropologist/social worker and statisticians. All of the
methodological work on PRECIS has involved researchers using either Visual Analogue Scales or Likert
scoring systems to enable individual raters to compare decisions and measure how pragmatic or
explanatory a domain is. Most groups have used a 5 point scale with “0” or “1” being very explanatory
to “4” or “5” being very pragmatic. Most groups have also used PRECIS for consensus decision making.
These research groups all concluded that PRECIS was useful in designing trials, or to assess pragmatic
or explanatory trials included in systematic review or indeed in the study by Elder to compare two
different interventions in the pilot of a non-randomised study [72]. All of the groups stated that PRECIS
needed some modifications, many developed training and had team discussions of the domain scoring
and suggested that there was a need for more guidance on how to use the tool – one [69] quoting the
protocol [52] which the authors hoped would produce further guidance on the domains in the tool.
There was also some redundancy in PRECIS domains (Bratton combined *Experience of practitioner
expertise for comparison* and *Experimental intervention* as postulated no difference in expertise in
these trials [57]) and Witt removed *Primary Analysis* in the trial on Ayuverdic medicine [54] though no
explanation was given for this particular deletion. However in this trial the PRECIS model had 12
domains as her team then added patients’ compliance and practitioners’ adherence domains for both
Ayurvedic intervention and conventional care controls[54]. In another trial using PRECIS Bratton added the domain blinding [57]. For many there were unclear data on inter-rater reliability for the PRECIS domains indicating uncertainty in scoring and again confirming the need for guidance on the PRECIS domains. A summary of their key findings on how PRECIS was used and how the tool was modified is in Table 3[39, 53-55, 57, 58, 61-64, 67, 69, 72-74].
# Table 3 Summary of existing work with PRECIS involving 15 groups in chronological order

<table>
<thead>
<tr>
<th>Reference</th>
<th>Scale</th>
<th>RCTs</th>
<th>Protocols</th>
<th>No of raters</th>
<th>Consensus scoring</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Riddle (2010)</td>
<td>0-4cm circle Visual Analogue Scale (VAS)</td>
<td>X</td>
<td>1</td>
<td>7</td>
<td>YES (individual scoring then consensus score)</td>
<td>Useful to focus trial design discussion. PRECIS scoring: Initial, personal ideal and then post meeting. 1 day face-to-face meeting to discuss trial. PRECIS facilitates discussion. Trial design changed: domain “practitioner expertise” – needed to do psychologist training and domain “practitioner adherence” – rigorous on-going assessment to check intervention as intended. Visual analogue scale cannot be used if “online ratings”.</td>
</tr>
<tr>
<td>Bratton (2011)</td>
<td>No scoring</td>
<td>3</td>
<td>x</td>
<td>2</td>
<td>Joint discussion</td>
<td>Blinding inserted and combined experience of practitioner expertise for comparison and experimental intervention as postulated no difference in expertise in these trials. Useful for pinpointing weaknesses in design and considering applicability.</td>
</tr>
<tr>
<td>Tosh (2011)</td>
<td>1-5</td>
<td>X</td>
<td>10</td>
<td>3</td>
<td>NO</td>
<td>“Useful tool”. Cumulative scores for all 10 PRECIS domains. Experimental 0-15, Pragmatic &gt;35, 31-19 interim where trial balances pragmatic and explanatory domains. Scoring depends on rater’s perspective. 0 for missing information</td>
</tr>
<tr>
<td>Glasgow (2011)</td>
<td>0-4 scale</td>
<td>3</td>
<td>X</td>
<td>9</td>
<td>NO</td>
<td>PRECIS improved “transparency” in trial design decisions, encouraged others to use. Domain most variation: Primary analysis. Trialists rate own trial more pragmatic than other raters. Not clear if original criteria are sufficient to provide a comprehensive profile, used with RE-AIM to consider generalizability of results and translation in to practice</td>
</tr>
<tr>
<td>Reference</td>
<td>Scale</td>
<td>RCTs</td>
<td>Protocols</td>
<td>No of raters</td>
<td>Consensus scoring</td>
<td>Comments</td>
</tr>
<tr>
<td>-------------------</td>
<td>-------</td>
<td>------</td>
<td>-----------</td>
<td>--------------</td>
<td>-------------------</td>
<td>---------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Koppenhaal (2011)</td>
<td>1-5%</td>
<td>20</td>
<td>X</td>
<td>2</td>
<td>YES (individual scoring then consensus score)</td>
<td>“Useful estimate by estimating quantitatively how pragmatic each RCT is”. Chose PRECIS as explanatory/pragmatic continuum and visual analogue scale. Domains most variation: Practitioner expertise (comparison), primary analysis. Tried 1-10 score but too difference between consecutive scores not meaningful, still concerned arbitrary – important to reduce subjectivity 2 raters (3rd rater if not consensus). Weighting could be important; eligibility criteria important but flexibility of the comparison intervention may be less important. Problems using PRECIS due to reporting so CONSORT guidelines not being followed.</td>
</tr>
<tr>
<td>Witt (2012)</td>
<td>1-5</td>
<td>10</td>
<td>x</td>
<td>5</td>
<td>YES (individual scoring then consensus score)</td>
<td>PRECIS useful but needs further development. CONSORT guidelines for reporting pragmatic trials should be expanded. Recognised that PRECIS originally intended for trial design but useful tool for appraising published RCTs for systematic reviews.</td>
</tr>
<tr>
<td>Bratton (2012)</td>
<td>VAS or 1-10 scale converted into percentages</td>
<td>Ongoing</td>
<td>6</td>
<td>NO</td>
<td>“Useful tool for designing, conducting and reporting trials.” Strongest consensus ‘flexibility of the comparison intervention’ and ‘practitioner adherence’ domains. Most disagreement on ‘eligibility criteria’ and ‘participant compliance’ – ‘</td>
<td></td>
</tr>
<tr>
<td>Reference</td>
<td>Scale</td>
<td>RCTs</td>
<td>Protocols</td>
<td>No of raters</td>
<td>Consensus scoring</td>
<td>Comments</td>
</tr>
<tr>
<td>-----------------</td>
<td>-------</td>
<td>------</td>
<td>-----------</td>
<td>--------------</td>
<td>--------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Selby (2012)</td>
<td>1-20</td>
<td>X</td>
<td>1</td>
<td>6</td>
<td>YES (individual scoring then consensus score)</td>
<td>PRECIS useful to help interdisciplinary co-investigators rate their study design. Used two rounds of Modified Delphi process used to reach consensus. 20-point numerical scale approximated a continuous scale allowing “easier, more accurate and more stable coding of the response using email” – extreme anchor points 1-20, discouraged rating the domains beyond the numbers provided. Against aggregate scoring.</td>
</tr>
<tr>
<td>Harden (2013)</td>
<td>0-4</td>
<td>17</td>
<td>X</td>
<td>3</td>
<td>YES (individual scoring then consensus score)</td>
<td>Used PRECIS with RE-AIM to consider generalizability of results and translation in to practice. Recommend. “0” was completely pragmatic, on outside of wheel and “4” was completely explanatory (closest to the centre of the wheel) reverse to other authors. Three authors scored 9 trials – once &gt;90% agreement then remaining 8 trials randomly assigned to 2 authors.</td>
</tr>
<tr>
<td>Witt (2013)</td>
<td>1-5</td>
<td>X</td>
<td>1</td>
<td>7</td>
<td>YES</td>
<td>PRECIS gives good representation of planning to produce treatment that is patient centred and individualised but minimises bias. PRECIS adapted by removing primary analysis and including patients’ compliance and practitioners’ adherence domains for both Ayurvedic intervention and conventional care control.</td>
</tr>
<tr>
<td>Reference</td>
<td>Scale</td>
<td>RCTs</td>
<td>Protocols</td>
<td>No of raters</td>
<td>Consensus scoring</td>
<td>Comments</td>
</tr>
<tr>
<td>---------------------</td>
<td>-------</td>
<td>------</td>
<td>-----------</td>
<td>--------------</td>
<td>------------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Krist (2013)</td>
<td>0-4</td>
<td>x</td>
<td>1 (at end of design stage)</td>
<td>3</td>
<td>No</td>
<td>Scored trial design using PRECIS immediately prior to protocol publication. All of the authors (13) and collaborators participated in designing the study. The authors used PRECIS to demonstrate how pragmatic the trial was with three domains fully pragmatic (score 4) and the remaining seven domains (score 3) suggesting the results would be highly relevant and generalizable. And were more pragmatic than the PRECIS trial previously published by Glasgow 2011. *details from personal communication with A Krist</td>
</tr>
</tbody>
</table>
| Sanchez (2013)      | 1-5   | 113  | x         | 7 total (2 for each study) | No (independent scoring) | Recommend PRECIS to detect gaps in key translational issues. Used PRECIS to determine how pragmatic trials were in systematic review. Used PRECIS with RE-AIM to assess generalizability of results and translation in practice.  
  Average ratings of domains 2.7-3.6  
  Training involved pilot of 4 trials to develop consensus.  
  Weighted Percentage agreement scores for PRECIS domains ranged from 63.9 to 78.5 %, with a median of 73.9 % - good inter-rater reliability. |
<table>
<thead>
<tr>
<th>Reference</th>
<th>Scale</th>
<th>RCTs</th>
<th>Protocols</th>
<th>No of raters</th>
<th>Consensus scoring</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yoong (2014)</td>
<td>0-4 5 point scale</td>
<td>55</td>
<td>x</td>
<td>2</td>
<td>YES (individual scoring then consensus score)</td>
<td>Useful and promising tool to determine how pragmatic trials were in public health systematic review – Cochrane review of interventions.</td>
</tr>
<tr>
<td>Exploring the pragmatic and explanatory study design on outcomes of systematic reviews of public health interventions: a case study on obesity prevention trials</td>
<td><strong>Explanatory if average scores for ten domains between 0 and 1.7; combined explanatory/pragmatic if the average score was 1.7 and 2.2 and mostly pragmatic if average score was .2.2 up to 4.0.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Kappa 0.23 to 0.75. Difficult to code the <strong>flexibility of intervention delivery</strong> for trials that consisted of multiple components and were delivered by various practitioners, as is common in public health interventions. Lack of information a problem in 6/10 domains – used “2” if missing data.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Considering effect sizes: overall smallest effect size in pragmatic trials (non-significant) and largest effect size in explanatory trials but overlap in effect sizes for pragmatic, combination and explanatory trials.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Recommend clearer domain guidance and write using PRECIS into review protocol to ensure optimal data collection as well as contacting primary authors.</td>
</tr>
<tr>
<td>Reference</td>
<td>Scale</td>
<td>RCTs</td>
<td>Protocols</td>
<td>No of raters</td>
<td>Consensus scoring</td>
<td>Comments</td>
</tr>
<tr>
<td>-----------------</td>
<td>-------------</td>
<td>------</td>
<td>-----------</td>
<td>--------------</td>
<td>-------------------</td>
<td>------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Damschroder (2014)</td>
<td>1-5 (restrictive = 1 vs non-restrictive = 5)</td>
<td>x</td>
<td>1 (at end of design stage)</td>
<td>9</td>
<td>Yes (one rater then meeting to discuss scores)</td>
<td>Scored trial design using PRECIS immediately prior to protocol publication. Ratings subjective based on published guidance. The authors used PRECIS to demonstrate how pragmatic the trial was with seven domains as highly pragmatic (score ≥ 4). *details from personal communication with L Damschroder, using PRECIS-2 for next trial.</td>
</tr>
<tr>
<td>Elder (2014)</td>
<td>0-4</td>
<td>x</td>
<td>x</td>
<td>14 (but only 6 described as experienced researchers had their scores used)</td>
<td>No (independent scoring using response clickers at presentation)</td>
<td>Found useful to tool to highlight areas for discussion in study design - used PRECIS to determine how pragmatic study Kentucky Pain and Research Outcomes Study (KUPROS) was. Recommend a priori or post hoc. Cumulative scores for 12 PRECIS domains. Scale of 0 to 48; a higher score indicates a more pragmatic approach. Study was not randomized nor a clinical trial and was atypical because of its examination of 2 different interventions without the intent of comparing them so used modified PRECIS wheel with 12 spokes (as opposed to 10) inserting the “flexibility of the experimental/comparison intervention” and “practitioner expertise (experimental/comparison)” components for PRECIS substituted massage therapy (MT) and progressive muscle relaxation (PMR) for “comparison” and “experimental,” respectively. Study 30.17 which authors’ state is 63% of pragmatic-ness.</td>
</tr>
</tbody>
</table>
Strengths and Limitations of the systematic review
All of the authors of methodological articles who were contacted provided additional information and
many even gave their published study results for further analysis, demonstrating their interest in the
PRECIS tool. Many also stated they were interested in being involved further in the project to develop
PRECIS-2.

A search was not formally undertaken for grey literature on methodological work on PRECIS although
all authors who were contacted were asked if they knew of any other work in this field using the PRECIS
tool. In addition during a later stage of the project all authors who had cited PRECIS were asked if they
knew of any work using the original PRECIS tool. I was aware of the original tool being used in several
theses but this work was not incorporated (Private communication Nicola Blencowe - surgeon, Grace
Thompson – music therapist). I am therefore content that work to improve and develop the PRECIS
tool was based on all relevant methodological studies available at the time.

Strengths and weaknesses of the first PRECIS tool
As mentioned in Chapter 1, as indicated by the number of citations there was, and is, demand for this
sort of tool and the national funding body for health research in the UK has included it as one of eight
Useful papers. To date, (December 2014) fifteen research groups have done methodological work
based on, or about PRECIS indicated a significant amount of interest in the PRECIS tool. These research
groups all concluded that PRECIS was useful in designing trials [40, 53, 54, 57, 58, 61, 62, 73], or to
assess pragmatic or explanatory trials included in systematic review [39, 55, 63, 64, 67, 69]. PRECIS
has also been used with a non-randomised study comparing two complementary therapies for chronic
back pain [72].

All of these studies, however, highlighted weaknesses in the tool and suggested modifications. The
study teams all used a scoring system (which they invented themselves), developed training and team
discussions of the domain scoring and suggested that there was a need for more guidance on how to
use the tool. There was also some redundancy in PRECIS domains (Bratton combined experience of practitioner expertise for comparison and experimental intervention as no difference in expertise in these trial domains was postulated [57]) and Bratton also added a domain - blinding [57]. Inter-rater reliability was also unclear, important because PRECIS scores are dependent on the rater’s background and perspective and having to make a judgement call [55, 57]. Finally the original PRECIS tool has never been formally validated, which may prevent some potential users from becoming actual users.

Conclusions and Implication of findings

This systematic review of the published methodological work aimed to support trial development at the design stage by addressing weaknesses with PRECIS. This work was the basis for discussion in producing an improved and validated version of the PRECIS tool, PRECIS-2. Moreover, the new, modified PRECIS tool was then used to compare the internal validity of, and effect estimates from, a set of matched (by intervention) trials, some taking an explanatory approach, others a more pragmatic approach. A modified PRECIS tool had to be created to enable separation of explanatory and pragmatic trials in a consistent way.

The four related research questions that stem from this systematic review and are the basis of this thesis are:

1. Which domains should be removed or added to PRECIS?
2. What, if any, scoring system will we use for a modified PRECIS tool?
3. What is the validity and inter-rater reliability of the modified PRECIS tool?
4. Do pragmatic trials have lower internal validity and/or lower effect estimates than explanatory trials of the same intervention?
Chapter 3: Critique of potential methods to develop PRECIS-2

Introduction

Deciding how to undertake a study is critical to its success. This chapter will explore the different methods that were considered to improve the original PRECIS tool and validate PRECIS-2. Just as PRECIS was created as a tool to design clinical trials in a more transparent way, to design how to improve PRECIS and validating the new tool required careful thought and consideration. We wanted to determine the best and most methodologically sound way to undertake this task so that our end result, PRECIS-2 would be scientifically rigorous and easy to use.

Aims and Objectives

Aims

To determine the optimal way to improve PRECIS and test the validity and reliability of PRECIS-2.

Objectives

The aim of determining the best methodology to improve PRECIS and test the validity and reliability of PRECIS-2 was achieved through a number of objectives:

1. Systematically review the literature citing PRECIS

2. Decide who we wanted to consult to improve PRECIS and create PRECIS-2

3. Research the different methods to consult this purposively selected sample of participants

   Option 1. Interviews

   Option 2. Focus groups

   Option 3. Brainstorming

   Option 4. Nominal group technique

   Option 5. Delphi panel

   Option 6. User testing
Option 7. Additional data collection techniques

4. Select the best methodology to develop PRECIS-2

5. Determine the best sequence for the different research methods for the development phase of PRECIS-2

6. Determine essential requirements to test the validity and reliability of PRECIS-2

Data collection

There are a variety of ways to gather views from participants. In considering options we looked at a mixture of qualitative methods, with the possibility of incorporating mixed methods into our research methodology to improve PRECIS and test the validity and reliability of PRECIS-2. The priority was to gather information on what methodological research had already been undertaken on the original PRECIS tool [48, 49] so a systematic review (Chapter 1) was the basis for the work to develop PRECIS. This had to be kept up-to-date to ensure our work was informed by current methodological work on PRECIS.

Opinion seeking

Participants

Who we involved in our work was crucial to the project. Essentially, we wanted to have a wide consultation and involve anybody who could contribute to the improvement of the PRECIS tool. There were thus several groups that could be involved:

1. Everyone who had cited PRECIS – suggesting knowledge of the tool
2. The original developers of the tool – giving additional insight into the tool
3. Additional people who used the tool in lecturing and/or were trialists recommended by people in groups 1 and 2.
4. Clinicians, researchers and policy makers who may be interested in using the tool
Possible approaches

For the first phase of our work, to improve PRECIS, we decided to undertake qualitative research to gather information on the direction our work should take. As Green and Thorougoud state “…the aims of most qualitative analysis are to both reflect the complexity of the phenomena studied, and to present the underlying structures that “make sense” of that complexity (p175 [75]).”

There were seven main options for gathering opinions on how to improve and develop PRECIS-2:

- One to one interviews (Option 1) including user testing (Option 6)
- Group interviews:
  - Option 2: Focus groups
  - Option 3: Brainstorming
  - Option 4: Nominal group technique
  - Option 5: Delphi process
  - Option 7: Other options

Each of these options is now discussed in turn.

Option 1: Interviews

General

The first option that was considered to gather opinions was to consult key informants on a one to one basis. Using this technique it would be important to consult a variety of people about PRECIS to ensure the tool would be as useful and appealing to as many users as possible. Thus, careful selection of the sample would be important. Another issue would be determining how many interviews to carry out. If the interview sampling procedure was methodologically rigorous then interviews should continue until there was data saturation. Thus, in gathering ideas to improve PRECIS this would mean that were no more new or relevant suggestions to be gleaned from consulting others [76]. With purposive sampling
though it has been found that twenty interviews is usually enough to collect all the views on a particular subject [77].

To ensure uniformity in the interviews, the same interviewer should be used (KL) and the interview should be structured with a topic guide including a set of questions that are usually asked in the same order. In this type of research it is possible to directly compare the data obtained from interviews. If the interview was not structured, and one question was asked by the interviewer to one interviewee but not another participant then there would be variability which would make it harder to draw conclusions from the results, as the answers may not be representative of the group.

Of central importance, in using the interview technique was to prepare the questions that would elicit the answers that were needed to address the research question. KL based questions on problems that had arisen from the systematic review and methodological work using PRECIS (Chapter 1). The questions in the topic guide were framed as open questions to encourage interviewees to openly give their opinions and beliefs on how PRECIS could be improved. It was important not to use questions that indicated the direction the researchers (KL, ST, MZ, FS) anticipated the project would take.

As important, however, was the technique employed by the interviewer KL. Suggestions for good interview skills included not interrupting when interviewee speaking, ensuring longer gaps between questions to encourage opinions to be fully explained, adequate prompting to elicit as much information as possible, and non-judgemental interviewing technique [77]. Equally important was establishing a rapport. A commonality with all interview situations is that the researcher, as interviewer, has an input on the data collected.

**Pros**

Semi structured interview is a technique which can reveal issues that might not be uncovered using a topic guide and structured interviews. This can optimise the information from a mixed group of
individuals. Everyone has different experiences and this can be hard to tap into unless the interviewer has an idea of possible answers and has a clear idea of what they want to know the answer to. Using one to one interviews at the beginning of the process to develop PRECIS, semi-structured interviews may be helpful to gain as much insight as possible into developing the tool.

**Cons**

1. Usually interviews are no more than 45 to 60 minutes long. Any longer and both the interviewer and the interviewee lose concentration. One of the crucial difficulties in using this research method to gather information with a mixed sample of users about PRECIS was the overall time it would take. Appointments would need to be organised with stakeholders and then further time is needed to assimilate the opinions from the individuals to determine if there was consensus or indeed a brainwave idea had been propositioned.

2. It is possible that a structured approach to interviews may constrain discussion, rather than a completely open discussion. However, we used open questions to give interviewees the opportunity to speak on issues we may not have explicitly invited them to comment on. There are downsides to this too. Even though open questions are used to elicit an honest response there is a danger that the researcher (KL) may influence the answer through inadequate framing of the question and by appearing to judge the response. Thus influencing the interviews in the direction that KL wanted them to take. Ensuring all questions are pilot tested and using an observer during the interview should help reduce this possibility.

3. As the individuals we were interested in recruiting were experienced there may be problems recruiting from a researcher (KL) that is less well known. This may be overcome through using the names of the more experienced supervisors for the project (ST, MZ, FS) [78].

4. Establishing a rapport can also be tricky as the power balance between interviewer and interviewee [79] can create problems in interviews. In this case, many of the researchers are extremely busy and while they have some familiarity to a greater or lesser extent on PRECIS, they are more experienced than KL so there may be a power imbalance which could be a problem in the interview process if KL is
not confident in asking the questions. This “elite interviewing” [80] is different to the usual scenario of researcher and a member of the public for instance where the power imbalance would be more likely to be reversed.

5. The setting and how the interview is carried out may be important. As many of the people that KL would consider interviewing are international researchers then personally interviewing in a one to one interview would be impossible. Thus alternatives like Skype would be important. This may also allow interviewees to be more comfortable and accessible for interview, as they will be answering questions in their own “space” so the only unknown is the questions that will be asked. It is much easier to decline or minimise interaction if interviews take place online and not face to face. Thus the building of the relationship, preparation for the interview and recruitment, as well as the interview technique of the interviewer KL may become very important in drawing out as much information as possible.

Conclusions

Due to organisational time to prepare and pilot test the topic guide, arrange interviews with participants the researcher KL did not know, faster ways of assimilating information were considered involving discussions with more than one person at a time. Thus we decided that we would have to organise a group interview to gather our data on how to improve PRECIS. This would allow faster progression of the work to develop PRECIS. However one to one interviews were used later, when we knew exactly what we wanted to ask participants through “User testing” (see later Chapter 6)

Option 2: Focus groups

General

Focus groups have been used for over a hundred years since 1926 when Emory Bogardus described group interviews in his qualitative research [81]. They are frequently used in health and social sciences and marketing to gauge public opinion about a new product. Like one to one interviews they involve exploring a topic but in a group with a facilitator. This enables individuals to hear each other’s opinions and discuss with each other in an interactive setting. Ideally groups would be 6 to 12 individuals to
facilitate responses from all of the group participants [82]. Of importance is creating a positive atmosphere to encourage sharing of information and responses to questions.

To facilitate optimal sampling of ideas, focus groups should be undertaken until there was data saturation and no further ideas were generated [76]. As in one to one interviews, a topic guide would assist consistency in the questions asked of the different groups and ensure the time was used well by the facilitator who was not distracted by irrelevant issues that arose during discussion. As described before in one to one interview, the framing of questions could be significant. So whether or not we used a more informal approach and semi-structured agenda with issues that we wished to cover with focus group participants was also of consideration.

Although, group interviews can be described in several ways, selecting which format to use to generate the data we needed was important (Figure 3).

**Pros**

1. In focus groups, ideas that were generated by one participant could then be developed by another and discussed. That is one of the key advantages of a focus group that participants can respond to each other’s views and discuss them, moving the discussion into areas that might not have been envisaged beforehand [83]. The focus group facilitator (KL) would, however, have to guide the discussion so the group kept to the remit and the questions posed.

2. Focus groups can enable exploration of topics that are difficult to discuss in one to one interviews, e.g. abuse, however discussing PRECIS would not involve such issues.

3. Running a series of focus groups would have speeded up the process of information gathering on PRECIS, when directly compared to one to one interviews. They are thus a fast way of exploring a topic compared to one to one interviews.
## Advantages

<table>
<thead>
<tr>
<th></th>
<th>NGT</th>
<th>Brainstorming</th>
<th>Delphi</th>
<th>Focus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficult for dominant participants to control</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Possibly</td>
</tr>
<tr>
<td>Avoids 'quick decision making'</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Possibly</td>
</tr>
<tr>
<td>Generates a high number of comments/ideas</td>
<td>Yes</td>
<td>Possibly</td>
<td>Yes</td>
<td>Possibly</td>
</tr>
<tr>
<td>Provides support to allow identification of personal problems and self disclosure</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Allows measurement of importance of ideas/items to individuals</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Possibly</td>
</tr>
<tr>
<td>Avoids pursuit of a single train of thought ('focus-effect')</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Encourages minority concerns/options to be voiced</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Possibly</td>
</tr>
<tr>
<td>Participants value social interaction i.e. group cohesiveness</td>
<td>Possibly</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>High degree of task completion</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Possibly</td>
</tr>
<tr>
<td>Ease of administration</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Possibly</td>
</tr>
<tr>
<td>Need for experienced leader</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Figure 3 The relative strengths of focus groups, Brainstorming, NGT and Delphi (from Gallagher et al [84])

### Cons

1. As many of the participants were not known to the researchers (in particular KL), arranging a series of focus groups in a timely manner posed difficulties.

2. Careful planning would need to be taken to ensure the focus groups were a success with the optimal number of people (6-12), a good balance of people so they all know each other, and use of a topic guide to ensure everything is covered and the time is used well.

3. An important issue to consider would again be the purposive sampling of participants to gather a wide range of responses from individuals. There can be difficulty ensuring good group dynamics among the participants. Focus groups with strong personalities can inhibit quieter individuals who have
equally important contributions. Ensuring there was not too much heterogeneity in group members and everyone contributes could be assisted, however, by the focus group facilitator (KL) and co-facilitator (ST, FS or MZ).

4. The setting for the focus groups was also important, online meetings for focus groups were considered but there are difficulties in discussing important issues (e.g. what makes a PRECIS domain more explanatory) with researchers where there has been limited contact beforehand. A serious issue that limited use of focus groups was the expense of getting participants together (with no more than 6 to 12 participants in a group). All of the individuals that KL wished to consult were spread around the world so consulting them in a face to face focus group was not possible without some modification of the methodology.

Conclusions

There were clear advantages of running a focus groups over one to one interviews but as face to face focus groups were hard to organise due to the financial costs, other alternatives had to be considered. However, gathering a select group of key informants in a face to face meeting appeared to be an important step in developing PRECIS-2.

Option 3: Brainstorming

General

The creative thinker Alex Faickney Osborn, manager of an advertising agency, used the term “using the brain to storm a creative problem” in a book published in 1942 called “How to think up” [85]. Since then brainstorming has become widely used and most people, regardless of language, understand the term “brainstorming” [86]. It is often thought of as another type of group interview which has many similarities to focus groups but is typified by much more free flowing discussion to enable more creative thinking to produce solutions to problems [82]. However, even the use of the online social networking service Twitter can start the process of brainstorming as people “tweet” and interpret and give their opinion on the topic being discussed [86].
Face to face brainstorming works best if there is a semi structured group interview with an agenda and topic guide [82]. For the development of PRECIS-2 this was based on published weaknesses by users of the original tool.

**Pros**

1. The key strength of a brainstorming meeting is the freedom to be creative in thinking up ideas on a specific topic (Figure 3).

2. The participants included in the brainstorming group are highly influential in determining if the aim of the meeting will be met. This will depend on who is invited and who can attend. A highly motivated interested group can interact with each other to produce an intense atmosphere that is the ideal environment for creative thinking, listening to each other and moving the discussion forward, clarifying uncertainties.

3. Using a facilitator (KL) who was not an “expert” in the area and as familiar with the development of the original PRECIS tool as the co-assistants, and was a junior researcher, potentially gave a more neutral approach encouraging all the participants to fully express their opinions and not be inhibited by status [82].

**Cons**

1. Preparing a group interview requires careful organisation to ensure the aim of the brainstorming meeting is fulfilled. If not then the effort of gathering a group of experts may not be worthwhile (Table 3.1).

2. The facilitator is not just a note taker but of central importance in the brainstorming to ensure that the conversation does not get too focussed on one area and all alternatives are considered as mentioned previously in focus groups [84].

**Conclusion**

Following a systematic reviewing of the methodological work on PRECIS and having considered previous options for gathering ideas, KL selected a “brainstorming format” to ensure a more free flowing discussion to develop PRECIS-2. Brainstorming is usually connected with the generation of
ideas – however, in the development of PRECIS-2 the evidence base for this work was the systematic review, meaning there were already ideas to fuel the improvement. We did, however, want brainstorming with a range of stakeholders to pull these ideas together (rather than generate the ideas in the first place) and translate them into an innovative product that would be easy to use as well as genuinely helpful in the creation of relevant research.

KL organised a local (Dundee, Scotland) brainstorming meeting to discuss plans for developing PRECIS. As KL prepared and facilitated the local University of Dundee brainstorming meeting this gave KL experience organising the larger brainstorming meeting in Toronto. This was an international brainstorming meeting (Toronto, Canada) including people from North America, both USA and Canada where the tool originated. For this second brainstorming meeting, KL invited researchers whom had been involved in the original PRECIS tool development, researchers who had undertaken methodological work using PRECIS, and potential users of the PRECIS-2 tool we were developing – medical practitioners and policy makers. A third and final meeting with local trialists in Dundee, after the Toronto meeting, assisted with discussion of themes and proposed models for PRECIS-2. This further forum for open discussion allowed the Steering group (KL, ST, FS) to test out ideas with a different audience in the UK and openly discuss whether or not they were realistic and helpful for the project.

**Option 4: Nominal group technique**

**General**
This technique was first described in the late 1960s in the USA for exploratory health studies but it was also used for aerospace design projects [87]. Nominal group technique (NGT) is a consensus panel approach and entails carefully managing the input from a group of participants to answer a question; participants would individually generate ideas which would then be discussed and then participants would anonymously vote to prioritise options for improving PRECIS [82, 88, 89]. The group is “nominal” as the individuals in the group are gathered purely for the purpose of answering the question and
individuals do not normally work with each other Table 3.1 [82, 84]. This process is very structured and can involve several round of voting until consensus is reached [89] or indeed several NGT groups at different locations meeting to vote on the same question [87].

**Pros**

1. Nominal group technique allows participants in the group to be anonymous and prevents dominant members having undue influence (Figure 3) [82, 84].

2. Many of the individuals we would be inviting to participate do not usually work together so this research method could be the most effective way for participants to contribute as individuals independently give their views and do not need to collaborate.

**Cons**

1. Problems with employing this technique are that it takes time to select and organise participants to meet several times. Several meetings may be required to vote on best strategies to improve PRECIS.

2. The nominal group technique is very structured and the formality of voting and grading different options to improve PRECIS may not suit everyone. We did not want to prioritise options for developing the tool for designing trials as we did not believe it was constructive use of time for participants to indicate one suggestion was better than another. We thought it was likely that there would be several options to improve PRECIS that would be employed which could be sifted through during the brainstorming meeting and then after the meeting by the steering group.

**Conclusions**

We decided that we did not want the structure and formality of a nominal group technique and organising several meetings in Dundee and Toronto would not be feasible and a constructive use of participants’ time.

**Option 5: Delphi process**

**General**
The Delphi process is another consensus panel approach and was first used during the Cold War in the 1950’s and 1960’s by the Rand Corporation to gather expert anonymised consensus opinion from military personnel on the impact of technology on warfare[90]. The name Delphi stems from Ancient Greece and the priestess who worked in the temple of Apollo at Delphi. She was the most important Oracle in Greece who could predict the future. The name Delphi is perhaps a misnomer as it could imply the process is arbitrary in nature with advice from only one person instead of well-grounded “suggestions” from a panel of experts.

Using a Delphi consensus technique, participants go through several rounds of questions on a particular topic with anonymised feedback on answers given by individuals in the group. Participants are then invited to respond and can change their opinions until there is consensus in the group [82]. A modified Delphi may start with a first round of preselected topics based on a literature review but consensus can still be reached [89]. It is possible, however to simply have two or three rounds of a modified Delphi, and not use the Delphi technique for consensus due to limited time [91]. This technique has been used in different settings to pool consensus opinion from experts. A 3-round Delphi was used to generate consensus from an international group of patients, researchers and clinicians on a Core Outcomes Set (COS) for trials on Non Specific Lower Back Pain (NSLBP) as a heterogeneous set of outcome measures was hindering progress in pooling results and meta-analysis and thus treating this debilitating condition [92]. In contrast, in primary care practice, a 2-round email Delphi asked General Practitioners to prioritise statements that they would be more likely to undertake in a pragmatic trial to test part of an intervention for a new feedback system to address problems of high risk prescribing behaviour[93].

A recent paper by Bloor (2013) suggested that a modified Delphi using an electronic poll is a useful addition to mixed research methods. Using purposive sampling to give a balanced view with interested participants and a clear remit gave timely feedback and can give added value to a research study. Bloor
et al also suggested the facilitator (or steering group) is key to deciding which opinions to develop further or ignore based on alternative evidence from other sources [94].

**Pros**

1. Undertaking a Delphi survey could be used to relatively quickly gather anonymous qualitative and quantitative data on views to improve PRECIS [95, 96].

2. A Delphi survey is cost effective and a good vehicle for consulting informed individuals, allowing everyone to express their views without being unduly influenced by others (Figure 3) [97].

3. Avoids organising a face to face meeting in the first instance.

4. Using an electronic survey for the Delphi questionnaire KL could also take this opportunity to enrol participants for all phases of the project and using the Delphi process develop a rapport with participants.

**Cons**

1. Using the Delphi process to elicit responses requires a well-structured questionnaire to get meaningful results to improve PRECIS.

2. There is a need large enough sample size to produce useful and informed opinions that can be used in future stages of the project to develop PRECIS-2. There is uncertainty what is the ideal sample size but KL needed to ensure invited everyone who could contribute to the project to improve PRECIS, if not there is a possibility of limiting ideas for improvement.

**Conclusion**

We decided to employ the Delphi process prior to the main brainstorming meeting in Toronto, as we agreed with the peer review from the funder for the work (CSO) advised this would give a portfolio of research methods to better inform the development of PRECIS-2. This methodology enabled us to prepare up-to-date information on the opinions of users of the PRECIS tool so we could maximise our time with the individuals we were consulting in the brainstorming meeting in Toronto and focus on issues where there was no consensus. We thus used the Delphi process following the Dundee
brainstorming meeting. We would use the systematic review of the literature on the methodological work published on PRECIS as a basis to collate opinions on suggestions to improve the tool (Chapter 2). We believed two rounds were adequate as we had planned brainstorming meetings and we were aiming for general agreement for the Delphi and not aiming for consensus.

**Option 6 - User testing**

**General**

This research methodology is based on product development and builds on the interview technique discussed earlier. This is, however, slightly different to a simple interview technique involving posing a series of questions from a topic guide. It involves presenting information to “users” and asking them the “think out loud” as they consider the information in front of them [98]. The meeting is usually audiotaped and notes are taken to assist in analysis, often by an additional person. This technique is used to check understanding and ease of using a product for the purpose it was designed. Feedback can then be analysed and then the product improved before further testing. Testing is therefore iterative until the product is user friendly and there is a high degree of satisfaction using the product with no problems interpreting information[99].

User testing gives a greater understanding of a product and it is a way of getting further suggestions for improving a product. It also lets the tester know if this is a product that the user would want to use in addition to being able to use, two different things. The tester in particular is looking for (1) “show stoppers” where is a complete lack of understanding; (2) big problems where the “key information is too long and difficult to read”; (3) minor issues or cosmetic issues; (4) positive feedback; and (5) suggestions for improvement [99].

It is important that the people in user testing are representative of the people who will be using the end product. The sample size should be dictated by saturation of findings, as in group interviews with
focus groups, user testing should be stopped when think there are no more changes that need to be made to the product [99].

Pros

1. User testing can check understanding with everyone who will use a product as there may be a range of experience.

2. User testing with potential users, outside the steering group, can overcome subjectivity due to overfamiliarity with a product.

3. User testing gives insight into how others will use a product and how they anticipate it will assist them.

4. Instant feedback on what user is thinking; it is immediately clear if the information is misunderstood and there is a high degree of DIS-satisfaction.

5. Time for user testing is no more than 60 minutes which is similar to a meeting time, making it easier for user tester participation.

6. Once user testing material is prepared, with clear questions to obtain maximum amount of information in a short space of time, this is an efficient research method.

7. Technique can be used online with Skype or face to face whichever is easiest for the user.

Cons

1. Preparing user testing material to make it as simple as possible needs to be carefully done to minimise frustration and ensure that the product is tested and not the way of presenting the material.

2. There is no interaction from the “tester” - the person presenting the product to the “user”, so there can be no deviation from the material being tested. The “tester” is simply there to user test and take notes.

Conclusion

We decided not to do interviewing on a one to one basis at the beginning of the project as it would be an inefficient use of time. We decided, however, to use individual interviews at the end of the tool development, using user testing, to check the model that was being created for PRECIS-2. Interviewing
is one to one and by using the technique of user testing KL could consult potential users of the PRECIS-2 tool and get their feedback on the version of the tool that was under development.

In User testing, PRECIS-2 participants – users – looked at the questions, tool domains, descriptions, PRECIS-2 tool, with the interviewer (KL). Comments and feedback through this process created different iterations of PRECIS-2 as we progressed towards the creation of the final product [98, 99]. Further information on this methodology is discussed in Chapter 6: User testing PRECIS-2 models.

**Option 7 - Additional data collection techniques**

There are a number of other methods to determine how to improve PRECIS and while these were not used in the project to improve PRECIS we could have considered them. Investigating creative directions for groups to use the available research evidence, in this case on PRECIS, could have been explored through mirroring some of the procedures used by Fourie [100]: mind mapping, force field analysis:

1. “Mind mapping” is a technique for organising ideas that could have been used to assist brainstorming [http://en.wikipedia.org/wiki/Mind_map](http://en.wikipedia.org/wiki/Mind_map). It is a way of sorting information and from the perspective of PRECIS-2 could have been used to graphically display the various ideas for developing the tool, perhaps for and against making changes to create PRECIS-2.

2. “Force field analysis” - Analyzing the Pressures For and Against Change to make decisions could also have been helpful in deciding how we were going to improve and modify PRECIS. [http://www.mindtools.com/pages/article/newTED_06.htm](http://www.mindtools.com/pages/article/newTED_06.htm).

3. Another technique “De Bono's 6 hats” is a tool to help participants brainstorm and could have been used to help decide on what PRECIS-2 should look like. It has been used by successfully by investment companies but uses colours of different hats to consider the facts (white), the positives (yellow), negatives (black), intuition/feelings (red), creating new ideas (green) and managing the process for developing the tool (blue) [http://www.debonogroup.com/six_thinking_hats.php](http://www.debonogroup.com/six_thinking_hats.php).
We decided against using any of these more complicated ways of brainstorming as there was limited information about using these methods in health research. In addition, the participants would almost certainly be unfamiliar with these techniques, making it harder to engage participants. Instead we used methodology that had been well tested in health care research (Option 3).

**Order of the PRECIS development phases**

Figure 4 indicated the order of the different research methods using to improve PRECIS and develop PRECIS-2, concluding with reliability and validity testing of PRECIS-2. We initially decided to start the project with brainstorming (Option 3), in Dundee with local expertise from the University of Dundee to discuss possible changes, this also enabled KL to gain experience facilitating. The Delphi process was then used to crystallise ideas with expert anonymised input (Option 6). We believed this would be a useful additional research tool to collect ideas to improve PRECIS based on the systematic review of the methodological work on the PRECIS tool (Chapter 2). These three research processes would increase the likelihood of success for the brainstorming meeting in Toronto - the latter involved some of the original developers of the tool and others who had undertaken methodological work on PRECIS.

Using a brainstorming meeting to brief everyone on the facts and then letting the conversation flow, allowed everyone in the room to participate and discuss the direction the work to improve PRECIS would take. The ideas generated by this second Brainstorming meeting in Toronto were then accepted or disregarded by the steering group (ST, MZ, FS, PD) lead by KL. As the brainstorming meeting in Toronto provided a lot of rich material we felt it was important to have a local brainstorming meeting in Dundee - post-Toronto - inviting some of the original participants or substitutes who we believed could also guide the PRECIS-2 tool development. Thus we had three brainstorming meetings to inform decision making before user testing of different PRECIS-2 models commenced, prior to validity and reliability testing.
Figure 4 Flow diagram of research methods during PRECIS-2 development
Validity and reliability testing

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 would test the face validity and inter-rater reliability. The methodology for all validity testing is based on sample size calculations. It was thus essential that the sample size was large enough, had statistical power, to detect if the effect size is statistically significant or not [101]. These calculations do not vary but are dependent on the statistical test, in this case Intra-class Coefficient [102]. We wanted to test the inter-rater reliability of the PRECIS-2 tool. 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. Further details on sample size calculations and study methodology for validity testing, including inter-rater reliability and discriminant validity will be discussed in Chapter 7.

Conclusions

Through combining a mixture of qualitative techniques and quantitative research we believe that we determined the best option for the design of the PRECIS-2 to be selected. The diversity of our methods supports and gives strength to the overall study. This is important as suggested by Barbour (1999), using different techniques can reinforce the results from research and make sure that “no stone is left unturned” if there is wide enough consultation with suitably qualified participants [103]. Using the Delphi process, Brainstorming, User testing prior to Validity and Reliability testing gave a comprehensive multi-method research project with the results from each stage feeding into the next to create PRECIS-2. Our intention was that this tool will be simple and easy to understand for trialists - both new and experienced - to design trials fit for purpose.
Chapter 4: Gathering improvements for PRECIS from experts - A Delphi process

Introduction to the Delphi

PRECIS was published in 2009 as a conceptual tool that was recognised to be in need of further development to place it on a more solid empirical footing. Subsequent users of the tool (Table 3 Chapter 2) who had published methodological work using PRECIS stated how useful the tool was in developing their trial design or in assessing trials for systematic reviews but all of these groups made suggestions to improve the PRECIS tool. In particular, the choice of PRECIS domains and their description were debated and the lack of a scoring system in PRECIS forced users to invent their own, to enable use of the tool. These were good pointers to direct work on developing the next version of the tool but we wanted to determine if there were any additional ideas to improve the original PRECIS tool. As there were an increasing number of international researchers, methodologists, trialists and health professionals who were citing the tool, we decided this would be an ideal group to invite to assist with this task. With such a diverse group, we hoped for an assortment of ideas to develop the original tool to make PRECIS-2 as easy to use as possible and to make it more useful for creating clinical trials fit for purpose and assisting trialists to consider the end users and applicability.

To determine PRECIS improvements and priorities we decided to select a Delphi approach (see Chapter 3 Methods) since it enabled us to obtain anonymous input from the purposively selected participants using an electronic survey. These ideas could then be discussed in a subsequent brainstorming meeting. We considered this simpler and easier than embarking on the more time consuming formal structured nominal group technique. The latter is a staged process, at its simplest is a managed discussion to prioritise ideas the group supports but at its most complicated it can involve time consuming qualitative analysis of themes from interviews which are then prioritised through a series of group discussions [88].
We were aware that by asking a diverse group of people for their opinion on ideas to improve the original PRECIS tool we would get different, perhaps conflicting ideas. The steering group’s role was to sort out the ideas into what would be feasible and genuinely improve the tool for everyone using it, and not get side-tracked with ideas that would have limited applicability to all trialists. A modified Delphi of only two rounds could be accomplished relatively quickly and enable further discussion with an agenda for discussion that would not have been easily reached in several brainstorming sessions, with a more limited group of participants. We were not aiming for consensus in using a Delphi process; it was modified to two rounds to allow quick ascertainment of the agreement of the key issues that participants believed would improve the original PRECIS tool. These could then be discussed in the next brainstorming step in developing PRECIS-2. We thus did not set any a priori decisions about reaching consensus, for instance from 70% of participants [104].

However, we were wary of simply undertaking a consensus Delphi due to the risk of developing the tool in a direction that would minimise its use either due to complexity or including items that were not of interest to potential users, if misled by the majority opinion as described by Mullen (2003) [105]. In other words, consensus is not always what is required for a useful tool. We wanted “to find the correct answer, whether or not it is an outlier, rather than a unanimously agreed wrong answer”[105].

Dave Sackett, an important collaborator and proponent of the original PRECIS tool, quoted Linus Pauling to sum up our views "Truth is not often best determined by a majority vote" (personal communication). Thus we decided to conduct two rounds of an electronic modified Delphi survey over a three month period and use this during further brainstorming in face to face meetings with past and future users and designers of the PRECIS tool in order to bring independent voices to bear on what might otherwise be an unwarranted consensus.
The Delphi technique [106] provided a framework to consult “experts” with knowledge of PRECIS who might be interested in improving the tool in an anonymous but relatively quick way, which then avoids the huge expense of travel. Dalkey and Helmer first proposed this methodology to get anonymous feedback from experts using questionnaires thereby avoiding face to face discussion [90]. Anonymity in a Delphi survey enables participants to freely give their opinions without being influenced by others. However, while the Delphi technique can be used to establish consensus, it can also be used in the first round to determine the degree of polarisation and inform the group of divergent opinions [105]. Subsequent rounds give participants the opportunity to change their opinions and reach agreement [88]. This technique has been used in health research in a variety of studies as the basis for informing decision making. For instance, a similar modified 3-round Delphi to PRECIS was undertaken for CONSORT to determine the minimal information that should be reported in journals on Randomised Controlled Trials [107]. A 3-round consensus Delphi was also used for SPIRIT (Standard Protocol Items: Recommendations for Interventional Trials) Initiative in 2007, with the primary aim of increasing the transparency and completeness of trial protocols [108]. Recently, a 3-round online Delphi was used to determine the priority research agenda of the 48 UK clinical trial units (CTU) registered at that time [104].

All of the aforementioned have enabled anonymous opinions to be gathered to inform further work but pre-defining the number of Delphi rounds - often called a modified Delphi.

**Aims and Objectives**

**Aim**

The aim of the Delphi was to collect ideas for discussion to improve the PRECIS tool and to recruit participants to each phase of the project to produce PRECIS-2 (user testing and validity testing).
Objectives

The Delphi process consisted of the following stages.

1) Recruit participants who could give useful informed contributions

2) Explore the views of Delphi participants in a 2 round modified Delphi to generate a range of ideas to address the issues uncovered by the review (see Chapter 2) of researchers, methodologists and health professionals who have cited the PRECIS article

3) Make concrete recommendations to improve PRECIS-2 that could be discussed in a brainstorming meeting with interested trialists and policy makers some of whom had been involved in the original PRECIS (2009 tool)

4) Recruit participants who could be contacted at subsequent stages of the work

Methods

Participants eligibility criteria for the Delphi

To optimise input to a modified DELPHI and collate independent views we invited all of the contact authors who had cited PRECIS. However, we excluded the original developers of PRECIS and authors that had evaluated the utility of PRECIS, as they were consulted in brainstorming sessions. We believed that by inviting all researchers who had cited PRECIS we were able to increase the likelihood of including in the consultation many, if not most, of the people who had thought about PRECIS and how to use it, but had not been influenced by the developers’ preconceptions. A modified Delphi enabled us to collect these opinions and help us understand how we could go about improving the tool, as suggested by Pill (1971, p62)” and “one asset of the use of a group is the diversity of opinion they bring to bear thus minimising the possibility of overlooking some obvious facet of a question” [109]. Ninety authors who had cited PRECIS from Australia (7), Europe (38), North America (47), and South America (1) were contacted to participate in two rounds of a modified Delphi. Anonymity between participants was maintained throughout the modified Delphi. We also encouraged participants to share the
invitation with colleagues who were interested in assisting with the PRECIS project, believing they may have additional views but similar careers to the group that had cited the original PRECIS tool.

**Design of PRECIS questionnaires for the Delphi**

Firstly, KL established a steering group (KL, ST, FS, MZ, PD) led by KL to coordinate the Delphi and different stages of the project. Secondly, we designed the first round of a modified Delphi consensus method [96] to discuss improving the current PRECIS tool and create PRECIS-2. We generated a web-based questionnaire using the SurveyMonkey® Easy-to-use web-based survey tool “select” version (i.e. not the free version). We decided to use this version of SurveyMonkey®, so that we had question numbering, page numbers, survey completion progress bar and skip logic, and could brand our survey with a logo, custom url (uniform resource locator or web address) and print out the questionnaire. We wanted to ease the burden of participants assisting in the project and make the survey quick and simple.

The initial round of a modified Delphi process can be a structured questionnaire if there has been a systematic review of the literature and there is adequate information to base the survey on [95, 97]. Our questionnaire was based on problems that had been identified by eight groups that had used PRECIS prior to the Delphi survey [39, 40, 53, 55, 57, 58, 70, 73] (see Table 3, Chapter 2) and the report from a brainstorming meeting at the University of Dundee in June 8th 2012 (Table 4 taken from Chapter 5 Brainstorming). (Seven groups subsequently published using PRECIS following our Delphi Survey [54, 61, 62, 64, 67, 69, 72]) Topics that were discussed in the survey included Scoring, Domains and Design. There was an introduction to the survey and all participants were advised that their assistance in participation in the survey would be acknowledged in any future publication if they left contact details. We were careful in designing the questionnaire not to lead respondents by, for example, keeping questions open to gain maximum information, allowing us to develop a full range of PRECIS alternative models for user testing and “explore or expose underlying assumptions”[110].
The survey was designed to be completed in approximately 10 minutes. It was not compulsory for participants to answer any of the questions. All participants had an opportunity to leave comments as well as multiple choice answers of yes, no or not sure (if appropriate). Responses from individuals were known only to the moderator (KL), so participant anonymity and confidentiality of responses was ensured. Anonymity is important in the Delphi process as it prevents rank and position, strong personalities and peer pressure having the detrimental effect that can occur in face to face meetings [105]. Reasons for declining participation were collated if solicited.

**DELPHI survey – Round 1 Design**

The first round (Appendix – Chapter 3 Delphi) included 15 questions: two introductory questions to gather information about the participants’ experience of using PRECIS to design a clinical trial and whether they liked the current tool; eight questions about the format of the PRECIS tool – wheel or table, scoring system and domains and weighting; and finally five questions about assisting in a future round of the modified Delphi, and assisting with further development of the tool with acknowledgement of help in a future publication if the participants left their name.

The second round included eight questions: five questions giving feedback on the first round as the basis for further questioning about the format of the PRECIS tool – using both wheel and table format,
incorporating suggestions for a scoring system, aggregate scoring which was discussed in brainstorming and mentioned in first round in free text comment, additional domains and weighting; and finally two questions concerning whether or not participants wanted a summary report of the Delphi or wanted to add something that had not been covered previous questions. We also included a question on asking guidance that had arisen from PRECIS utility publications and experiences of the steering group using PRECIS in workshops.

**Pilot for the Delphi**

For both rounds of the modified Delphi, we pilot tested the SurveyMonkey® questionnaire twice, with the PRECIS steering group and then with the four participants involved in the initial Dundee brainstorming meeting (June 2012). Amendments were made until the steering group were satisfied that the Delphi survey included the main points identified in publications of the eight groups that had used PRECIS and following the Dundee brainstorming feedback meeting. In addition, the survey was clear and easy to complete in less than 10 minutes. The first round had 15 questions and the second round eight questions.

For the pilot we also developed a logo for the PRECIS questionnaire (Figure 5) thus personalising the survey with the wheel that the original developers of the tool had used. We did, however, add “Pragmatic” on the rim to clarify the continuum from pragmatic to explanatory (the 2009 tool only had an “E” in the hub of the wheel spokes for the domains)
All the pilot testers found the surveys easy to complete; it took 5 to 10 minutes for both rounds and they confirmed it covered all salient points. During the pilot for the first round we decided to have open responses for most questions and also not have compulsory questions so participants could skip questions if they wished, to facilitate completion. We also decided to be more open to the viewpoints of the participants so for example, we originally had seven choices for scales for scoring with ranking in the pilot for the first round but we decided to leave the question open and not lead participants but simply ask "If PRECIS was to have a scoring system, what sort of scoring system would you give it?"

One of the pilot testers did state that some of the questions were not really yes/no, more of a continuum of preference/aversion. One pilot tester also helped rephrase a question in Round 2 i.e. Question 6 - "What do you think about trialists determining the weighting for the relative importance of each PRECIS domain, for their trial, on a trial by trial basis". The options are Yes or No which was subsequently amended to "Do you agree that trialists should determine..." which was more appropriate for the yes or no nature of the question.

First round of the Delphi

On 9th October, 2012, we invited by e-mail all those who have cited PRECIS from the Thorpe 2009 [48, 49] publications (Appendix Box 4.1). (We excluded those that had published work that had evaluated the utility of PRECIS or been involved in the development of PRECIS). Email addresses were obtained from the published articles. We asked for a reply within 14 days. We sought advice to improve PRECIS through completion of the SurveyMonkey® questionnaire and assistance with other phases of PRECIS-
2 development. We also asked participants to forward the e-mail invitation to interested colleagues or let us know if they knew of anybody who was interested in participating.

We sent out invitation reminders after two weeks, on 24th October 2012 and another reminder to those who had initially said they were interested but who had not completed the survey after a further two weeks, on 7th November to optimise recruitment (Table 6). Participants were given a further week, until 12th November to participate, so the survey was conducted over a total of five weeks. Sending reminders has been shown to increase recruitment by 4% for each reminder [78]. In addition, to increase response to the electronic questionnaire we used the following strategies [111, 112]:

1. Short questionnaire, less than 10 minutes to complete
2. Non-monetary incentive for survey – acknowledgement of input in publications
   a. Mention of payment for raters who went on to participate in Phase 3 work.
3. Interesting topic – relevant to person invited as they had cited the paper on PRECIS
4. Used a white background in survey
5. Personalised invitations that used name and specifically mentioned paper that had cited the PRECIS publication
6. Heading was PRECIS – Pragmatic Explanatory Continuum Indicator Summaries so did not include survey in e-mail title.
7. Participants were given a clear deadline
8. We obtained the Chief Scientist Office logo to use in e-mail invitations to indicate backing for the research project.
9. The signature of KL, leading the project was included in the e-mail invitation (female, shown to result in higher response rate)
10. We sent out more than one reminder including a link to the survey, to increase the response rate.

We noted the response rate for the first round, pre and post reminder before moving onto the
second round. Following the first round of the Delphi, the steering group reviewed the responses and selected the topics to go forward to the next round.

**Selection criteria for the Delphi**

Topic selection for the second round questionnaire was undertaken by KL and ST based on responses from the first round (e.g. where there was greatest uncertainty) and adding in issues raised by methodological work that had not been covered in the first round. This was subsequently discussed with the steering group. Aside from the first two questions that were introductory questions to gauge the participants’ views on PRECIS, all the issues raised in the first survey were expanded in Round 2 of the modified Delphi. Many of the questions in the first round were open questions to elicit suggestions so these were then presented back to participants in Round 2 for further discussion based on the free text responses [110]. We chose to use the participant’s own words to avoid over-interpretation and ensure the Delphi was participant directed rather than led by KL and the steering group. In addition, issues that had been discussed in brainstorming that had been raised in methodological research of PRECIS, were added to ensure comprehensive discussion of all aspects of PRECIS that we were currently aware of i.e. guidance to use PRECIS and aggregate scoring that had not been raised in the first round. The aim was to generate ideas for further brainstorming.

**Second round of the Delphi**

We sent out the follow-up SurveyMonkey® questionnaire on 26th November to those people who had indicated in Round 1 that they would like to participate in the second round. In this second round, we presented the responses collated in the first round, both quantitative and qualitative, as the basis for follow-up questions. This included the six questions from format of PRECIS-2 to weighting domains (Table 5). The second and final round of the modified Delphi survey was distributed within a short a time as possible after completion of the first round to maintain interest.
### Table 5 Questions for second round of Delphi to improve PRECIS

<table>
<thead>
<tr>
<th>Questions based on results from first round of electronic Delphi survey on PRECIS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1.</strong> 40% of respondents to our first PRECIS survey said that they would prefer to use a table to the PRECIS wheel, do you think we should offer both formats for the PRECIS tool?</td>
</tr>
<tr>
<td><strong>2.</strong> The original tool has no scoring system. 62.5% of respondents think PRECIS needs a score. Anchor points can allow for scoring online and prevent scoring beyond the line. If we have a score, which do you prefer?</td>
</tr>
<tr>
<td>- Split into three: mildly pragmatic, moderately pragmatic, extremely pragmatic</td>
</tr>
<tr>
<td>- Split into three: pragmatic, unclear, explanatory</td>
</tr>
<tr>
<td>- 5 point Likert scale (1=very explanatory; 2=mostly/rather explanatory; 3=equally pragmatic/explanatory, 4=mostly/rather pragmatic, 5=very pragmatic)</td>
</tr>
<tr>
<td>- 1 to 7 Likert scale (with 1 meaning completely explanatory and 7 completely pragmatic)</td>
</tr>
<tr>
<td>- 0 to 9 Likert scale (with 0 meaning completely explanatory and 9 completely pragmatic)</td>
</tr>
<tr>
<td>- 1 to 10 Likert (with 1 meaning completely explanatory and 10 completely pragmatic)</td>
</tr>
<tr>
<td>- 1 to 20 Likert with Visual Analogue properties</td>
</tr>
<tr>
<td><strong>3.</strong> Do you think we need to improve the guidance information given in the original Thorpe et al PRECIS paper to make it easier to score individual domains? Should we have a list of criteria within each wheel element so it is easier to assign a score to each criterion and thus make it easier to determine how pragmatic or explanatory a particular domain is?</td>
</tr>
<tr>
<td><strong>4.</strong> Should we have an aggregate score with cut offs for how pragmatic a trial is, based on the area enclosed by the shape?</td>
</tr>
<tr>
<td><strong>5.</strong> Which of the following domains do you think we should include as extra domains in PRECIS?</td>
</tr>
<tr>
<td>- No extra domains</td>
</tr>
<tr>
<td>- Budget, costs</td>
</tr>
</tbody>
</table>
6. Do you agree that trialists should determine the weighting of the relative importance of each PRECIS domain, on a trial by trial basis? For example, some trialists might consider all domains equally important, others may think eligibility criteria are more important than the other domains.

A further reminder was sent to the Delphi Round 2 participants on 4\textsuperscript{th} December (Table 7) to encourage maximum participation, requesting responses by 11\textsuperscript{th} December. Response rates were noted and a report was prepared for all Round 2 participants and distributed on 7\textsuperscript{th} February 2013.

Table 6 Round 1 Delphi reminder schedules for recruitment by e-mail by stage of invitation

<table>
<thead>
<tr>
<th>Stage of invitation</th>
<th>Dates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial invitation</td>
<td>9\textsuperscript{th} October</td>
</tr>
<tr>
<td>First reminder (2 weeks after initial invitation)</td>
<td>24\textsuperscript{th} October</td>
</tr>
<tr>
<td>Second reminder (2 weeks after first reminder)</td>
<td>7\textsuperscript{th} November</td>
</tr>
</tbody>
</table>
Quantitative and “Free text” data for the Delphi

Quantitative data consisted of the number of participants responding to each question, with a yes, no, unsure response. These quantitative questions were based on published methodological research on PRECIS. We also used free text to efficiently collate opinions and additional information to develop PRECIS-2. Thus, to elicit suggestions for a scoring system an open question was asked i.e. “If PRECIS was to have a scoring system, what sort of scoring system would you give it?” and to encourage suggestions for adding or deleting domains the PRECIS tool participants were asked first a quantitative question “Are there any PRECIS domains you think should be added?” and “Are there any PRECIS domains you think should be removed?” with qualitative follow up questions if Yes, “which domains and why?” All questions about the PRECIS tool had a comments section to encourage free text responses to gather background information on the reasons for answers. All of these comments were considered for the second round of the Delphi or for future brainstorming. Wherever possible we used participants’ wording from “free text” data to inform PRECIS development [96]. This was done to optimise the usability of the PRECIS tool being developed and improve uptake of the tool; it has also been common practice since the original Delphi [105].

Ethics

The study was approved by the University of Dundee Ethics Committee, UREC 12107

---

Table 4.4 Round 2 modified Delphi reminder schedules for recruitment by e-mail by stage of invitation

<table>
<thead>
<tr>
<th>Stage of invitation</th>
<th>Dates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial invitation</td>
<td>26th November</td>
</tr>
<tr>
<td>First reminder(2 weeks after initial invitation)</td>
<td>4th December</td>
</tr>
</tbody>
</table>

---

Table 7 Round 2 modified Delphi reminder schedules for recruitment by e-mail by stage of invitation
Results

Round 1 of the Delphi

Delphi participants

This was an international, diverse group of methodologists, researchers and medical practitioners (Table 8). The only commonality of the participants was their citation of PRECIS [48, 49]. Two of the participants who were the contact person for a publication that cited PRECIS, were erroneously included as they did not meet our a priori selection criteria, having either been involved in the original PRECIS design or had published using PRECIS in another publication. So these two participants were more familiar with the tool.
<table>
<thead>
<tr>
<th>Name</th>
<th>Country</th>
<th>Research Area</th>
<th>Further phases</th>
<th>Other info</th>
<th>Answer to question 1 “Have you used PRECIS tool to design a clinical trial?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Luca Boni</td>
<td>Italy</td>
<td>Epidemiologist, trialist in oncology</td>
<td>Said yes but didn’t invite as had 19 and colleague Gianni had done user testing</td>
<td>Invited to participate by colleague Gianni Virgili</td>
<td>No</td>
</tr>
<tr>
<td>Eric Brass</td>
<td>USA</td>
<td>MD, Trialist, Director Centre for Clinical Pharmacology</td>
<td>1st and 2nd round, user testing, Theta Brainstorming, validity testing</td>
<td>12 if consider large clinical trials</td>
<td>No</td>
</tr>
<tr>
<td>Vilemine Carayanni</td>
<td>Greece</td>
<td>Design trials considering economic evaluation</td>
<td>User testing</td>
<td>2 – economic evaluation</td>
<td>No</td>
</tr>
<tr>
<td>June Carroll</td>
<td>Canada</td>
<td>Doctor, family physician</td>
<td>1st round Delphi</td>
<td></td>
<td>No</td>
</tr>
<tr>
<td>Kalipso Chalkidou</td>
<td>UK</td>
<td>Medical doctor</td>
<td>1st and 2nd round Delphi</td>
<td>Original PRECIS developer (first author CER article)</td>
<td>No</td>
</tr>
<tr>
<td>Jean Paul Chippaux</td>
<td>France</td>
<td>International health, director of research</td>
<td>1st round Delphi</td>
<td></td>
<td>Anonymous</td>
</tr>
<tr>
<td>Tim Coats</td>
<td>UK</td>
<td>Emergency medicine doctor, trialist</td>
<td>1st round Delphi</td>
<td>Don’t know - anonymous</td>
<td></td>
</tr>
<tr>
<td>Sandra Eldridge</td>
<td>UK</td>
<td>Senior Statistician, Trialist, Director of Pragmatic Clinical Trials Unit</td>
<td>1st round, possibly 2nd round anonymous</td>
<td></td>
<td>No</td>
</tr>
<tr>
<td>Louis D Fiore</td>
<td>USA</td>
<td>MD, internal medicine, epidemiologist, trialist</td>
<td>1st and 2nd round, User testing</td>
<td>About 50 trials</td>
<td>No</td>
</tr>
<tr>
<td>Name</td>
<td>Country</td>
<td>Research Area</td>
<td>Further phases</td>
<td>Other info</td>
<td>Answer to question 1 “Have you used PRECIS tool to design a clinical trial?</td>
</tr>
<tr>
<td>---------------------</td>
<td>------------</td>
<td>--------------------------------------------------------------------------------</td>
<td>-------------------------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>--------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Christian Gold</td>
<td>Norway</td>
<td>Music Therapist, trialist</td>
<td>User testing</td>
<td>About 8 trials some quasi RCT</td>
<td>Yes - I have used it as an idea, not very systematically. But it has been helpful both in guiding decisions and in justifying them.</td>
</tr>
<tr>
<td>Amanda Graham</td>
<td>USA</td>
<td>PhD, Psychologist, trialist</td>
<td>1\textsuperscript{st} and 2\textsuperscript{nd} round Delphi</td>
<td>Not in the design of a clinical trial, but in publications and presentations about the trial to justify/clarify certain design decisions.</td>
<td>No</td>
</tr>
<tr>
<td>David L Hahn</td>
<td>USA</td>
<td>MD, trialist</td>
<td>2\textsuperscript{nd} round Delphi</td>
<td></td>
<td>No</td>
</tr>
<tr>
<td>Peter Hall</td>
<td>UK</td>
<td>Oncologist, design and running of oncology clinical trials, economic analysis</td>
<td>User testing</td>
<td>4 trials total: 2 trials health economics, 1 Phase 2, 1 Phase 3</td>
<td>No</td>
</tr>
<tr>
<td>Gisele Huf</td>
<td>Brazil</td>
<td>MD, trialist</td>
<td>1\textsuperscript{st} and 2\textsuperscript{nd} round, user testing</td>
<td>5 trials</td>
<td>No</td>
</tr>
<tr>
<td>Eric S Johnson</td>
<td>USA</td>
<td>Researcher, PhD</td>
<td>Toronto brainstorming (web)</td>
<td></td>
<td>No</td>
</tr>
<tr>
<td>Rodger Kessler</td>
<td>USA</td>
<td>MD, psychologist</td>
<td>1\textsuperscript{st} round, user testing</td>
<td>10-15 trials</td>
<td>Anonymous</td>
</tr>
<tr>
<td>Maarten O Kok</td>
<td>The Netherlands</td>
<td>Physiotherapist, psychology, health policy</td>
<td>1\textsuperscript{st} round Delphi</td>
<td></td>
<td>Anonymous</td>
</tr>
<tr>
<td>Jerry A Krishnan</td>
<td>USA</td>
<td>MD, PhD</td>
<td>1\textsuperscript{st} and 2\textsuperscript{nd} round Delphi, Toronto brainstorming, Validity testing</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Name</td>
<td>Country</td>
<td>Research Area</td>
<td>Further phases</td>
<td>Other info</td>
<td>Answer to question</td>
</tr>
<tr>
<td>----------------------</td>
<td>-------------</td>
<td>--------------------------------------</td>
<td>---------------------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>--------------------</td>
</tr>
<tr>
<td>Ingrid Liodden</td>
<td>Norway</td>
<td>Nurse practitioner</td>
<td>2&lt;sup&gt;nd&lt;/sup&gt; round Delphi</td>
<td></td>
<td>No</td>
</tr>
<tr>
<td>Noemi Lois</td>
<td>UK</td>
<td>ophthalmologist</td>
<td>1&lt;sup&gt;st&lt;/sup&gt; and 2&lt;sup&gt;nd&lt;/sup&gt; round Delphi</td>
<td></td>
<td>No</td>
</tr>
<tr>
<td>Franz Portzsolt</td>
<td>Germany</td>
<td>MD, Haematologist/oncologist trialist</td>
<td>2&lt;sup&gt;nd&lt;/sup&gt; round, User testing</td>
<td>40 years in trials: 20-50</td>
<td>No</td>
</tr>
<tr>
<td>John Powers</td>
<td>USA</td>
<td>MD, trialist</td>
<td>1&lt;sup&gt;st&lt;/sup&gt; round Delphi, Toronto brainstorming (web)</td>
<td>An excellent tool but journal editors don't seem to understand the difference between pragmatic and explanatory trials and reject pragmatic trials with &quot;negative&quot; results.</td>
<td>Yes</td>
</tr>
<tr>
<td>Marianna Purgato</td>
<td>Italy</td>
<td>Systematic reviewer, psychiatrist</td>
<td>1&lt;sup&gt;st&lt;/sup&gt; round Delphi</td>
<td></td>
<td>No</td>
</tr>
<tr>
<td>Jean Raymond</td>
<td>Canada</td>
<td>MD, oncologist, trialist</td>
<td>1&lt;sup&gt;st&lt;/sup&gt; round Delphi</td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td>Clare Relton</td>
<td>UK</td>
<td>Medical doctor, trialist</td>
<td>1&lt;sup&gt;st&lt;/sup&gt; and 2&lt;sup&gt;nd&lt;/sup&gt; round, user testing</td>
<td>4/5 homeopathy trials</td>
<td>No</td>
</tr>
<tr>
<td>Kim Reynolds</td>
<td>USA</td>
<td>Trialist</td>
<td>1&lt;sup&gt;st&lt;/sup&gt; round, 2&lt;sup&gt;nd&lt;/sup&gt; round Delphi</td>
<td></td>
<td>No</td>
</tr>
<tr>
<td>Jodi Segal</td>
<td>USA</td>
<td>MD, Trialist, lecturer</td>
<td>2&lt;sup&gt;nd&lt;/sup&gt; round, User testing, validity testing</td>
<td>No trials, 9 observational studies, systematic reviews</td>
<td>No</td>
</tr>
<tr>
<td>Peter Selby</td>
<td>Canada</td>
<td>MD, addictions, trialist</td>
<td>Toronto brainstorming (web)</td>
<td>Publication using PRECIS it was used to inform the design rather than explicitly go through the process</td>
<td>Yes</td>
</tr>
<tr>
<td>Name</td>
<td>Country</td>
<td>Research Area</td>
<td>Further phases</td>
<td>Other info</td>
<td>Answer to question 1 “Have you used PRECIS tool to design a clinical trial?</td>
</tr>
<tr>
<td>---------------------------</td>
<td>------------------</td>
<td>------------------------------------------------------------------------------</td>
<td>---------------------------------------------</td>
<td>------------------------------</td>
<td>--------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Grace Thompson</td>
<td>Australia</td>
<td>Early career researcher, Lecturer, Music Therapy</td>
<td>1\textsuperscript{st}, 2\textsuperscript{nd} round, user testing</td>
<td>2 trials</td>
<td>Anonymous first round</td>
</tr>
<tr>
<td>Sabine N Van der Veer</td>
<td>The Netherlands</td>
<td>Medical informatics, trialist</td>
<td>Validity testing</td>
<td></td>
<td>No</td>
</tr>
<tr>
<td>Tjeerd van Staa</td>
<td>UK</td>
<td>Health informatics</td>
<td>1\textsuperscript{st} round Delphi</td>
<td></td>
<td>Don’t know - anonymous</td>
</tr>
<tr>
<td>Gianni Virgili</td>
<td>Italy</td>
<td>Ophthalmologist, systematic reviewer, hospital care management</td>
<td>User testing</td>
<td>Systematic reviewer</td>
<td>No</td>
</tr>
<tr>
<td>Vivian Andrea Welch</td>
<td>Canada</td>
<td>Clinical Epidemiology, Methodologist, Cochrane Systematic reviewer</td>
<td>2\textsuperscript{nd} round, brainstorming, validity testing</td>
<td></td>
<td>No</td>
</tr>
<tr>
<td>Hywel C Williams</td>
<td>UK</td>
<td>Dermatologist, trialist</td>
<td>2\textsuperscript{nd} round</td>
<td>Publication using PRECIS (Bratton)</td>
<td>No</td>
</tr>
</tbody>
</table>
Figure 6 Flow diagram of participants in 3 phases
There were initially 90 e-mails sent out to an international group of researchers but two respondents forwarded to colleagues so an additional three people were invited to participate in the Delphi survey: Italian (2), and Australian (1); all with similar background to the others invited. The Australian was recommended by a colleague as they had been using PRECIS in their postgraduate research. One Italian and one Australian scientist accepted the invitation to participate in the Delphi. Five potential participants took the time to respond with reasons for not participating, including, too busy (2), retiring (1), unable to help (1), not interested (1) and three potential participants were away with automatic responses stating irregular access to the internet for the time of the survey. Nine e-mail addresses were “dead”. After checking they were correct in publication citing PRECIS no further contact was attempted with these researchers, due to uncertainty about communicating with correct person, and time constraints. The participation rate was therefore 37% (34/93) with a response rate of 47% (42/90).

Table 9 Round 1 modified Delphi number recruited by e-mail by stage of invitation

<table>
<thead>
<tr>
<th>Stage of invitation</th>
<th>n</th>
<th>Recruitment rate % (n = 93)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial invitation</td>
<td>27</td>
<td>29.0</td>
</tr>
<tr>
<td>First reminder (2 weeks after initial invitation)</td>
<td>4</td>
<td>4.3</td>
</tr>
<tr>
<td>Second reminder (2 weeks after first reminder)</td>
<td>3</td>
<td>3.2</td>
</tr>
<tr>
<td>Total</td>
<td>34</td>
<td>36.5</td>
</tr>
</tbody>
</table>

Prior to the first reminder 27 people had responded (started the survey), with four more after the first reminder and an additional three participants after the second reminder, one received two days after the deadline which was also included (Table 9). A total number of 34 individuals started the survey and 28 (82.4%) finished the survey. The results can be seen in Table 10. A total number of 31 were interested in participating in round 2, after responding positively to the question in the Delphi “Are you interested in a further round of Survey Monkey?”. Questions skipped varied from 1 to 22 per respondent.
### Round 1 Results of the Delphi

#### Quantitative results

Table 10 Round 1 Questions and results for 34 participants

<table>
<thead>
<tr>
<th>No.</th>
<th>Question</th>
<th>YES</th>
<th>NO</th>
<th>UNSURE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Have you used PRECIS (PRagmatic Explanatory Continuum Indicator Summaries) Tool to design a clinical trial?</td>
<td>6</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Do you like the current PRECIS tool, which is a wheel with separate domains covering different aspects of trial design?</td>
<td>28</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Would you prefer a table instead of the PRECIS wheel to &quot;score&quot; how pragmatic the different domains are?</td>
<td>14</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>The original tool has no scoring system - do you think PRECIS needs a score?</td>
<td>20</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>If you don't think there should be a score, are you happy using a Visual Analogue Scale and putting a mark on a spoke to indicate how pragmatic a domain is?</td>
<td>11</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Are there any PRECIS domains you think should be added?</td>
<td>10</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Are there any PRECIS domains you think should be removed?</td>
<td>3</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Should we have weighting of PRECIS domains, so some domains count more in the overall PRECIS &quot;score&quot; or in determining how pragmatic a trial is?</td>
<td>YES</td>
<td>NO</td>
<td>UNSURE</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>10</td>
<td>Is there anything you would like to add about using PRECIS that has not been covered in previous questions?</td>
<td>5</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Are you interested in assisting in a further round of Survey Monkey?</td>
<td>31</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Are you interested in assisting in Phase 2 - user testing of different PRECIS models, one to one or small group testing? This will take up to 1 hour</td>
<td>22</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Are you interested in assisting in Phase 3 - testing PRECIS-2 with a sample of 15 trials over a 2 month period?</td>
<td>22</td>
<td>9</td>
<td></td>
</tr>
</tbody>
</table>
“Free text” Results

While all questions gave participants the opportunity to comment on the question, several questions in particular sought suggestions from participants and were open questions, 6 and 7. However, the comments field linked to all questions produced comments on the domain descriptions (included below).

**Question 6. If PRECIS was to have a scoring system, what sort of scoring system would you give it?**

This question received 21 comments, from “too difficult to answer” to “the simplest possible.”

The scoring system suggestions given by participants have been categorised into two groups: pragmatism (so thinking purely of grading trial domains on how pragmatic they are in comparison with usual care; and options for scoring domains from “explanatory to pragmatic”. This group of “free text” suggestions from “explanatory to pragmatic” were then categorised into six options that could then be discussed in the second round of the Delphi Survey (Table 11).
Table 11 Scoring System suggestions for PRECIS-2

<table>
<thead>
<tr>
<th>Scoring System suggestion</th>
<th>Number of responses</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pragmatism</strong></td>
<td></td>
</tr>
<tr>
<td>Split into three: mildly pragmatic, moderately pragmatic, extremely pragmatic – degree from typical care.</td>
<td>1</td>
</tr>
<tr>
<td><strong>Explanatory to Pragmatic</strong></td>
<td></td>
</tr>
<tr>
<td>Split into three: pragmatic, unclear, explanatory</td>
<td>5</td>
</tr>
<tr>
<td>5-point Likert scale (1=very explanatory; 2=mostly/rather explanatory; 3=equally pragmatic/explanatory, 4=mostly/rather pragmatic, 5=very pragmatic)</td>
<td>2</td>
</tr>
<tr>
<td>1 to 7 Likert scale</td>
<td>1</td>
</tr>
<tr>
<td>0 to 9 Likert scale</td>
<td>1</td>
</tr>
<tr>
<td>1 to 10 (with 1 meaning completely explanatory and 10 completely pragmatic)</td>
<td>5</td>
</tr>
<tr>
<td>Likert with Visual analogue properties ie 20 point with anchor points to allow for scoring online and prevent scoring beyond the line.</td>
<td>1</td>
</tr>
</tbody>
</table>

**Question 7. Are there any PRECIS domains you think should be added?**

This question produced 16 comments on the existing PRECIS tool with suggestions for domains and discussion of the current domains’ weaknesses or inadequacy or misinterpretation of tool and domain descriptions. A few of the comments (out of 16) simply said the participants hadn’t thought about it, or had discussed in previous questions for instance “2. Do you like the current PRECIS tool, which is a
wheel with separate domains covering different aspects of trial design?”. Two of the responses (out of 16) mentioned that the participants would have liked to see another picture of the PRECIS wheel with the domains to remind them what they were “as I don’t keep them all in my head”.

Overall, the 10 suggestions for new additional domains (in no particular order) using the participant’s own words from the Delphi survey were diverse: cost analysis, consent, recruitment, scarce resources, intervention adverse events, consideration of the research question and the trial’s primary objective, blinding and use of the trial results (Table 12).

Table 12 Ten additional domain suggestions for PRECIS-2

<table>
<thead>
<tr>
<th>Additional Domain suggestion</th>
<th>Number of responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Budget, costs</td>
<td>1</td>
</tr>
<tr>
<td>Availability of drug</td>
<td>1</td>
</tr>
<tr>
<td>Risks</td>
<td>1</td>
</tr>
<tr>
<td>Feasibility of masking outcome measure</td>
<td>1</td>
</tr>
<tr>
<td>Appropriateness of methodology for question asked</td>
<td>1</td>
</tr>
<tr>
<td>Main objective of the trial e.g. inform practice or register a new drug</td>
<td>1</td>
</tr>
<tr>
<td>Strength of the rationale (1)</td>
<td>1</td>
</tr>
<tr>
<td>Integration into the healthcare system (1)</td>
<td>1</td>
</tr>
<tr>
<td>Source of the patients or the setting for recruitment</td>
<td>3</td>
</tr>
<tr>
<td>Information/consent domain</td>
<td>1</td>
</tr>
</tbody>
</table>
Suggestions for modifications of descriptions for domain – free text from participants

**Eligibility** There were concerns that the eligibility criteria do not specify how representative the population involved in the trial is. “The most extremely pragmatic approach to eligibility would seek only to identify study participants with the condition of interest from as many sources (e.g., institutions) as possible”. One clinician suggested “a purely pragmatic trial should not include “all participants with the condition of interest” e.g. only enrol patients with severe hypertension if research question aimed at this group so do not include everyone with hypertension. And do not include all people with asthma if only looking at asthma of a certain severity.”

**Eligibility** “actual measure of comparability of trial population with target population (ie did the trial actually recruit in an unselected manner from the target population)”

**Practitioner expertise** “Practitioner expertise is stated but it is unclear whether this reflects the typical expertise of practitioners who would deliver the program in real world circumstances.”

**Compliance** “Does participant and practitioner compliance fully capture an assessment of the fidelity of the intervention delivery?”

**Flexibility of experimental intervention** “Does the tool currently capture an intentionally tightly controlled/defined intervention so does not have to score low on the pragmatism scale.”

**Analysis** – “Cluster randomized consider setting in addition to patient and provider characteristics e.g. setting with an electronic health record, setting not typical of a community setting, at least not in USA. A systematic reviewer may consider findings closer to the explanatory or ideal.”
Question 8. Are there any PRECIS domains you think should be removed?

The suggestions for removal included three domains: *Analysis* (1 response) and *Flexibility of interventions/comparison* and *Practitioner Expertise intervention/comparison* (both from the same person). While only four researchers said yes to removing domains the six comments included two that said they had answered in previous question and included response “too difficult question”.

**Analysis** - One person stated that this domain could be removed as Intention to Treat approach (ITT), in which all participants must be accounted for and analysed in the group to which they were allocated, whether or not they finish the trial or crossover, should be the same for explanatory and pragmatic trials.

*Flexibility of interventions/comparison* and *Practitioner Expertise intervention/comparison*. Two participants believed that these were often the same for the intervention and comparison.

Question 9. Should we have weighting of PRECIS domains, so some domains count more in the overall PRECIS "score" or in determining how pragmatic a trial is?

This question produced 15 comments (Table 13) – opinions were mixed: from keep it simple to others specifying which domains had most weight in their opinion e.g. eligibility criteria and primary outcome.
Table 13 Qualitative feedback from Delphi participants on question 9 on PRECIS (15 comments in no particular order)

<table>
<thead>
<tr>
<th>Question 9</th>
<th>Should we have weighting of PRECIS domains, so some domains count more in the overall PRECIS &quot;score&quot; or in determining how pragmatic a trial is?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Not sure of purpose of quantification. In some respects an ITT analysis is &quot;pragmatic&quot; compared to an &quot;as treated&quot; analysis even in an explanatory trial. Keep it simple. Readers don't understand much of these principles as it is.</td>
</tr>
<tr>
<td>2.</td>
<td>My understanding is that this alerts the trial designer as to whether a choice they are making is less or more pragmatic. I don't think anyone needs to say &quot;YES&quot; - this is pragmatic.</td>
</tr>
<tr>
<td>3.</td>
<td>Main things for me are the eligibility criteria and the stringency of the protocol.</td>
</tr>
<tr>
<td>4.</td>
<td>No - learn from all of the flawed attempts from trying to quantify scales that measure study quality. It is the components and overall pattern that is important.</td>
</tr>
<tr>
<td>5.</td>
<td>Because, in my opinion, some domains have a different weight. for example, practitioner expertise vs. outcomes or primary analysis</td>
</tr>
<tr>
<td>6.</td>
<td>I think a more descriptive accounting might be better than a weighted number that often gives a false sense of objectivity.</td>
</tr>
<tr>
<td>7.</td>
<td>e.g the primary outcome and the eligibility criteria are much more crucial than other domains</td>
</tr>
<tr>
<td>8.</td>
<td>I think it may depend on the intervention. Some PRECIS components may be more important for some interventions</td>
</tr>
<tr>
<td>9.</td>
<td>For a given trial, there is often a certain weighting. But I am not sure if there is one that can be generalised.</td>
</tr>
<tr>
<td>10.</td>
<td>An overall score won't be as useful as separate scores for each of the 10 domains.</td>
</tr>
<tr>
<td>11.</td>
<td>Depends on application, but suspect weighting would imply quantitative meaning not justified -- goal should be to inform trialists and readers</td>
</tr>
<tr>
<td>12.</td>
<td>I think that it is hard to weight the domains and I do not think that &quot;pragmaticness&quot; is a quality which can be specified - I prefer the idea of the cogwheel diagram of the individual domains without a weighting.</td>
</tr>
<tr>
<td>13.</td>
<td>maybe the &quot;compliance&quot; items</td>
</tr>
<tr>
<td>14.</td>
<td>I would need to think about this a bit more.</td>
</tr>
<tr>
<td>15.</td>
<td>Some items have clinical and 'ethical' impacts. Others do not, I am not sure you need numerical weights though</td>
</tr>
<tr>
<td>Question 10</td>
<td>Is there anything you would like to add about using PRECIS that has not been covered in previous questions?</td>
</tr>
<tr>
<td>-------------</td>
<td>-----------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>1.</td>
<td>I don’t dare to suggest that PRECIS will be needed only for 5 more years. After 5 years most people will understand ...</td>
</tr>
<tr>
<td>2.</td>
<td>I don’t think of PRECIS only as a tool for designing a trial. It can also be used when a trial is running, or even when folk are writing up their trial when considering whether to use terms such as “pragmatic”. I am also aware of the framework by Dekkers on external validity which is close to this area</td>
</tr>
<tr>
<td>3.</td>
<td>The PRECIS authors are a talented group of individuals from a variety of academic and quasi-academic (Kaiser) settings. The next generation of PRECIS would benefit from inclusion of non-academic practice-based researchers, including non-MDs (e.g., nurses, pharmacists, dentists).</td>
</tr>
<tr>
<td>4.</td>
<td>I would like to see additional version for designing a range of pragmatic studies beyond RCTs (e.g., comparative cohort studies, prognostic cohorts that predict risk, etc.).</td>
</tr>
<tr>
<td>5.</td>
<td>I think that PRECIS is a great tool and am keen to see it be more widely used</td>
</tr>
<tr>
<td>6.</td>
<td>Maybe PRECIS should be better used not as a resource to help the design of a trial, but rather as a diagnostic tool to draw conclusions about its “pragmatism”</td>
</tr>
<tr>
<td>7.</td>
<td>Wondering if ‘reversing’ the direction of the intensity (pragmatism) would not have been better. I was looking for a way to see at a glance what was ‘extra’ Extra tests Extra selection Extra expertise Extra visits Extra questionnaire beyond what was normal care</td>
</tr>
<tr>
<td>8.</td>
<td>I am not sure the tool currently captures that an intentionally tightly controlled/defined intervention (when the tight control is part of the intervention, in a sense) does not have to score low on the pragmatism scale...Not sure this means a new domain is needed but I think this is problematic.</td>
</tr>
</tbody>
</table>
Question 10. Is there anything you would like to add about using PRECIS that has not been covered in previous questions?

This question elicited eight comments (Table 14) indicating that the PRECIS tool was viewed positively but also suggesting there was interest in using the tool to assess: trials in systematic reviews; to design cohort studies and providing ideas to improve the next version of the tool. The comment from one participant suggesting the input of health professionals that were practice based was noted with interest as the steering group and brainstorming group in Dundee comprised dentists, practising General Practitioners and was led by a nurse (KL) with a current license to practice.

Round 2 Results of the Delphi

Participants

Thirty two people had asked to be contacted to participate in Round 2 of the Delphi, so the participation rate in this stage was 72% (23/32). A total number of 23 started the survey (Table 15) with 20 (87%) completing all the questions. The results can be seen in Table 16.

Table 15 Round 2 modified Delphi number recruited by e-mail by stage of invitation

<table>
<thead>
<tr>
<th>Stage of invitation</th>
<th>n</th>
<th>Recruitment rate % (n = 32)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial invitation</td>
<td>19</td>
<td>59.4</td>
</tr>
<tr>
<td>First and only reminder (2 weeks after initial invitation)</td>
<td>4</td>
<td>12.5</td>
</tr>
<tr>
<td>Total</td>
<td>23</td>
<td>71.9</td>
</tr>
</tbody>
</table>
Quantitative results for Round 2 of the Delphi

Table 16 Round 2 Questions and results of the Delphi

<table>
<thead>
<tr>
<th>No.</th>
<th>Question</th>
<th>YES (%)</th>
<th>NO (%)</th>
<th>Unsure (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>40% of respondents to our first PRECIS survey said that they would prefer to use a table to the PRECIS wheel, do you think we should offer both formats for the PRECIS tool?</td>
<td>13 (56.5)</td>
<td>7 (30.4)</td>
<td>3 (13.0)</td>
</tr>
<tr>
<td>2</td>
<td>The original tool has no scoring system - 62.5% of respondents think PRECIS needs a score. Anchor points can allow for scoring online and prevent scoring beyond the line. If we have a score, which do you prefer?</td>
<td>5 point Likert scale (9 – 45%)</td>
<td>7 point Likert scale (4 – 20%)</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Do you think we need to improve the guidance information given in the original Thorpe et al PRECIS paper to make it easier to score individual domains?</td>
<td>18 (78.3)</td>
<td>1 (4.4)</td>
<td>4 (17.4)</td>
</tr>
<tr>
<td>4</td>
<td>Should we have an aggregate score with cut offs for how pragmatic a trial is, based on the area enclosed by the shape?</td>
<td>6 (26.1)</td>
<td>10 (43.5)</td>
<td>7 (30.4)</td>
</tr>
<tr>
<td>5</td>
<td>Which of the following domains do you think we should include as extra domains in PRECIS?</td>
<td>Source of the patient or the setting for recruitment 12 (52.2%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Do you agree that trialists should determine the weighting of the relative importance of each PRECIS domain, on a trial by trial basis? For example, some trialists might consider all domains equally important, others may think eligibility criteria are more important than the other domains.</td>
<td>8 (34.8)</td>
<td>7 (30.4)</td>
<td>8 (34.8)</td>
</tr>
</tbody>
</table>
**Question 2 detail**

**Question 2.** The original tool has no scoring system - 62.5% of respondents think PRECIS needs a score. Anchor points can allow for scoring online and prevent scoring beyond the line. If we have a score, which do you prefer?

There were 20 responses to this question with one participant skipping the question (Figure 7).

![Question 2 results for second round of Delphi](image)

**Figure 7 Question 2 results for second round of Delphi**

With a clear majority, 45%, favoured a 5-point Likert scale, participants in the second round of the Delphi have indicated this would be the choice for the majority of potential users of PRECIS.

**Question 5 detail**

**Question 5.** Which of the following domains do you think we should include as extra domains in PRECIS?

With 23 responses on this domain, the most popular choice for an additional domain was - “Source of the patients or the setting for recruitment” - 52.2% of respondents thought we should add a domain considering context “Source of the patients or the setting for recruitment” (Figure 8).
Figure 8 Question 5 results for second round of Delphi

There were three other domains that were equal second for consideration for inclusion with 26.1% (respondents could choose more than one domain): “No extra domains”, “budget/costs”, “Main objective of the trial e.g. inform practice or register a new drug”.

Qualitative Results Round 2 of the Delphi

Question 1. 40% of respondents to our first PRECIS survey said that they would prefer to use a table to the PRECIS wheel, “Do you think we should offer both formats for the PRECIS tool?”

There were 18 comments on this question, the question that triggered most “free text” responses. Respondents felt quite strongly about the benefits of both a table and a wheel, however there were a
few comments that using both formats could be confusing.

One comment to consider in informing the development of PRECIS-2 was discussion of the design of the original 2009 PRECIS model: “the spokes are visually confusing at 180” - making it look like e.g. Practitioner expertise was diametrically opposed to participant compliance or that somehow these are two opposite ends of the same factor.” If the PRECIS wheel remains, we should consider modifying the placement of the spokes to remove this potential confusion.

**Question 2.** The original tool has no scoring system - 62.5% of respondents think PRECIS needs a score. Anchor points can allow for scoring online and prevent scoring beyond the line. If we have a score, which do you prefer?

There were two comments that the scoring system was pushing the use of the tool to more or less pragmatic and that without further guidance on what explanatory and pragmatic meant it was not possible to choose a scoring system.

**Question 3.** Do you think we need to improve the guidance information given in the original Thorpe et al PRECIS paper to make it easier to score individual domains? Should we have a list of criteria within each wheel element so it is easier to assign a score to each criterion and thus make it easier to determine how pragmatic or explanatory a particular domain is?

There were nine comments on this question with 78.3% respondents saying we need to improve the original guidance; some thought that examples were most helpful perhaps better than “an exhaustive list of criteria”, but only one person mentioned the four trial examples in the original PRECIS paper by Thorpe et al (2009).

The purpose of PRECIS was also highlighted in how much guidance should be given: “If there is an expectation of external comparison and uniformity then considerable guidance required. If the purpose is to challenge the design and facilitate critique/interpretation/discussion then less is needed”. But another respondent highlighted the “problem not wise to overload the reader with too much information...”. 
Another point made by a respondent, that is relevant to designing clinical trials, is the education and background of the person using PRECIS to design trials and their varying need for guidance.

One comment that could also be addressed in additional guidance for PRECIS-2, rather than adding a new domain: “Many trialists would prefer to design more pragmatic trials, but it is cost or time prohibitive given their funding. How do trialists trade-off between domains when faced with time or budget constraints? What is the highest priority for investing in pragmatic evidence?”

Question 4. **Should we have an aggregate score with cut offs for how pragmatic a trial is, based on the area enclosed by the shape?**

There were 15 comments on this question, the question that triggered the second most “free text” responses. The respondents believed that the individual domain scores were important and that the key problem with aggregate scoring is that it masks extreme scoring in domains which is visually very clear in the wheel.

**Question 5. Which of the following domains do you think we should include as extra domains in PRECIS?**

There were seven comments for this question and nobody skipped answering the question.

Budget and costs had been voted by six people for inclusion in PRECIS. One person commented that costs could also be included in Participant compliance, perhaps suggesting incentives. But another comment was specific to “drug costs”.

An additional inclusion for PRECIS-2 that was suggested is an Information/consent domain. Comment: “how closely (or not) does the information that patients (and clinicians) receive match what is communicated in routine healthcare? In routine healthcare patients are not told that their treatment will be randomly selected (rather than chosen by themselves and/or their clinicians; nor are they told about that treatments that they may not then be randomly selected to). If a study is to be truly pragmatic then surely patients and clinicians should receive it in routine healthcare – as happened in Zelen/randomised consent designs – in a staged and patient relevant manner. That is to say –
patients/participants are not told about treatments that they might not receive nor are patients/participants told that their treatment WILL be decided by chance. Instead – only those patients/participants randomly selected to the intervention group are told about the intervention, and only those randomly selected to the intervention are told that they HAVE BEEN randomly selected (rather than that they WILL BE randomly selected). The information that people in trials receives determines how they behave and how they experience the intervention, and is important to know when attempting to interpret the results of a trial and understand the generalizability (or not) of the trial results.”

Question 6. Do you agree that trialists should determine the relative importance of the weighting of each PRECIS domain on a trial by trial basis?

Opinion was divided on weighting PRECIS domains. There were nine free text responses indicating that the “no” group felt strongly with comments “no – recipe for a mess”, “this would destroy any standardisation” and “how would this work in terms of comparing across trials or when the tool is filled in for the same trial by different individuals?”. 
Overall Results

A summary of the results from Round 1 and 2 of the modified Delphi on PRECIS-2 Scoring, suggestions for Domain changes, additional guidance, tool format and areas needing further discussion from can be seen in Table 17.

Table 17 Combined results from Round 1 and 2 of modified Delphi

<table>
<thead>
<tr>
<th>Scoring</th>
<th>The original tool has no scoring system – 19 (63.3%: N=30) of respondents in the first round of the Delphi thought that PRECIS needed a score. Anchor points could allow for scoring online and prevent scoring beyond the line. 9 (45%: N=20) of respondents in the second round chose 5 point Likert scale.</th>
</tr>
</thead>
</table>
| Domains | • Add Source of the patients OR Add Setting for recruitment OR include in Eligibility Criteria 12 responses (52.2%)  
• Add Cost analysis 6 (26.1%)  
• Remove Analysis  
• Consider main objective of the trial eg register a drug or inform practice 6 (26.1%)  
• Integration into the healthcare system 5 (21.7%) |
| Guidance | 78.3% said that we need to improve the guidance information given in the original PRECIS paper (N=23) |
| Format | • The original PRECIS tool was a wheel: 13 (56.5%: N = 23) of respondents said that we should offer a table (with rationale for domain scores and a wheel format for the PRECIS tool  
• The PRECIS wheel rim should have text PRAGMATIC “donut “labelling with E in the centre  
• The wheel domains should be logically arranged and clockwise instead of anti-clockwise  
• Improve labelling of domains |
| For Discussion in brainstorming | • Blinding  
• Trade-off between domains when time or budget constraints  
• Highest priority for investing in pragmatic evidence?  
• Weighting |

Discussion

Main findings

Probably because only authors who had cited PRECIS were invited, there was a good response rate and a great deal of interest and enthusiasm in the project was generated over a short period of time (just over 2 months). SurveyMonkey® proved a straightforward tool to use, to collate individual responses and generate reports. Participants responded both by offering suggestions to improve
PRECIS and to help in future phases of the project: user testing and validation. It is significant that almost all of these respondents had not undertaken methodological work with PRECIS so they had not used PRECIS in designing clinical trials, however the participants were positive about the tool; this is backed up by increasing citations of the two PRECIS articles published simultaneously in 2009.

In the second round of the Delphi, the participation rate was 72% (23/32) compared to 37% in the first round indicating the interest from this particular group. The first round of participants may have included some individuals who were curious to take a look at the survey content rather than an altruistic approach to contributing to the research as in the second round; 20 finished the survey out of the 23 that started (87%) compared to 27 out of 34 (79.4%) in the first round.

Delphi respondents were clearly in favour of a scoring system for PRECIS which is supported by almost all of the groups (9 out of 11) that had undertaken methodological work with PRECIS – eight at the time of the modified Delphi ([39, 40, 53, 55, 57, 58, 63, 73]). Only Bratton (2011 and 2012) in two studies had used the Visual Analogue Scale in the original PRECIS wheel [57, 58]. Participants in the modified Delphi were mostly interested in using PRECIS with a score and using a Likert scale of 1-3, 1-5 and 1-7 scoring. This would assist trialists to carefully consider their decisions for each domain in PRECIS AND make it easier to compare their “scores” with their fellow trialists. This should thus simplify the process enabling consensus decision making in designing clinical trials in a way that was not possible before, especially when trialists are not sitting round a table together but separately involved in e-mail discussion [60].

The reality of splitting the PRECIS score from 1 to 7 from 1 most explanatory to 7 most pragmatic would be difficult to put into practice. It is easy to define the anchors: 1 as very explanatory, 4 explanatory = pragmatic and 7 as very pragmatic but harder to define the areas in between for 2, 3 and 5, 6. It is far easier to imagine using Likert scales of 1 to 3 (1 explanatory, 2 explanatory = pragmatic, and 3
pragmatic) or indeed scales of 1 to 5 (1 as very explanatory, 2 as rather explanatory, 3 explanatory = pragmatic, 4 rather pragmatic and 5 very pragmatic) but this will be discussed in a subsequent chapter 4, Brainstorming meetings and in Chapter 5 User testing.

Interestingly, half of the participants in the first round of the Delphi did not respond positively to the qualitative aspect of Question 6 “If PRECIS was to have a scoring system, what sort of scoring system would you give it?” Responses included participants stating that we should have given options for scoring, not wishing to state their preference saying “ordinal scale” while others stated “too difficult question” and “this question is quite vague”. Prior to the pilot, in the original version of the questionnaire for the first round of the Delphi, we had had six suggestions for scoring but had decided to leave it open to encourage genuine feedback. However, even in the pilot, testers had stated a preference for multiple choice options for scoring.

The issue of guidance arose in comments that were not always specific to the question asking about guidance. It appeared that not all participants completely understood the domain descriptions in the 2009 paper on PRECIS. The feedback strongly suggested that we should focus on providing clear information on how to use PRECIS-2 to assist trialists. Getting the balance right however could be important to facilitate using the tool and not make it too complex. Involving people with a range of backgrounds would assist in trial design but their design decision making needs using PRECIS may differ, depending on whether or not they treat patients and their experience as trialists. Thus, in developing PRECIS-2 and considering guidance we not only need to be aware of who is going to use PRECIS-2 but also consider the minimum amount of guidance that will enable everyone to use PRECIS. We should also test PRECIS-2 using a diverse group of trialists.

Two other aspects that should be covered in advice to use PRECIS include masking/blinding of outcomes and explicit views on pragmatism as well as blinding for the comparison and experimental
intervention which is not included in the original Thorpe 2009 paper. This has been discussed extensively following the original publication but should be clarified in future guidance information for PRECIS-2.

It is interesting that two of the domains that have been suggested are controversial (in the case of blinding and cost analysis) but are not necessarily on the continuum of pragmatic to explanatory. When thinking about defining a pragmatic trial using usual or real world conditions, blinding may not always distinguish an explanatory or pragmatic trial. It is generally accepted that trials should aim to have high internal validity and lack of blinding is one of several potential sources of bias including generation of random allocation, concealment of random allocation, incomplete outcome data and selective reporting. Triple blinding of participants, personnel and outcome assessors would support high internal validity in a clinical trial, regardless of how pragmatic it is. It is not however, always possible to blind in a trial, easier for drugs but not possible for pragmatic trials with therapy interventions like physiotherapy. For instance in this trial, described as pragmatic by the authors, comparing usual care to family therapy for teenagers that have been hospitalised for anorexia nervosa, it is not possible to blind participants or therapists to the intervention but the outcome assessors were blinded to randomization [113]. So all that could be done to blind was done and the trialists then have to assess if the fact that the participants receiving and delivering care not being blinded introduced bias. This issue will be considered, in brainstorming (Chapter 5) and when the external and internal validity of clinical trials is considered in chapter 7.

There appeared to be consensus in the feedback on not considering aggregate scores for the PRECIS-2 domains. We believe it is unlikely that most PRECIS users will use an overall pragmatism score (adding up all the scores for each domain) without considering that the individual domain scores may be masked by an aggregate score. For instance, using a score of 1 to 5 for each domain a total score of 30 would not necessarily mean the 10 domains with individual scores of 3 but some domains may have
higher scores and others lower indicating greater or lesser pragmatism. Tosh et al, however, did publish using PRECIS to create the Pragmascope to consider how pragmatic 10 protocols were, using a 1 to 5 scale for each of the 10 PRECIS domains [55]. The aggregate scores could thus range from 0 (explanatory) to 50 (totally pragmatic) [55]. Tosh et al determined after considering 10 protocols that a more explanatory study undertaken in more ideal circumstances would have scores of 0 to 30 and a total score of >35 would be more pragmatic with 30-39 indicating a mixture of pragmatic and explanatory domains [55]. Selby discussed composite scores and said they should be avoided as trials that have completely different domains scores could have the same overall score – PRECIS enabled raters to consider the differences – and aggregation would mask them [40]. Glasgow backed up this view [56] and indeed Spigt and Kotz [38] were critical of another tool for assessing how pragmatic or explanatory trials were published by Gartlehner that added up scores of seven items [37]. Finally in the original work to develop PRECIS, Sackett discussed a summary number could indicate “Pragmaticness” but it would “hide the individual spoke scores which may have extreme values” [65].

Due to the feedback received in our Delphi and as we focussed on improving and validating PRECIS to design clinical trials, we did not consider aggregate scoring further in this project. Of note is that three of the Delphi participants were systematic reviewers so while this will not be pursued in this project, using PRECIS for systematic reviews may be considered in subsequent work for scoring trials included in systematic reviews. One of the these Delphi participants stated “I think PRECIS is a useful tool for synthesising and comparing trials in systematic reviews, so this would be interesting to see more consideration of how PRECIS might be used in this context.”

Source of the patients or the setting for recruitment was a suggestion for a domain lifted verbatim from the first round of the Delphi. This information was not in the Thorpe 2009 paper though included in the suggestions for CONSORT extension for Pragmatic trials published in 2008. This domain and information could be included in a new PRECIS-2 model but we needed to decide if we should include
this in the domain *Eligibility criteria* or as a separate domain. This will be discussed in Chapter 5 Brainstorming.

Cost was also a domain that was suggested but again it was difficult to determine how this could be applied on an explanatory/pragmatic continuum. Many trials have cost analysis work carried out in parallel with a randomized trial of the intervention to determine whether putting a successful intervention into practice will be cost effective. The implication of this is that there may be no estimate of cost of the intervention at the time that the trial is being designed, which would make a cost element practically impossible to include as a domain. In any case, even if cost estimates were available, we contend that the PRECIS-2 instrument that we are developing should not be concerned with feasibility or affordability of the intervention, as we are interested in developing the tool to help design trials that will give the results intended and encourage results are more likely to be applicable, irrespective of the cost of the intervention itself; which we view as a separate issue in the social decision to use, or not use, an intervention once it has been shown to be effective. One person also thought costs could also be included in participant compliance, perhaps suggesting incentives. We found this confusing, as it implies that explanatory would mean adding incentives and usual care without incentives thus be pragmatic- which seems a fairly specific domain for what is intended to be a generically useful tool.

“Main objective of the trial e.g. inform practice or register a new drug”. Like “cost, budgets” it is very difficult to determine the full spectrum of pragmatism in this domain suggestion. Perhaps a more reasonable way of incorporating these aspects into PRECIS would be to include guidance in a document for using PRECIS-2. For instance considering costs: these may be negligible if “usual practice” but in some countries where interventions are being tested these may have significant costs although the rest of the trial is highly pragmatic. High costs, however, may be worth it if the treatment is effective, but how can you rule “pragmatism” on this domain? This is another area which is discussed further in
Chapter 5 on brainstorming in addition to the information/consent domain that was suggested for inclusion.

The response to the final question on weighting of PRECIS domains indicated that half of respondents thought weighting should be done on a “trial by trial basis” or were unsure. We will thus consider mentioning weighting in the guidance on using PRECIS-2. The PRECIS wheel as it currently stands is visually clear but as one respondent stated, if we introduce weighting on a trial by trial basis, it would not be clear to those looking at the design of a specific trial.

One domain that we decided not to give as an option for round 2 was the Consent domain into the PRECIS questionnaire for Round 2. This issue was raised by one trialist in a free text response to question 2 in the first round – “Do you like the current PRECIS tool...?”. The reason this was suggested is that the patient information given in most current RCTs differs dramatically to the information which patients are given in routine real world healthcare. With concerns that when patients are aware of random treatment choices they may seek healthcare elsewhere and in real world healthcare there is lack of patient information. It is a serious issue and debate with calls for more naturalistic trials, undertaken at point of care [114, 115]. We decided, however, this was not an issue that was part of a pragmatic to explanatory continuum and therefore we would not discuss this as part of PRECIS-2.

**Strengths**

The key strength in our modified Delphi approach was that we tapped into an expert group of well-informed participants who were knowledgeable about PRECIS. There was a breadth and wealth of experience that was clearly indicated by the qualitative research that we received in free text responses. In addition, the participants were highly motivated and enthusiastic; many wishing to participate in later stages of the project (71% -22 out of 31 responses for both user testing and Validity testing). A few of those that could not participate sent apologies, were very positive about the development of PRECIS and wished to be kept up to date with the project.
Unlike the tool developed by Gartlehner to distinguish efficacy from effectiveness studies we used a literature review as the basis for a modified Delphi to inform brainstorming with experts. Gartlehner et al simply used a literature review, along with methodological and clinical concerns, to discuss with an expert panel the creation of a tool with seven domains [37, 116]. We believe that our development of PRECIS-2 has a stronger methodological basis.

**Limitations**

**Invitation sender**

It is possible that there was reduced participation as the e-mail Delphi invitation was not sent out by a well-known methodologist or researcher. Thus if recipients were busy they may have overlooked the invitation to participate in this research. Perhaps a higher response rate could have been attained if the e-mail had been sent via one of the supervisors for the PRECIS research or they had been copied into the invitation e-mail – though this would have resulted in an increase in e-mails and administration dealing with responses to the invitation. We did, however, have all the names of those involved in the Delphi invitation at the end of the e-mail and it was sent:

“On behalf of Shaun Treweek, Frank Sullivan, Merrick Zwarenstein, Peter Donnan (PhD Supervisors and co-applicants on CSO grant CZH/4/773 *MAKING CLINICAL TRIALS MORE RELEVANT: IMPROVING AND VALIDATING THE PRECIS TOOL FOR MATCHING TRIAL DESIGN DECISIONS TO TRIAL PURPOSE*)”

**Sample size and response rates**

There is limited literature to determine the optimal sample size for a modified Delphi survey - the original studies have used small panels of 8-12 but some panel sizes have been very large with several thousand people, more like a quantitative survey [105]. A large sample size could lead to difficulty analysing the data, particularly with open responses and qualitative data [96]. We believe that our sample size was large enough to produce a range of useful and informed opinions, which have been helpful in guiding reformulation of PRECIS.
There are a range of response rates in published Delphi surveys from a very low 8% to a much more acceptable level of 70% [105]. A similar modified 3-round Delphi to PRECIS was undertaken for CONSORT to determine the minimal information that should be reported in journals on Randomised Controlled Trials [107]. Hopewell et al had a sample size of 109 with a response rate of 61%. Another similar study also had a high response rate of 123 out of 167 (74%) with 104 accepting the invitation (62%) to participate in a 3-round consensus Delphi for SPIRIT (Standard Protocol Items: Recommendations for Interventional Trials) Initiative in 2007, with the primary aim of increasing the transparency and completeness of trial protocols [108]. We consider our first round response rate of 47% (42/90) with participation rate of 37% (34/93), followed by second round response rate of 72% (23/32) to be favourable given the purposive sampling and the time period for our modified Delphi.

**Participants**
Using the Delphi consensus technique is the first but important part of the process to improve and validate PRECIS but we acknowledge that the results are the opinions of the group we invited to participate. If different experts had responded perhaps we would have had different results in percentages, however we are confident that by inviting everyone who had cited PRECIS and most likely to know about the tool to participate if they chose to do, so we have gathered informed opinions. One of the limitations of consulting only authors who had cited PRECIS and then responded to the invitation to participate in the Delphi process is the danger of a skewed set of opinions influencing the development of PRECIS-2. While this is possible, out of the 90 citations at the time the reason for the citation varied greatly. Some of the multidisciplinary authors had simply cited PRECIS when indicating pragmatic design in discussion papers or completed trials, others had used the original reference as part of a group of references to discuss their design choices: “the assessment report could provide structured information on expected relevance of pivotal trials to use in practice along the lines of, for example, the PRECIS framework” [117], and in some publications it was unclear why they had included the original PRECIS article. Authors who had undertaken methodological work were excluded from the
Delphi so participation was from authors with less in-depth knowledge of the tool. We did, however, think it would be too hard to discuss issues related to PRECIS-2 with participants who did not know anything at all about PRECIS as the learning curve would be too great.

The responses however, of those we did ask may have been limited by time and experience of both the PRECIS tool and clinical trial experience [118]. As it is generally accepted that expert opinion does not constitute the highest level of empirical evidence [15], the next stage of brainstorming and subsequent user testing (which would include consultation with participants who did not know anything about PRECIS) would then sort out the “truth” of the Delphi work. Thus we believed the limitation of the participants that were selected for the Delphi was acceptable, given that it was one of several forms of consultation used in the project.

Conclusions: proposed changes to PRECIS

The following proposals to improve PRECIS, which included free text responses, were discussed by the Steering Group (ST, FS, MZ) led by KL to be taken forward as recommendations for discussion at the brainstorming meeting in Toronto, March 11th 2013 (Chapter 4). These ideas were understandable, feasible but in particular the domain suggestions needed discussion through brainstorming to clarify if they would be helpful to trialists and how they would work in practice to improve PRECIS-2 considering the pragmatic to explanatory continuum. Additional input was also sought from the Dundee brainstorming group (June 2012) (see Chapter 4) to finalise the models for Phase 2 user testing (Chapter 6).

- Use Excel and conditional formatting to produce both a wheel and a table with explanations for decision making
- Use 5-point Likert scale
- Insert additional domain Source of the patients or The setting for recruitment OR include in Eligibility Criteria (with specific guidance on this area)
• Insert the following as additional domains OR include in advice for designing pragmatic trials:
  Information/consent;  
  Main objective of the trial e.g. inform practice or register a new drug”;
  Budget/Costs.
• Domains to consider removing:  Analysis
• Design -  Practitioner expertise currently diametrically opposed to  Participant compliance – confusing.
• Revise information on using PRECIS with additional examples and table with key points.
  Include:
  o How do trialists trade-off between domains when faced with time or budget constraints?
  o What is the highest priority for investing in pragmatic evidence?”
  o Masking/Blinding
  o Weighting

Conclusion

We successfully collected and selected a number of ideas for changes to the PRECIS tool to take forward to brainstorming meetings, drawn from expert consensus generated in a two-round electronic Delphi survey. We also recruited participants to assist in further phases of the improvement and validation of PRECIS-2.
SUMMARY
Two rounds of a Delphi were undertaken in two months. A total of 90 participants who had cited PRECIS (2009) were invited with a first round participation rate of 37% (34/93) and second round of 72% (23/32). 62.5% thought PRECIS needed a scoring system: 1-3, 1-5 and 1-7 most popular. Suggestions for new domains included 52.2% wanting “Source of the patients or the setting for recruitment”, 26.1% (respondents could choose more than one domain): “No extra domains”, “budget/costs”, “Main objective of the trial e.g. inform practice or register a new drug”; and 78.3% respondents said we needed to improve the original guidance for PRECIS.

CONCLUSION
Anonymous information was efficiently gathered from participants using an electronic survey. The Delphi provided results for further face-to-face brainstorming. In addition, an enthusiastic group of supporters was enlisted for further work on the development of PRECIS-2
Chapter 5: Brainstorming

Introduction
Brainstorming at its best: “A well facilitated group has the feel of an everyday discussion, with participants interacting, joking and arguing with each other, rather than through the facilitator.” (p117) [82]. As mentioned previously (Chapter 3) one type of group meeting that we selected to use in developing PRECIS-2 was the brainstorming technique, ideally as just described, allowing a more free flowing discussion with a group of participants. A key issue requiring early consideration was who to invite to the brainstorming sessions. We were also aware that we needed to invite more than 25% of the sampling frame to allow for unavailability on the date or participants dropping out having accepted the invitation [82]. Ideally groups would be 6-12 participants [82].

As we had decided to hold three different brainstorming meetings in the project to improve and validate PRECIS we had to carefully consider who we invited. There were two different groups we wanted to be represented to ensure optimal input into designing the new tool: firstly people who were familiar with the original PRECIS tool [48, 49], and knew the benefits of what we were trying to achieve by improving the original PRECIS tool [40, 53]; and secondly, future users of the tool (trialists and healthcare professionals) that were interested in producing and using health research that is highly relevant and easily translatable into everyday practice [14, 119]. For the latter, a simple training package would be needed to bring these participants up to speed with the aims of the project to enable them to contribute.

Green et al state that the brainstorming facilitator does not have to be an expert in the field that is being investigated; in fact this can be an asset ensuring impartiality of any interaction [82]. (KL) was not an expert in trials or facilitating research group meetings but with assistance from supervisors (ST, FS and MZ) the advantages of having a non-expert with group facilitating expertise through the
assistant facilitators ensured the quality of the brainstorming meetings would not be reduced. In addition, the experience of the PhD supervisors enabled fruitful debriefing following meetings to ensure both the experience of facilitating and the content of the meetings was discussed [82].

For brainstorming, it is important to ensure all participants felt able to contribute and discuss with their fellow group. Dominant characters should not be allowed to control the meeting [84]. The agenda was of central importance and that participants were clear on their remit and kept on target by the facilitators and not allowed to “focus” on one particular topic at the expense of time discussing other important issues on the agenda [84]. It was important that the structure of the meeting was carefully considered and included an introduction to allow participants to introduce themselves and to ensure everyone was aware of the goal of the meeting and the intended outcome. A topic guide helped ensure nothing was missed. Prompts and probes were used by the facilitator (KL) and co-facilitators (ST, MZ, FS) to elicit information, for instance, “anything else” but leading questions that would encourage participants to answer in a way directed by the questioner were avoided [82]. The number of ideas therefore generated in the brainstorming meetings were determined by several factors: the participants and their interaction with each other, the facilitators, the time allowed for the brainstorming meeting, and the interest in the topic being discussed. “Social interaction” between participants and group “cohesiveness” was important in determining the outcome from brainstorming meetings and should not be overlooked [84] as described earlier [82].

This chapter will describe the three brainstorming meetings that we used to discuss the creation of PRECIS-2. Each of the three brainstorming meetings will be discussed separately as each played its part in moving from the original tool PRECIS to the tool that was then used in validity and reliability testing and each meeting had different aims and objectives, participants and results.
First brainstorming meeting

Aims
To discuss using the PRECIS (Pragmatic Explanatory Continuum Indicator Summaries) tool for improving the design of trials.

Objectives
The aim of the first brainstorming meeting was achieved through a number of objectives:

1) Recruit local (University of Dundee) trialists, methodologists, academic healthcare practitioners.
2) Use a training package (PowerPoint) to educate brainstorming participants about PRECIS.
3) Determine what participants thought about using the PRECIS tool to score two trials prior to the meeting (if possible), one more explanatory—a non-pharmacological intervention [120] involving radial versus femoral access for coronary intervention. The second trial was more pragmatic, and involved a pharmacological intervention treating asthma with leukotriene antagonists as first line or add on asthma controller therapy [121]
4) Determine if the PRECIS tool could be improved through discussion of: domains, scoring, weighting, strengths/weaknesses, how easy was the tool to use on scale of 1-10 (1 = very easy, 10 = very difficult).

Methods
Participants
On the 11th May, 2012, 15 participants (Table 18) were invited (Appendix Box 5.1) to attend a meeting on Friday 8th June 2012 for up to two hours. The participants were based at the University of Dundee and the Tayside Clinical Trials Unit (TCTU). Participants were principally suggested by ST and FS as researchers who may be interested in assisting with the project as PRECIS may be relevant to the work these researchers were involved in.
Pre-meeting preparation

Participants who had agreed to attend were sent an email two weeks prior to the brainstorming meeting with a zipped file including:

- PRECIS wheel with 4cm long spokes (Figure 9)
- The pdf of the original Thorpe 2009 [49] article on PRECIS
- PRECIS training package with 14 PowerPoint slides
- The pdf of Jolly 2011 [120] – more explanatory trial, non-pharmacological intervention
- The pdf of Price 2011 [121] – more pragmatic trial, pharmacological intervention

The PRECIS training package had been pilot tested by the three PhD supervisors (MZ, FS and ST) but this was the first time that it was used by people who were unfamiliar with the PRECIS tool. It included some background information on PRECIS and its development from the ideas of Schwartz and Lellouch [28] on pragmatic and explanatory trials, the importance of considering how pragmatic or explanatory a trial design is as a continuum NOT a dichotomy, a table describing the 10 PRECIS domains descriptions by Sackett [33]. It also contained two worked examples using PRECIS considering the 10 domains on two very different trials, one more pragmatic trial on music therapy for autism [122] and the other a more explanatory trial treating asymptomatic adults with elevated coronary calcium scores with Atorvastatin, Vitamin C and Vitamin E [123].

The participants were asked to read the training package and the original article on PRECIS by Thorpe et al. and then look at the two trials by Jolly [120] and Price [121] prior to the meeting. If there were time constraints, they were encouraged to take a look at one of the trials. A blank PRECIS wheel was included in the package so that they could plot their PRECIS wheel for each of the trials, considering the 10 different domains mentioned in PRECIS.

Participants were told in the e-mail that their remit at this brainstorming meeting was to discuss their experience of using PRECIS. The e-mail also indicated that we were interested in discussing the scoring
system (as the original tool had no scoring system) and keen to discuss the 10 PRECIS domains: was there anything individual participants thought should be included and anything they thought could be missed out? Should there be weighting of domains, so some domains more important? Finally, we were interested to know how long it took individuals to use PRECIS on a trial and how easy it was to use on a score of 1-10 (1 = very easy; 10 = very difficult).

Figure 9 The original PRagmatic Explanatory Continuum Indicator Summary from Thorpe et al 2009 [48, 49]

Protocol for the meeting
An agenda was prepared and followed (Appendix Chapter 5 Box 5.2). An audio transcript of the meeting was taken after gaining permission from the participants and assuring that any comments would be anonymised. We also used a topic guide to ensure all questions were asked during the meeting and nothing was left out. In addition, notes were taken by KL and ST.
**Table 18 Participants invited to participate in first brainstorming meeting**

<table>
<thead>
<tr>
<th>Participants</th>
<th>Job description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annie Anderson</td>
<td>Professor of Public Health Nutrition Trialist, interested pragmatic designs and applicability</td>
</tr>
<tr>
<td>Heather Cassie</td>
<td>CSO research fellow, dentist</td>
</tr>
<tr>
<td>Janet Clarkson</td>
<td>Director of the Effective Dental Practice Programme, Honorary Consultant in Paediatric Dentistry, Dental Health Services Research Unit (DHSRU)</td>
</tr>
<tr>
<td>Peter Donnan</td>
<td>Professor, Co-Director TCTU, statistician, methodologist, trialists (co-applicant CSO grant)</td>
</tr>
<tr>
<td>Vikki Entwistle</td>
<td>Professor of Values in Healthcare, Social Dimensions of Health Institute at the Universities of Dundee and St Andrews.</td>
</tr>
<tr>
<td>Bruce Guthrie</td>
<td>Professor Population Health Sciences, Academic GP, trialist Primary care, multi-morbidity, polypharmacy, applicability</td>
</tr>
<tr>
<td>Fiona Hogarth</td>
<td>TCTU Assistant Director, aware of design issues, first point of contact for investigators coming to Tayside CTU</td>
</tr>
<tr>
<td>Martyn Jones</td>
<td>Reader, Leads the Nursing, Midwifery and Allied Health Professions Training Scheme Nursing. School of Nursing and Midwifery</td>
</tr>
<tr>
<td>Nora Kearney</td>
<td>Professor of Nursing and Cancer Care Research Dean, School of Nursing and Midwifery</td>
</tr>
<tr>
<td>Thomas Lamont</td>
<td>Dentist and Trialist on HTA trial (IQuaD – Improving the Quality of Dentistry)</td>
</tr>
<tr>
<td>Roberta Littleford</td>
<td>Senior Trial manager, TCTU.</td>
</tr>
<tr>
<td>Joseph Liu</td>
<td>Senior Research Fellow, Dental Health Services Research Unit</td>
</tr>
<tr>
<td>Stephen McSwiggan</td>
<td>Senior trial manager, Tayside CTU.</td>
</tr>
<tr>
<td>Professor Nigel Pitts</td>
<td>DHSRU Director, Professor of Dental Health (University of Dundee), Honorary Consultant in Dental Public Health with NHS Tayside Dentistry</td>
</tr>
<tr>
<td>Frank Sullivan</td>
<td>Professor - Director Population Health Sciences, Supervisor for PhD</td>
</tr>
<tr>
<td>Brian Williams</td>
<td>Director of Nursing, Midwifery and Allied Health Professions Research Unit, behavioural scientist, interest in complex intervention trials and pragmatic designs</td>
</tr>
</tbody>
</table>

**Facilitators**
The facilitators for the brainstorming meeting were KL and ST. FS (supervisor) was a participant.

**Analysis**
The transcript of the meeting was created by an independent transcriber [75] and used along with notes to consider feedback on the tool, in particular on: first impressions, ten PRECIS domains, scoring, weighting, strengths/weaknesses and how easy the tool was to use. This information was then used to inform the development of the Delphi.
Results

Participant attendance

Four participants plus two facilitators (KL, ST) (Table 19) attended the brainstorming meeting. No replies were received from six of the invited participants, although two were on holiday when invitations were sent out. One of the participants asked if the invitation could be forwarded to another trialist, a dentist who would be interested, so in total 16 participants were invited. This participant (TL) did attend the meeting. Professor Bruce Guthrie, was unable to attend on the day but was keen to be involved and sent comments and his PRECIS “wheels” for the two trials we discussed at the meeting.

In addition, apologies were received from: Annie Anderson (Professor of Public Health Nutrition), Jan Clarkson (School of Dentistry), Martyn Jones (Reader, Department of Nursing and Midwifery), Frank Sullivan (Director Population Health Sciences).

Table 19 Participants who attended the first brainstorming meeting on PRECIS

<table>
<thead>
<tr>
<th>Participants</th>
<th>Job description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heather Cassie</td>
<td>PhD student, dentist</td>
</tr>
<tr>
<td>Peter Donnan</td>
<td>Professor, Co-Director TCTU, statistician, methodologist, trialists (co-applicant CSO grant)</td>
</tr>
<tr>
<td>Fiona Hogarth</td>
<td>TCTU Assistant Director, aware of trial design issues, first point of contact for investigators coming to Tayside CTU</td>
</tr>
<tr>
<td>Thomas Lamont</td>
<td>Dentist and Trialist on HTA trial (IQuaD - A randomised controlled trial comparing oral hygiene advice and periodontal instrumentation for the prevention and management of periodontal disease in dentate adults attending dental primary care.) (Suggested by Heather Cassie)</td>
</tr>
</tbody>
</table>

Initial impressions

Generally the participants were positive about PRECIS. “…very useful…learning tool”; “…makes me consider “is that going to be applicable?” and “…make sure everything’s talked about”.

129
As one participant mentioned it is subjective and up to the judgement of the individual rater: “...it’s very open to the interpretation of the person reading it as to whereabouts quite on the scale you think”.

The PRECIS tool also educated the participants in considering the meaning of pragmatic and explanatory: “...it’s very interesting you can have a trial that is both pragmatic and explanatory in the different areas (domains)...but we’re trying to...make them as pragmatic as possible to make them as valid as possible in actual general practice so it (PRECIS) is very interesting” and “...useful to focus my mind on the process of what eventually gets up to these more pragmatic trials”. One comment indicated that it superficially could be quite simple but the design details of differentiating between more pragmatic and more explanatory was not easy. In addition, the idea of a continuum had not been picked up: “I quite liked the idea of explanatory vs pragmatic (does it work under ideal vs does it work under real)...I find it all rather disintegrates into vagueness when considering the detail.”

How easy is PRECIS to use?

All found PRECIS on a scale of 1 to 10 (0 = very easy; 10 = very difficult) usable (Table 20). The median score was 5, with a range 3 to 6. Most of the participants took 15 minutes to score each trial [120, 124] with 30 minutes for one rater.

<table>
<thead>
<tr>
<th>Name</th>
<th>Approx. Time</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>HC</td>
<td>30mins</td>
<td>6</td>
</tr>
<tr>
<td>PD</td>
<td>15mins-20mins</td>
<td>&lt;5</td>
</tr>
<tr>
<td>FH</td>
<td>15mins</td>
<td>&lt;5</td>
</tr>
<tr>
<td>TL</td>
<td>10-15mins</td>
<td>&lt;5 to 3</td>
</tr>
</tbody>
</table>

Training Package

All thought training package good and had no comments for improvement. “...worked examples ...were very useful...” helped with breaking down different domains into considering how pragmatic or explanatory a trial was....“ah right, that’s what I’m looking for here...”.
Scoring

Three out of four participants felt having a quantitative way of using PRECIS was important and using a score that had a midpoint was also crucial if people were not sure whether design more explanatory or more pragmatic. Groupings of scores might be an option to explore.

“It would be quite hard to show differences just with the diagram”

“From a statistical point of view I think about the measurement, should it be a Likert 1 to 5, should it be 1-10?”

“...I felt that it was very subjective.”

Alternative views

“I suppose if it’s going to become a comparative tool but if it’s a tool to sort of design your trial and it’s a starting point for a discussion, then I would be comfortable without a score I have to admit...”

One idea that was not backed up by the other participants was grades of colour instead of scores - “rather than scoring could it be a colour thing so it goes from black to white or ... green to red or ...?”

Cumulative score

Two of the participants were interested in using a composite score for the 10 domains but would think about weighting. Otherwise, if adding up all the domains, and some domains very different scores, the cumulative score could appear skewed.

Design

Design was the area all participants gave feedback on. On the original PRECIS tool, the more logical way of using PRECIS is to move round in an anti-clockwise direction, starting with Eligibility criteria moving to Flexibility of experimental intervention and finishing up with Primary analysis.

Suggestions included:

- clockwise movement more instinctive than anti-clockwise
“... I would... just go round clockwise would be my thing, but then to have them follow on, so if you said
... Eligibility criteria to start, then maybe Flexibility of experimental intervention, Flexibility of the
comparison intervention, you know sort of followed it round.”

- **Pragmatic label on rim** – donut shape on the rim with pragmatic written

“...there should be an outer circle that says “pragmatic”...

“...an outer wheel going all the way round...”

‘cause at the moment you’re doing the counter fact show if you like in the explanatory...it seems to be
the centre of attention...”

- **Domain labelling could be better**

“I think the labelling isn’t very good...“(multiple agreement).

“I think you needed the background” (multiple agreement)

“Flexibility of the intervention versus the comparison, and the Practitioner expertise on the two
arms...”

“Why explanatory was in the middle, I was thinking of visually you could have it the other way round
that pragmatic in the middle...”

“...I don’t see pragmatic listed anywhere...”

“...find the labels... bit terse...”

- **Ordered table instead of “spider diagram”**

“Try out an ordered table, where the ordering reflects the nature of the different spokes”

- **different colours for different domains**

“coloured cobwebs in my head of all the different trials and the different domains... a spectrum ... so
you know red’s always going to be eligibility criteria, you know then you could pick it out looking at
different trials...”
No domain information

One of the problems highlighted in using PRECIS retrospectively is reporting of trials. What do you do if there is no information in the publication on different aspects of the trial design meaning that you cannot judge how pragmatic that domain is?

Missing domains

Additional domains to consider where suggested by the statistician: power and sample size. And a domain that covered the intervention and “applicability suitability or expandability”, including Cost-effectiveness “the likelihood of the intervention actually being adopted”. Something that covered “clinical applicability or feasibility”, “clinical transferability”. “Is it affordable in that healthcare system, do people want to do it…”

“It doesn’t come out as cost effective...(but) it’s an effective drug.”

“It brings in the context though ‘cause this is a tool that could be used internationally…”

“..an intervention wheel health economics, transferability, applicability, quality of life, patient acceptability…”

Eligibility criteria

All agreed that the domain “Eligibility criteria is absolutely critical”.

In particular “expand the Eligibility criteria information in some way because that is both key and it would be interesting to think more carefully about the individual eligibility criteria because it could be highly pragmatic on some and very explanatory on others and it would be good to be explicit about that”.

In the paper that was discussed by Jolly [120] the “cut-offs” were age, so “patients greater than 60, so a 59 year old is immediately eliminated, “why?”; chest pain lasting longer than 10 minutes so if 9.9 minutes...not eligible but perhaps over 5 minutes is similar enough to be eligible.”

“It might be that actually 5 minutes is important to the patient rather than 9 minutes…”
“are they trying to get as many people into the trial, ’cause ultimately that’s going to, for me, impact on the generalizability of the results...”

**Multi-centre trials**

If trials are carried out in more than one centre (multi-centre) then there are more practitioners, thus there is most probably more variability in the intervention. The current guidance on PRECIS does not include any information about multi-centre trials or indeed context.

“(flexibility) it does depend on how many centres you’ve got, if there’s only two centres then you’re not getting much variability in practice...”

“whereas if it was multicentre instead of single centre then you might allow complete flexibility.”

“perhaps there should be another domain looking at context and where a trial was cluster randomised or whether it was multi-centre...”

“...may be some confounding issues with this (cluster randomisation).”

“...could a trial by definition ever be pragmatic if it was a single site, for example?”

“John Nichols in his presentation talked about the setting of giving treatment” [125]

“If 10 different sites and the usual practice is the same in all of them, if the definition of pragmatic is the usual practice, then it might not make a difference one way or the other whether it’s one site or 10 sites..” “If consistent.”

“A common criticism of site selection (for trials) is that there’s an awful lot of research done in academic centres, which are different...to non-University hospitals.”

**Flexibility of experimental and comparison interventions and Practitioner expertise**

(experimental and comparison)

There was a great deal of discussion about these domains and what should be in each. The participants in the meeting believed that; part time practitioners; number of procedures each day, each week; the skill level in procedures would all influence how pragmatic or explanatory these domains were.

“I struggled with the Flexibility of the comparison intervention...”
“If the population studied is broad/close to real world, then some of the other questions have very important implications but aren’t critical in the same way. For example, the tightness/implicit required fidelity of the intervention makes an intervention more or easier to implement, but the main implications are for what you’d need to do to replicate the results in real life. For example, angioplasty in acute MI trials or thrombolysis in stroke trials had very tightly defined interventions, and there was real doubt about whether the very stringent timing requirements could be achieved in routine practice, but in practice they could be once you accept that there’s no point doing these interventions unless you organise to deliver them as they were trialled. But that is rather different from the population studied issue. “

**Follow-up intensity**

This domain prompted discussion around follow up as part of the intervention or purely for data collection. Knowing the norm for routine data collection was important for different interventions.

“...if its brilliant pragmatic and perfectly pragmatic trial I would still like, as a statistician I would like quite a lot of data, so the number of collection points to me doesn’t make it, you know it...it’s not making it more explanatory because you are actually measuring more...” “if the follow-up is saving routine data then I don’t see any problem...”

However, if follow-up is trying to influence outcomes and reinforce of behavioural change for instance then it is classified as part of the intervention.

“...if you are explicit then I still think you could have a follow up as being at the pragmatic end, because you’re actually saying it’s not the follow-up, it’s part of the intervention...”

“But, if you do, if you’re not thinking of it in those explicit terms then I think you, you then move it to explanatory for the reasons that you can influence the outcome.”

**Participant Compliance with prescribed intervention**

In the surgical trial by Jolly [120] compliance by patients being operated on was discussed. How should that be measured in this domain? All three options were given: pragmatic, no score or explanatory.
“...that’s completely pragmatic, they (patients) have no say at all”

“..if you turn up and have it done you complied fully...”

“...miss off this domain scoring as no choice”

“...completely explanatory as well ‘cause it is, outwith their control”

**Outcomes**

In this PRECIS domain, there was discussion about meaningful outcomes and primary outcomes. The participants discussed in depth outcomes that were meaningful to patients. But one participant highlighted the different viewpoints in choosing outcomes to measure “...must get a dichotomy between the clinicians and the non-clinicians...the clinicians want a hard outcome that’s physiological Peak Expiratory flow Rate (PEV1) whereas the rest of the group wanted quality of life...”

**Primary Analysis**

There was discussion on the description of this domain, with a suggestion to call this domain Intention-to-treat (ITT) as usually explanatory and pragmatic trials use ITT; “Why can’t you label it ITT or not...rather than call it Primary analysis?”

**Weighting**

This topic prompted interesting discussion that suggested that it would be difficult to have fixed and equal weighting for different domains in PRECIS and that one option could be that trialists decided to give different weightings for the domains depending on the individual trial that was being designed. “As it stands you give equal weight to the flexibility so that you might think it’s better to give more weight to a more flexible intervention...”

“But as it stands, you’re saying that all the spokes are equal...”

“Would the weight of different things not change for what you’re actually looking at?”

“one particular trial might be the eligibility criteria might be a particular weight, whereas on another trial it might the follow up intensity...”
Summary

Suggestions that were discussed to modify the PRECIS tool are summarised in Table 21.

Other uses for PRECIS

There was discussion that PRECIS was useful for:

- “evaluating and appraising trials for decision makers…”
- “good training tool for students in how to appraise and understand trials…” in particular that “….amongst young clinicians early on in their careers that you might want to consider other things than lab-based measures.”
- “…so another way that we could use the tool locally to actually improve the quality of the (clinical research…what is the clinical impact of the research? …and to look at each of their trials and if you found that they were all around the middle to say to them. Do you think actually there is an argument now to take these studies out? You've done a lot of experimental ones and we know that there are small changes and there's, minor impact between these two drugs, but how does that actually impact on a population? So what we need to do is actually look at the way you were designing your trials and bring them out (make more pragmatic), which then might encourage different funding sources and more collaboration.”

Brainstorming meeting Participant interactions

From a qualitative perspective, on analysing for example the number of hits for “laughter” 46 or “laughs” 75 in the transcription, it is clear right from the first introductions that there was a good atmosphere. ST knew all but one (TL) of the participants, KL had met one of the participants beforehand (HC). Two of the participants (HC and TL) knew each other as both were based in the Dental school and the other two participants (PD and FH) had worked together on different trials.
Table 21 Possible PRECIS modifications

<table>
<thead>
<tr>
<th>Score</th>
<th>Design</th>
<th>Domains</th>
<th>Weighting</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-4</td>
<td>Pragmatic “donut” labelling round wheel</td>
<td>ITT instead of analysis</td>
<td>Fixed</td>
</tr>
<tr>
<td>1-5</td>
<td>Clockwise use instead of anti-clockwise</td>
<td>Add blinding</td>
<td>Variable</td>
</tr>
<tr>
<td>0-5</td>
<td>Change order of domains so group control and intervention</td>
<td>Add Setting</td>
<td></td>
</tr>
<tr>
<td>1-10</td>
<td>Improve labelling of domains</td>
<td>Add cost analysis</td>
<td></td>
</tr>
<tr>
<td>1-4cm</td>
<td>Ordered table rather than “wheel”</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Composite</td>
<td>Top score for explanatory rather than pragmatic</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Colours for domains</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Discussion

There was a significant drop out rate prior to the meeting, this was despite inviting more than 25% of the number than were needed for the brainstorming meeting. This initial brainstorming meeting nevertheless appeared to be successful on several counts. Firstly, the participants were keen to stay and discuss PRECIS and the meeting over ran, indicating the interest in the topic. Secondly, the agenda with topic guide was followed and every item was discussed with important feedback for future stages of the project to improve PRECIS. Thirdly, this was a successful training opportunity for KL in co-facilitating a brainstorming meeting. Fourthly, the training package received positive feedback. Finally, there was humour and laughter amongst participants who clearly enjoyed the discussion (p117) [82] and all of the participants were keen to be involved in subsequent pilot testing and/or future brainstorming meetings.

The success of the training package (although limited sample size testing) acted as a model for further development of a training package or “tool kit” and website for the tool PRECIS-2 that we were developing (See Chapter 7B on PRECIS-2 database). At the time of brainstorming, there was only information about the development of PRECIS on the internet entitled “Spokes” [65] but this has been used by some of the groups that have tested using PRECIS and described as “a training package” [68, 73] along with the original article on PRECIS [48, 49] despite it describing the early development of the PRECIS tool and is not a tailor made training package for the final PRECIS tool that was published in 2009 [48, 49].

Discussion of the design of the PRECIS tool created most suggestions to make the tool easier and more intuitive to use. All of them made sense and could be relatively easily incorporated including: improving labelling for the domains to make it more descriptive, instead of the explanatory side with “E” in the centre being emphasised participants also suggested “Pragmatic” should be on the rim with a defined edge to the circle; better grouping of domains, for example by placing intervention and
comparison together; and domains placed in a more logical clockwise order. Participants also suggested using a table to indicate the rationale for decisions about how pragmatic or explanatory they had decided to make a domain.

Scoring of domains was also a key discussion topic with suggestions from Likert scale 1-5 to 1-10, with either pragmatic or explanatory scores being the highest. Participants needed to have a centre point though if they believed a domain was equally pragmatic and explanatory. However, one participant did think that scoring was not so important if designing a trial and the tool was used by trialists to stimulate discussion.

Setting, context and multi-centre trials were issues that did not appear to be covered sufficiently by the original PRECIS tool. In addition, the domain Outcomes may need further guidance. For instance, Thorpe [48, 49] guidance does not mention composite outcomes, whereas Sackett does. For the training package we used Sackett’s [33] advice in using PRECIS for the Outcomes domain for the Jolly [120] article - as this involved composite outcomes we determined this domain was more explanatory. Weighting was an issue that was of interest but there was no overall agreement indicating we should discuss this further; one participant suggested weighting would have to be considered on a trial by trial basis.

Domains that were proposed for inclusion by the participants were cost effectiveness and applicability of different interventions. Blinding was also discussed as a possible domain to be inserted into PRECIS, as this had been done by Bratton when he used PRECIS to consider a tuberculosis treatment trial [126]. He stated that in an explanatory trial on the Blinding domain “Participants and investigators blinded where possible to minimise bias” and in a pragmatic trial “Only independent assessors and laboratories blinded to minimise bias.” [126]. In the brainstorming meeting a discussion on internal validity arose as blinding is one of the criteria to determine internal validity as defined in the Risk of Bias tool [46].
The discussion mentioned that more pragmatic trials were generally considered to have poor internal validity [43] an issue addressed in Chapter 8C. The participants were neither for nor against including blinding as an extra domain.

Using PRECIS for systematic reviewing was mentioned; users may face difficulties assessing different domains if there is inadequate reporting of information. This may be due to authors not following CONSORT guidelines e.g. for pragmatic trials [127]. This was subsequently considered in guidance for PRECIS users to assess how pragmatic or explanatory a domain is if there is inadequate information (see Chapter 8B). But as our primary aim is to improve PRECIS to design clinical trials this is not our primary consideration.
Study strengths and limitations of the first brainstorming meeting

The facilitator (KL) was not an expert in trials or facilitating. The former is an advantage and the latter could have been a limiting factor but ST had experience and ensured that the first brainstorming meeting went well, stuck to its remit and the maximum amount of information was obtained from the four participants in a timely relaxed fashion. Although a minimum of six had been anticipated for the brainstorming meeting, two unfortunately dropped out on the day so despite sixteen participants being invited there were only four attendees. The input from the four participants was diverse but could no doubt have been more varied if there had been additional participation from the invited experienced trialists who represented a range of disciplines. If this was the single consultation stage of the project having a small group would have been a problem. However, as this was one of several stages of consultation I do not believe this to be a critical problem; other forms of consultation provided diversity that was less present in this brainstorming meeting. The quality of the feedback enabled us to move on to the next step of the project to improve PRECIS and create PRECIS-2.

KL did not use software to assist in reading the transcription of the brainstorming meetings, for instance NVivo for qualitative analysis of the content of the brainstorming meeting. As the meeting involved only four participants giving feedback on using the PRECIS tool, we did not think this was necessary and would over complicate the process of extracting relevant material.

Our use of the brainstorming technique compares favourably with similar projects that used brainstorming successfully as one of a variety of methods to meet a goal [128, 129]. Brainstorming in small groups and as part of plenary session at a conference was used to determine barriers and facilitators to implementation of guidelines for throat infections. The lead author analysed notes from these meetings to improve management. This was part of a study using a literature review, focus groups, a pilot study, small group discussions and interviews [128]. In a similar way, we also had purposive sampling of participants for the brainstorming meetings, inviting participants whom we
believed could contribute relevant information as trialists and health care practitioners to improve the PRECIS tool, thus random sampling would not have been appropriate. The composition of this first Brainstorming meeting, small due to drop outs, may have been deleterious to the development of PRECIS as only those genuinely interested or perhaps curious attended. However, as other methods were employed in the development process, we believed this would compensate for the small numbers involved.

**Conclusions**

This first brainstorming meeting was considered to be a success by the facilitators as all objectives were met. We had recruited trialists who were clinicians and able to contribute to the development of the next version of PRECIS. We had also successfully educated the participants on the PRECIS tool for designing clinical trials. In addition through inviting local participants we had communicated our project to a wider audience. We were confident we would be able to build on this enthusiasm for our next local meeting.

**Second brainstorming meeting**

**Aims**

To discuss possible PRECIS-2 models based on the Delphi and published methodological work on PRECIS.

**Objectives**

The aim of the second brainstorming meeting was achieved through a number of objectives:

1) Recruit expert panel of researchers from North America that included:
   a. Original designers of the original PRECIS tool
   b. Researcher who had published methodological work using PRECIS
   c. Policymakers
d. Trialists

2) Discuss proposed models of PRECIS-2 - including domains, modifications of descriptions for domains, scoring, weighting,

Methods

Participants
On 12th November 2012 KL invited all the non-Canadian participants (18 in total) individually to an expert panel brainstorming meeting in London Ontario in Canada or Toronto in Canada, on 27th November. 19 Canadians were invited using individual e-mail invitations (Appendix for Chapter 5, Box 5.3). To assist in recruitment to attend the brainstorming meeting, MZ who lived and worked in Toronto and was well known for his work on pragmatic trials, involvement in the original PRECIS tool and the CONSORT extension for pragmatic trials, was copied in to four of the invitations. A Doodle poll http://www.doodle.com/4g5fcgk9vpt5eh29 was used to check availability from March 11-26th 2013. Invitations were personalised but also categorised into two groups depending on whether or not the invitees had participated in the Delphi. A confirmatory email, including the agenda was sent out 11th January 2013 (Appendix for Chapter 5, Box 5.4)

Meeting organisation
The venue for the brainstorming meeting was important and MZ secured an invitation to use the facilities at the Theta Pharmaceutical building, centrally located in Toronto, in the University of Toronto with easy access to transport. A local organiser assisted in producing a headed participant list (Table 22) and an Agenda. Items on the agenda included a presentation of the project using a 17 slide PowerPoint presentation which would act as a topic guide for discussion of the Delphi phase and potential models for PRECIS-2, concluding after lunch with a discussion of the elaboration paper for PRECIS-2. This was distributed by e-mail to participants. As several participants were unable to attend in person but interested in online participation a free – non-payment system “Cisco WebEx” was tested prior to the meeting between Canada and Scotland and then the link was sent out to potential users (Appendix for Chapter 5, Box 5.5).
The meeting room also had flipchart which was used by MZ and an i-phone was used to record the meeting following agreement by the participants. A transcriber in Scotland was on standby to work on the recording to assist in post meeting analysis. Notes were taken by KL and ST to ensure maximum information on views were gathered from the meeting for post meeting analysis.

Initially we considered having a two hour meeting but decided we had sufficient discussion for a full day of brainstorming. Further invitations (Rodger Kessler (Delphi) 18th February; Janet Martin (local suggested by MZ) 28th February; Ba' Pham (former colleague of MZ) 11th February; Valeria Rac and Lusine Abrahanyan (host MK suggestion) 22nd Jan 2013) were sent when the date of the second brainstorming meeting in Toronto, Canada was finalised.

No preparation prior to the meeting was expected of the participants; we anticipated that all of the participants were aware of the original PRECIS tool as demonstrated by their attendance at the meeting.

**Incentives**

Unfortunately we were unable to offer travel costs, only local taxis from the airport would be funded by our travel grant with the Chief Scientist Office. Lunch and refreshments were provided.

**Reminders**

Reminders were sent on 27th November to American invitees. Reminders were then sent on an individual basis depending on when they initially responded or were invited.

**Facilitators**

Initially, the proposed facilitators for the brainstorming meeting in Canada were KL and MZ. At the time invitations were sent out ST was unable to attend but became available shortly before the brainstorming meeting.
At the meeting, each had a role, MZ as the local Canadian introduced the meeting, ST acted as Chair, KL presented items on the Agenda in the morning session and ensured topic guide with key questions was covered in discussion. In the afternoon MZ took notes on flip chart to assist final discussion on PRECIS-2 models and facilitate post meeting discussion with ST and KL. Final round and thanks to participants was given by MZ.

Post-meeting brainstorming

ST, MZ and KL had set aside over two days to continue with brainstorming following the initial meeting using the flip chart notes and notes taken by all three facilitators while the meeting content was fresh. This time was used while all three facilitators were in the same place to expedite the production of ideas for PRECIS-2 models which would then be user tested – the basis for this would be the elaboration paper which would explain how PRECIS-2 could be used and the detailed explanation of the domains which were considered through a continuum of pragmatism from very explanatory to very pragmatic.

Results

Participants

There were 23 participants (Table 22) who had stated they would attend the meeting. This included webinar participants that would “drop in” on the web link when they were able. On the day, one participant was not able to attend in person due to illness but joined the brainstorming meeting using the web link. Two other participants did not attend on the day and four web participants did not log in to join the meeting. Funding to attend and pay for travel expenses was an issue for three participants but due to the nature of the meeting one participant (published methodological work on PRECIS who had participated in the Delphi) managed to find funding to attend. In total 14 individuals participated in the meeting, in addition to KL the facilitator, MZ the convenor and ST the co-facilitator (Figure 10).
### Pragmatic Explanatory Continuum Indicator Summary (PRECIS) Meeting

**Monday, March 11, 2013**  
University of Toronto  
Leslie Dan Pharmacy, Room 310

<table>
<thead>
<tr>
<th>Name</th>
<th>Position / Organization</th>
<th>Email</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Convenors</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Merrick Zwarenstein</td>
<td>Senior Scientist, Institute for Clinical Evaluative Sciences, Sunnybrook Health Sciences Centre</td>
<td><a href="mailto:merrick.zwarenstein@ices.on.ca">merrick.zwarenstein@ices.on.ca</a></td>
</tr>
<tr>
<td>2. Shaun Treweek</td>
<td>Chair in Health Services Research, University of Aberdeen</td>
<td><a href="mailto:streweek@mac.com">streweek@mac.com</a></td>
</tr>
<tr>
<td>3. Kirsty Loudon</td>
<td>2nd year Medical Research Council student</td>
<td><a href="mailto:k.loudon@dundee.ac.uk">k.loudon@dundee.ac.uk</a></td>
</tr>
<tr>
<td><strong>Participants (Toronto)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. David Moher</td>
<td>Senior Scientist, Clinical Epidemiology, Ottawa Hospital Research Institute; Associate Professor, Department of Epidemiology and Community Medicine, Faculty of Medicine, University of Ottawa</td>
<td><a href="mailto:dmoher@ohri.ca">dmoher@ohri.ca</a></td>
</tr>
<tr>
<td>5. Dan Riddle</td>
<td>Professor, Departments of Physical Therapy and Orthopaedic Surgery, Virginia Commonwealth University, Richmond</td>
<td><a href="mailto:dlriddle@vcu.edu">dlriddle@vcu.edu</a></td>
</tr>
<tr>
<td>6. Jerry Krishnan</td>
<td>Professor Medicine and Public Health, Associate Vice President for Population Health Sciences, University of Illinois Hospital and Health Sciences System, Chicago</td>
<td><a href="mailto:jakris@uic.edu">jakris@uic.edu</a></td>
</tr>
<tr>
<td>7. Vivian Welch</td>
<td>Deputy Director, Centre for Global Health, Institute of Population Health, University of Ottawa, Ontario</td>
<td><a href="mailto:Vivian.Welch@uottawa.ca">Vivian.Welch@uottawa.ca</a></td>
</tr>
<tr>
<td>8. Joel Gagnier</td>
<td>Assistant Professor, Department of Orthopaedic Surgery, Department of Epidemiology, University of Michigan, Michigan</td>
<td><a href="mailto:jgagnier@med.umich.edu">jgagnier@med.umich.edu</a></td>
</tr>
<tr>
<td>9. Kevin Thorpe</td>
<td>Biostatistician/Trialist,</td>
<td><a href="mailto:kevin.thorpe@utoronto.ca">kevin.thorpe@utoronto.ca</a></td>
</tr>
<tr>
<td>No.</td>
<td>Name</td>
<td>Position/Institution</td>
</tr>
<tr>
<td>-----</td>
<td>-----------------------------</td>
<td>--------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>10</td>
<td>An-Wen Chan</td>
<td>Assistant Professor, Associate SGS Member, W Women’s College Research Institute, Toronto</td>
</tr>
<tr>
<td>11</td>
<td>David Kent</td>
<td>Associate Professor, Medicine, Neurology, Clinical and Translational Science Director of the Clinical and Translational Science (CTS) Program General Internist, Tufts Medical Centre, Boston</td>
</tr>
<tr>
<td>12</td>
<td>Rob Fowler</td>
<td>Senior Scientist, Sunnybrook Health Sciences Centre, Toronto</td>
</tr>
<tr>
<td>13</td>
<td>Janet Martin</td>
<td>Assistant Professor, Schulich School of Medicine and Dentists, University of Western Ontario</td>
</tr>
<tr>
<td>14</td>
<td>Valeria Rac</td>
<td>Director, Clinical Research, Toronto Health Economics and Technology Assessment Collaborative (THETA)</td>
</tr>
<tr>
<td>15</td>
<td>Lusine Abrahamyan</td>
<td>Collaborator, Toronto Health Economics and Technology Assessment Collaborative (THETA)</td>
</tr>
<tr>
<td>16</td>
<td>Ba Pham</td>
<td>Decision Modeler, Toronto Health Economics and Technology Assessment Collaborative (THETA)</td>
</tr>
<tr>
<td>17</td>
<td>John Powers</td>
<td>Collaborative Clinical Research Branch, NIAID (National Institute of Allergy and Infectious Diseases)</td>
</tr>
<tr>
<td>18</td>
<td>Eric Johnson</td>
<td>Investigator, Kaiser Permanente Centre for Health Research, Portland, Oregon</td>
</tr>
<tr>
<td>19</td>
<td>Peter Selby</td>
<td>Chief, Addictions Division, Centre for Addiction and Mental Health (CAMH), Toronto</td>
</tr>
<tr>
<td>20</td>
<td>Noemi Lois</td>
<td>Ophthalmology Department, Grampian University Hospitals-NHS Trust, Aberdeen</td>
</tr>
<tr>
<td>21</td>
<td>Jodi B. Segal</td>
<td>Associate Professor, Medicine, Director, Pharmacoepidemiology Program, Co-Director, Hopkins DEcIDE Network of AHRQ, Co-Director, Hopkins AHRQ-NRSA Training Grant for Comparative Effectiveness Research, Associate Director, Johns Hopkins Evidence Based Practice Center</td>
</tr>
</tbody>
</table>
Highlights of topics discussed

Introduction

The reason participants were present indicated their interest in PRECIS and developing the PRECIS-2 tool. The group had a shared interest in pragmatic trials and that the tool could assist policymakers and trialists by making decision making more transparent and inform knowledge translation. The physiotherapist (DR) who had used the PRECIS tool with a multidisciplinary team, each with completely different perspectives, had found the tool assisted in highlighting areas for discussion to create consensus in designing a trial. A local policymaker with an issue that needed resolving could envisage
using PRECIS to assist in deciding what was the best treatment to use in the hospital to improve care in the hospital and to reduce costs.

1. Scoring

There was discussion about the necessity of scoring which was countered by the methodological research on PRECIS in which all studies had adapted the tool by using a score. This was an issue that could be user tested with agreement that a scale of 1 to 3 did not give much granularity with pragmatic, explanatory and equally pragmatic/explanatory. Using an odd score naturally created a midpoint so using scoring systems of 1 to 5 “less discrimination”, 1 to 7 and 1 to 9 were discussed with “20 being too close to the continuum to assist with bringing agreement to reliability”.

The issue of scoring raised the issue of the reason for using the tool, retrospective use as a decision aid and to highlight areas for research or prospective use to design clinical trials using PRECIS for the purpose it was intended. It was agreed that users could use PRECIS-2 as they wished but the new tool may assist users in both designing trials and in systematic reviews, the intention though was for the group to consider improving the tool for the purpose of designing trials. As one of the participants had stated he was interested in using PRECIS for observational studies this was another potential use. Scoring for prospective use, the actual inter-rate reliability agreement of PRECIS scores was of less importance than the use of the score to indicate disagreement over pragmatism of a PRECIS domain. Cumulative scoring was also mentioned but the limitations of this indicating the individual domain score barred further work on this during this project by KL.

2. Using PRECIS-2 wheel and a table to present rationale

The brainstorming group, after initial clear indication at the beginning of the meeting to approve of using PRECIS to assist with transparent decision making, really liked the table to indicate the rationale for decision making that was behind the visual appearance of the PRECIS wheel for an individual trial. “The table really operationalizes the score” and with regard to transparency “Others could disagree
with what we had done and our scoring but nevertheless we would have made the reasons for our scoring very clear.” Thus there was agreement that the PRECIS-2 table was useful and “it can improve reliability”.

3. Need for further guidance to use the tool – agreement that was needed in future publications.

4. Suggestions for domains to add

One of the key issues discussing extra domains was considering if there was a continuum from pragmatic to explanatory. Out of this grew a discussion on the match between the intention of the trialists and real world (the intended setting) which the policymaker (JM) stated would be “a giant leap forward if we could get clinical trialists to do that”. This lead to discussion on the generalisability of trial results which happened in real world settings. However the pulmonologist (JK) stated what this meant to him: “I’m a pulmonologist so to me what is pragmatic may or may not at all be pragmatic for a general practitioner working in a rural part of the city of Illinois – just ’cause they have very different resources – so I like to work applicability rather than generalisability first, because the question is “to whom are you trying to apply this evidence, and what setting are you trying to apply this evidence?...” and the word applicability of trial results was proposed. DK rounded up the discussion post lunch with reminding the group we had discussed “decision relevant” trials. However, KT the statistician thought of trials quite simply as explanatory “can it work?” and pragmatic “does it work?”

The issue of pragmatism was a constantly occurring theme throughout the brainstorming meeting. The uncertainty of a definition was highlighted by the example of a trial in a highly specialised group of Intensive Care Units (ITU) – with discussion by the group if a trial in this setting could be pragmatic. There was not consensus but the group’s thinking changed from considering this as explanatory to pragmatic IF applicability meant that the results were only intended to be applied in ITU units that had similar patients, setting, resources and personnel with similar expertise. The trial would therefore not be pragmatic if the results were considered for ANY intensive care units which were not similar. DK summed up the two ways the group was thinking of pragmatism: “decision relevant” or “broadly
applicable in public health”. MZ reminded everyone that Schwartz and Lellouch define pragmatic as “decision relevant”. There was agreement that this definition may be dependent on your viewpoint (implementing patient care or regulating drugs) and may contradict the idea of widespread applicability.

**Blinding** – there was a great deal of debate about this domain with mixed views on whether or not this should be an extra domain with “some” views that explanatory trials all had triple blinding and pragmatic trials did not. Whereas others thought all randomised controlled trials to ensure good internal validity should all have blinding. One viewpoint (JG) was that blinding was about bias so if included we would have to include other aspects of bias e.g. allocation. Others thought in usual care there was no blinding of patient, doctor or outcome assessment so that would represent the most pragmatic trial. Others thought outcome assessment should always be blinded so we find out “what is the effect? Not what do I think the expected effect is?”. The first author of the original PRECIS paper (KT) reminded the group that the NASCET trial [130] in the paper was explanatory but there was only blinding of outcome assessment, the patient and the doctor knew that the operation was taking place. KT stated “I don’t think that blinding is itself a defining characteristic of an explanatory or pragmatic trial”. The ethical issue of blinding was also mentioned as this does not happen in usual care and is one of the arguments for un-blinded treatment in pragmatic trials. One participant, the policymaker (JM) suggested “blinding itself is along the continuum but very much in the context of the question you are asking.” This topic was also discussed as requiring further guidance in the elaboration paper and not an additional domain for PRECIS.

**Setting** – discussion on this potential new domain encouraged participants to realise this domain is part of Participant eligibility criteria, and for the different sites in a multi-centre trial Experimental intervention practitioner expertise and Comparison intervention practitioner expertise. The discussion then moved back to considering if the setting for the trials was different to the intended setting for
the trial. And the word “match” was introduced by MZ considering extra resources. The NASCET trial was used again by MZ to demonstrate that not all the surgeons in one of the trial settings were eligible to perform in the trial, having an additional domain made that clearer. This did not change the domain scoring for Experimental intervention practitioner expertise. KT then went on to state “pragmatism is generalisability within a particular setting.” There appeared to be consensus that either additional information was added to guidance on using PRECIS-2 or this could be a new domain.

**Recruitment** - the methods of recruitment were touched on with discussion of additional resources that would change the people that would usually receive an intervention. This domain discussion was then absorbed into a discussion on matching and pragmatism mentioned previously.

**Cost effectiveness studies versus cost of an intervention which influences applicability** - There was discussion that there is cost effectiveness analysis in both explanatory and pragmatic trials thus there was no continuum. And there was also discussion that costs for interventions change e.g. anti-retroviral treatments in Africa. However, this then lead back to the example given by the specialist (pulmonologist JK) regarding the resources needed to operationalise an intervention and “did the trial require large amounts of infrastructure that were crucial for the execution of the intervention and would not exist in that setting when the trial disappears”. This was discussed as being part of the design characteristics of the trial. An example was given (JK) of the ARDSNET trial that had tested an intervention that made a difference in Intensive Care Units but there was no funding to maintain the equipment and number of specialised personal to continue to implement the intervention [131]. The systematic reviewer (VW) then suggested that each of the PRECIS domains had resources attached to it, e.g. monitoring, compliance, expertise and frequency of visit. In a recent review “some trials are very well resourced and they have this intensive follow up, monitoring.” And the policymaker (JM) pointed out to the orthopaedic surgeon (JG) “can’t do the trial in orthopaedics without the manufacturer providing the device for free which could cause me to grade it differently on the domain
of pragmatism for resources”. Discussion followed on changing domain descriptions or having a separate domain.

Analysis – The removal of this domain was discussed as both pragmatic and explanatory trials tend to be Intention to Treat (ITT). JK suggested that the important element was ensuring enough power for important sub-group analysis. KT suggested that in the original PRECIS paper explanatory trial analysis of the primary outcome meant doing more than ITT of the primary outcome. The discussion then opened up to pre-specification of subgroup analysis could still be pragmatic whereas post hoc would be explanatory and whether or not this assisted decision making. It was agreed that the original PRECIS paper difference between pragmatic and explanatory trials was good as an explanatory trial analysis created a “pure sample” of compliant, high-risk, homogenous patients through restricting ITT. There was uncertainty about whether or not this domain should remain so this would require further thought.

5. Weighting – This issue was touched on as weighting on a trial by trial basis but none of the group could see how this could be operationalised. With aggregate scores for systematic reviews (VW) was interested in this but otherwise there was not really any reason for this.

6. PRECIS-2 design - There was agreement that the PRECIS domains were arbitrarily arranged and could be grouped with Eligibility criteria coming at the top, 12 o’clock position. There was also discussion that there was an “E” in the middle but writing “Pragmatic” on the rim emphasised Pragmatic design and that was not the point of the tool so a single “P” may be better. The group had lots of thoughts on “spydergrams” and coxograms otherwise known as polar charts that Florence Nightingale used, with some participants of the brainstorming not liking the slope from one domain to another and using the axis for the domains and not a “segment”. KL mentioned Microsoft Excel and conditional formatting
could be used to create the PRECIS-2 wheel. This could combine neatly with using the table to detail rationale.

7. Use of different colours for each PRECIS-2 domain - Colours for the different domains were discussed with awareness by one of the participant (KT) that “certain colours look bigger than other colours”. There was agreement not to have PRECIS domains with separate colours as too expensive to publish and discriminates against people who are colour blind. There could, however, be potential to have different shades of the same colour for different scores.

8. PRECIS-2 website - One participant (JK) raised the issue of developing a training module and that (JM) “an online tool...would clinch the uptake of the tool.” In addition, the reliability of the domains could be improved by constantly testing the domains with a group of people to verify the definitions and revise until everyone was in agreement. JK had done this with the original PRECIS tool and suggested this technique would increase the applicability of PRECIS-2. Another (DR) suggested “four excellent examples in the original article really weren’t used” but the best way to increase “the usability and the reliability of the tool would be a training manual with generic concepts”.

The group was also interested in developing a database of pragmatic trials and that there should be an agreed definition of pragmatism to determine which trials were included. This was included in the original proposal for proposed work, along with the PRECIS-2 website. The trials included could “self-identify” but there was also discussion of including trials with a range of pragmatism to assist trialists in designing trials. However, participants agreed it may be difficult to select which ones do this best. However VW pointed out that some trials are maybe designed one way but are not actually “fit for purpose” and take a more explanatory or pragmatic direction which happens “by accident”. ST suggested the PRECIS tool may “make people think more carefully about justifying their decisions.” Other trials are created intentionally as “hybrids” partly explanatory and partly pragmatic. This may,
however, be part of testing an intervention that trials move from being more explanatory to more pragmatic by altering the domains on the PRECIS wheel. There were other suggestions for inclusion criteria: for example, trials with more than 1000 patients. However, suggestions needed to be screened and creating PRECIS wheels for a huge database of trials indicating trial pragmatism was not the intention of the thesis.

9. Future work - There was discussion around ensuring the CONSORT statement for pragmatic trials was updated in line with developing PRECIS-2 to ensure consistency. This lead to a proposal (DM) that the group consider collaboration with the SPIRIT protocol development group. MZ also proposed that this brainstorming group could be the core group for an “active engaged collaborative community of trialists thinking about how we do (design) our trials.”

Post meeting discussion

KL, MZ and ST discussed the content of the brainstorming meeting immediately after the meeting enabling face to face discussion to thrash out ideas for new domains and how to use the PRECIS-2 tool. The basis for this discussion was the flip chart summary and notes taken during the meeting. Using a Word copy of the original PRECIS elaboration paper the trio considered exactly what would be changed based on the discussion in the second brainstorming meeting. The core focus of the time discussing the tool was which domains to include in PRECIS-2, how to use the tool and how to define pragmatism and applicability.

The key decisions made by this group (numbered to match discussion in the Theta meeting) included:

1. Scoring would be tested in user testing with scales of 1-5 and 1-7.

2. Both wheel and table would be used as agreed in meeting.

3. Need for further guidance to use the tool – agreement that update would include more information on how to use PRECIS-2.
4. PRECIS-2 Domains - Discussion on domains started with the idea that matching intention to expectation could assist redefining the PRECIS domains for the new tool and not just for the “Setting” domain that was discussed in the Theta meeting with participants. The first key issue to agree on was which domains should be included and how they would be defined. The following issues are the summary of the discussion which lead to definitions of the new PRECIS-2 domains (Figure 10).

- **Blinding** was discussed and it was decided to update information on blinding in the guidance in the elaboration paper and not make this a separate domain.

- **Recruitment** was discussed and as the trio agreed this aspect of “who gets into a trial” was important this should become a separate domain that could be teased out from Participant eligibility criteria.

- **Setting** was discussed and it was decided this should be a separate domain and not just be incorporated as additional guidance.

- **Resources for trials** e.g. the ARDSNET low tidal volume trial intervention that had stopped when trial stopped and had been discussed in the Theta meeting as being included as domain that would “incorporate this concept of resource” became a new domain considering the Organisation of a trial which would include personnel, resources and organisation to implement the intervention and the comparison intervention.

- **Analysis** – Analysis of primary outcome - was kept in as a domain so there were 10 PRECIS-2 domains but there was agreement to include this in user testing to consider what users thought about including AND if there were additional restrictions that could be suggested to differentiate between more pragmatic and explanatory trials.

The ten 10 PRECIS-2 domains that arose from discussion of the Theta brainstorming meeting – based on “the similarity between X and that what was likely in the setting to which the results will be applied” (Table 23). Prior to determining the detail of the new domains, the process for how to use the tool was discussed to create six steps (Box 5.8 Third Brainstorming meeting – presented to participants).
1. The similarity between the eligibility criteria to be met by recipients of the intervention and comparator and the likely selection criteria that would need to be met by care recipients in the setting to which the results will be applied.
2. The similarity between the recruitment path for recipients of the intervention and comparator and the likely method of identifying care recipients in the setting to which the results will be applied.
3. The similarity between the setting of the trial and the setting to which the results will be applied.
4. The similarity between the resources, provider expertise and the organisation of care delivery, in the intervention arm of the trial and those likely in the setting to which the results will be applied.
5. The similarity between the resources, provider expertise and the organisation of care in the comparison arm of the trial and those likely in the setting to which the results will be applied.
6. The similarity between the flexibility in how to deliver the intervention to recipients and the likely flexibility of how care will be delivered in the setting to which the results will be applied.
7. The similarity between the flexibility in how intervention recipients must engage with the intervention and the likely flexibility of engagement allowed for care recipients in the setting to which the results will be applied.
8. The similarity between the intensity of measurement and follow-up of trial participants and the likely intensity of measurement and follow-up in the setting to which the results will be applied.
9. The direct relevance of the trial's primary outcome to recipients of the intervention and comparator.
10. The comprehensiveness of inclusion of trial participants in the analysis of the primary outcome.

The elaboration paper was then edited by discussing in detail the new domains, Recruitment, Setting, Organisation of the Intervention and Organisation of the Comparison Intervention with each domain being developed and restrictions considered to make the trial move from being more explanatory to more pragmatic (just as in the original PRECIS paper).

In addition, the following issues, numbered to match discussion in the Theta meeting were also agreed on by the trio.

5. **Weighting** would not be discussed further in PRECIS-2 development.

6. **PRECIS-2 Design** – Agreement to change order, group domains more logically and test out spydergrams and coxograms in user testing.

7. **Colours** – this would be limited to one or two to ease copying information.

8. **Database** work starting shortly after return with work on website summer 2013.

9. **Use SPIRIT database** for trial protocols for validity and reliability work.
Discussion
The initial invitations were sent on individually so that potential participants would not be influenced by who was on the invite list. It was only at a later stage when arrangements for the meeting were being finalised that the participant list (Table 4.5) was sent out along with the agenda. There was active participation at the meeting by nearly all of the invitees but two of the less experienced early career invitees (suggested by the Director hosting the meeting) acted more as observers.

It is worth noting that although there was no funding to attend, five Americans flew to Toronto to participate. One, a physiotherapist, was the first to publish methodological research on PRECIS in 2010 [53] using the tool as it was intended, to design a trial on total knee surgery. Another Professor of Medicine and Public Health at the University of Chicago (JK) had recently published two articles on Comparative Effectiveness research [132, 133] while another participant (JG), an orthopaedic surgeon from Michigan, had co-authored the CONSORT extension for pragmatic trials and two highly cited articles on pragmatic trials published about the same time as the original PRECIS article [34, 134] articles).

An Associate professor from Tufts (DK), Boston had also flown up to join the meeting after writing a discussion paper on the dangers of extrapolating the results of pragmatic trials [135]. Two other participants from the US online had published articles citing PRECIS [51, 136]. The Canadians had a representative who was heavily involved in the Cochrane Collaboration and Deputy Director of the Centre for Global health cited PRECIS [137]. Another local academic primary care physician had undertaken a review of the methodological literature on PRECIS and had used the PRECIS tool to evaluate a smoking cessation trial [40]. So to conclude the brainstorming meeting had a diverse group of participants who all had an interest in pragmatic trials and the development of the tool as demonstrated by their enthusiasm to travel long distances without payment to attend.
Meeting content

This discussion around the intended purpose of PRECIS, prospective to retrospective use of the tool created the idea to use the publically available SPIRIT 150 protocol database to test the validity and reliability of PRECIS-2. This database had been constructed during a study to consider the variability of trial protocols but provided a ready-made, publicly accessible group of trial protocols that would enable retrospective use of PRECIS but using the tool in a similar way to how it would be used in designing new trials.

There was a lot of interest in creating a PRECIS-2 Website to increase dissemination of the tool and provide teaching materials to increase reliability for trialists using the PRECIS-2 tool. The development of “a training manual with generic concepts” was immediately taken on board by the trio developing PRECIS-2, with the idea to use the elaboration paper for this purpose to assist users. The concept of the PRECIS-2 website including a pragmatic trial database containing trials that take a pragmatic approach to design was also enthusiastically received. Participants in particular suggested including in the database, trials that started out being more pragmatic or explanatory and changed direction for one reason or another; these trials might provide useful examples to trialists intending on using the PRECIS-2 tool to study before using the tool to design their own trial.

The innovative element that emerged from the brainstorming was the idea of including in PRECIS-2 an assessment of the additional resources required for a trial intervention to be tested. This followed discussion of a successful trial (Acute Respiratory Syndrome Network - ARDSNET) that had failed in implementation post trial due to lack of resources [131]. This lead to the development of the Organisation domain in the meeting post Theta. Concurrently, unknown to the developers of PRECIS-2, a checklist was developed by Dorling to improve the quality of trial proposals to the National Institute for Health Research (NIHR) which included workforce, organisation and location to consider the generalisability of the research [138].
At the same time as there was discussion about PRECIS-2 domains, there was also reflection on what pragmatic meant; if an intervention was tested out in a specialised area but that was only the area it would ever be needed that would be usual care for this intervention and thus by inference it would be pragmatic. This was in contrast, to the more generally accepted idea that pragmatic meant usual care that would happen anywhere a patient needed treatment. These two ideas needed some thought by the participants and the steering group later in considering applicability, generalisability and external validity [21, 45, 139].

We believed that the work on PRECIS-2 development would encourage trialists to consider the external validity of their results and consider the applicability to patients out-with the study participants – as visually indicated by the PRECIS-2 wheel. If a small wheel with the rim closer to the centre, then trial results were less applicable than a trial that had domains that had scored 5 and was on the outer rim of the PRECIS-2 wheel. PRECIS-2 could assist in making trial decision making more transparent. Generalisability (as discussed by Dekkers [139]) can only occur if the results are externally valid or applicable to a specific target population and may then be generalised to a different group of people. Whereas, applicability is often considered by individuals looking at trial results;” is this applicable to me?” Or healthcare professionals asking “is this trial applicable to my care of a specific patient?” Thus applicability is more specific and narrowing down to a personal perspective while generalisability may be extending the external validity of a trial. Both are hugely dependent on the characteristics included in our tool to assist trialists consider the external validity of results. There is however, some overlap between the terms so there is no clear cut definition.

Considering the visual display of the PRECIS-2 domains, the first author of the original PRECIS paper suggested a recent publication that gave further information about spydergrams [140] giving backing to the idea that the PRECIS wheel is visually very attractive and “Shape is a very powerful visual signal
that immediately catches the eye” going on to say that “Comparing different forms can be very powerful, if the reader is trained to understand where the dimensions are plotted”. This highlights that with experience “at a glance” differences between trial designs can be observed.

**Study strengths and limitations of the second brainstorming meeting**

The key ingredient in a brainstorming meeting is the participants; this group interacted very well together and the atmosphere in the room was of excited enthusiasm and was maintained throughout the day. There was humour and the transcript indicates there was laughter during the meeting (with wise cracks like “I used to be indecisive but now I’m not so sure” during the afternoon session). The participants bounced ideas around and spontaneously reacted to each other thoughts with genuine interest in the topic and mutual respect.

A limitation to the meeting may have been the composition of the participants who attended when compared to those that did not – 60% of those invited did not participate with 23% not replying to the initial invitation or reminders. Those not attending on the day may have been due to realisation of time differences for the two invitees in the UK (10.00GMT) or being uncertain about the benefits for individual participation. As those present included policymakers, users of the tool, clinicians and some of the original authors of the PRECIS publication we do not believe this is of great concern.

This study might have been limited if there had not been a team of three (KL, MZ, ST) ensuring that the meeting flowed. Each of the team coordinating the meeting had their roles, predefined before the meeting but every member played their part. A brainstorming meeting like this would not have produced such exciting input without careful coordination to ensure the participants kept to their remit and finished off discussion of ideas that arose from a topic.
**Conclusion**

The second brainstorming meeting with participants was very constructive. Everyone gained from the meeting, learning from each other and giving their own perspective on improving PRECIS to assist in decision making for healthcare. There was definite agreement on using scoring to improve the tool, further guidance on using PRECIS to assist users and the aim of the tool. The follow up meeting with KL, MZ and ST was equally intensive with the production of PRECIS-2 with ten domains and a draft of six steps to assist trialists using the tool.

**Third brainstorming meeting**

**Aims**

To discuss possible PRECIS-2 model based on Delphi and 2nd Brainstorming meeting in Toronto March 2013

**Objectives**

The aim of the third brainstorming meeting was achieved through a number of objectives:

1) Discuss results from second brainstorming meeting in Toronto and subsequent work.

2) Determine what participants thought about the models for the next stage of user testing.

**Methods**

**Participants**

On 12th April KL invited (Appendix Chapter 4 Box 4.6) four participants to the informal brainstorming meeting in Dundee (Table 24). We did not widen our inclusion criteria and invite participants outside the original invitees to the first brainstorming meeting. Neither did we re-invite people who had previously not replied.
Table 24 Participants invited to participate in third brainstorming meeting in Dundee

<table>
<thead>
<tr>
<th>Participants</th>
<th>Job description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peter Donnan</td>
<td>Professor, Co-Director TCTU, statistician, methodologist, trialist (co-applicant CSO grant)</td>
</tr>
<tr>
<td>Roberta Littleford</td>
<td>Senior Trial manager, TCTU, trialist</td>
</tr>
<tr>
<td>Thomas Lamont</td>
<td>Dentist and Trialist on HTA trial (IQuaD – Improving the Quality of Dentistry) (Suggested by Heather Cassie)</td>
</tr>
<tr>
<td>Frank Sullivan</td>
<td>Professor - Director Population Health Sciences</td>
</tr>
</tbody>
</table>

Facilitators
The facilitator for the brainstorming meeting was KL, assisted by ST. FS (supervisor) would be a participant.

Pre-meeting preparation
Prior to the meeting KL prepared a 12 page PowerPoint presentation to guide the meeting. Topics covered included:

- Who is PRECIS-2 for?
- PRECIS-2 model with 10 spokes and domains names (Figure 11 PRECIS 2 model post Brainstorming meeting in Toronto March 2013Figure 11)
- Training instructions for PRECIS-2 including 5 steps
- Main PRECIS-2 difference
- PRECIS-2 DOMAIN differences
- Potential PRECIS-2 model to test indicating scoring 1-5 or 1-7.

A 4-page hand-out entitled “Information sheet – the PRECIS tool: designing trials that are fit for purpose” was also prepared for each participant to look at during the discussion. This contained “Advice on using PRECIS-2”, “The PRECIS-2 domains” – 10 domains with one sentence descriptions and finally details of three of “The Domains in detail” entitled “Participant Eligibility Criteria similarities”, “Recruitment path similarities” and “Setting similarities”.
No preparation prior to the meeting was expected of the participants; we wanted to get feedback from their first impression on PRECIS-2 so this meeting would act as a bridge to user testing which was due to start soon afterwards.

Protocol for the meeting

An agenda was prepared and followed as detailed below. The topic guide was the PowerPoint presentation; this was used to guide a free flowing discussion of the content. In addition, notes were taken by KL and ST to ensure maximum information on views were gathered from the meeting for post meeting analysis.

Agenda 29th April 2013

1. Introductions – facilitator and participants
2. PRECIS-2 development
3. Instructions on how to use PRECIS-2
4. PRECIS-2 proposed domains (Figure 11, Table 23)
5. Scoring
6. Presentation of PRECIS – spokes or segments?
7. Any other business
8. Close
Figure 11 PRECIS 2 model post Brainstorming meeting in Toronto March 2013
Reminders

Invitations to the meeting were resent on Friday 26th April for the meeting on Monday 29th April (Appendix for Chapter 5, Box 5.7) reminding participants of the lunch meeting.

Incentives

There were no incentives to attend other than lunch. To improve the likelihood of attendance the participants had been offered two meeting times.

Post meeting analysis

Notes, prepared by KL and supplemented by ST, were used to summarise the third brainstorming meeting discussion. In particular, participants’ feedback on the changes in PRECIS-2, the domains and scoring were analysed. This information was then used to prepare for the next stage of improving the PRECIS tool through User testing.

Results

All four participants accepted the invitation to attend the meeting – TL and PD had been present at the first brainstorming meeting, FS had been part of the process discussing the development of the tool but had submitted an apology for not being able to attend the first meeting. RL had been invited to attend the first meeting but unable to attend.

The discussion centred on: 1) the domains of the PRECIS-2 tool (Figure 11, Table 23) – in particular the Recruitment domain using trial examples participants were involved in 2) how trial designs can change over time; 3) scoring using polar charts or coxcombs rather than dividing the spokes of the wheel, and 4) future monitoring of the use of PRECIS and PRECIS-2.

1. PRECIS 2 Domains: There was discussion about domain overlaps but generally all present at the third brainstorming meeting thought we should keep all ten proposed domains for PRECIS-2. By including these domains participants agreed we were being explicit about the issues to discuss in
designing a clinical trial. Thus, if for instance, we removed the new domain *Setting* then it would be harder for trial designers to remember to include all the necessary details in the new domain of *Organisation*.

The participants went on to use examples of their own trial experience to consider the proposed PRECIS-2 domains and the restrictions or factors that we had included to make a trial more explanatory from the starting point of very pragmatic (usual care). The *Recruitment* domain was new to PRECIS-2 and received most comment. Comments are summarised below:

a) *Recruitment* domain: Incentives in *Recruitment* might also be part of a complex intervention trial and not just under the *Recruitment* domain so there may be overlap.

b) *Recruitment* Incentives: If do not use incentives then e.g. might take eight years to recruit through normal clinics and trialists do not have time for that – funders have specific time period for trial – so trialists must augment usual practice to recruit adequate sample size.

c) Example (1) NATTINA (NAtional Trial of Tonsillectomy IN Adults) NATTINA is a multicentre, randomised, controlled trial with feasibility study and internal pilot. NATTINA is looking to determine the cost effectiveness and efficiency of tonsillectomy versus conservative therapy (delayed surgery) in adults with recurrent acute tonsillitis over a 24 month follow up period. – [http://www.ncl.ac.uk/nctu/research/project/4998](http://www.ncl.ac.uk/nctu/research/project/4998) As adults with tonsillitis keen to increase their chance of treatment, recruitment has been more than five times what would normally be needed to recruit through usual referral procedures as patients self-referring.

d) Example (2) IQuaD (A randomised controlled trial comparing oral hygiene advice and periodontal instrumentation for the prevention and management of periodontal disease in dentate adults attending dental primary care) has no problems recruiting as routine (current practice) or theory-
based personalised (to the needs of the patient) oral health advice (OHA)

http://www.abdn.ac.uk/hsru/research/assessment/orthopaedics/iquad/

e) (Example (3) Bell’s Palsy trial [141] - early treatment with prednisolone or acyclovir for Bell's palsy. This trial involved GPs in primary care/A&E but patients were treated in hospitals in secondary care. In usual practice normally treatment in primary care only more serious patients end up in secondary care being treated by specialist neurologists.

f) Example (4) Early Cancer Detection Trial (ECLS) (Blood test - an antibody test which can detect cancer at the microscopic level before it becomes a radiological cancer.) http://www.eclsstudy.org/early-lung-cancer-test The Universities of Dundee and Glasgow, NHS Tayside and NHS Greater Glasgow and Clyde - are hoping to recruit 10,000 people from these areas, mainly recruiting by identifying potentially eligible individuals from GP medical records and through postal invitation letter including a summary of the study and Participant Information sheet (PIS)

g) FS key question for using trial results “Are study participants like my patients?” PRECIS-2 helps design trials that enable trialists to be more conscious of this question for the domain Eligibility Criteria.

h) With regard to the Analysis domain, participants believed should keep in PRECIS-2 tool so complete process. PD suggested could put in subgroup per protocol analysis which would still make trial pragmatic as long as always have Intention to Treat (ITT).

1. Trial design changes over time. Following discussion of trial examples and trial domains, the participants moved on to discuss the changes from initial study design to final trial design. They suggested there will always be modifications to trials as a trial is rolled out so the domains and wheel in PRECIS-2 will change. There is always uncertainty in exactly how different aspect of a
trial will function in practice, participants suggested considered publishing initial PRECIS-2 wheel in protocol, then in post-trial publication could have final PRECIS-2. There could perhaps be a history of the PRECIS-2 wheel due to different decisions for the trial. But it is important to know the reasons for trial design and why a trial has been designed in a particular way, PRECIS-2 could help make explicit these reasons and help the user of the result match trial design with the setting in which results will be applied – thus making trial decisions more overt.

2. **Domains, spokes or segments?** There was discussion around how to use a scoring system in PRECIS-2. The original tool had no scoring system but in its development scoring had been used on the “spokes” and there had been published use of scoring systems. An alternative way of scoring the 10 domains on PRECIS-2 was suggested using Coxcomb diagrams or polar charts. It was suggested that scoring using this system may be easier to visualise and better to have segments for whole domain rather than spokes on a wheel which went in and out to different domain “spokes”. However, problems with colour interpretation if using different colours for different scores so suggested perhaps best if have grey as neutral colour for segments. This would also help printing out PRECIS-2 wheel if black and white printer and this system was used.

3. **For the future** – think about how PRECIS use could be monitored if use published in more General Medical Journals or more specialised medical journals.

**Discussion**
The participants invited all accepted the invitation and actively took part in brainstorming, expressing their appreciation for being invited to participate in this research post-meeting. All were positive about the development of PRECIS-2 in assisting in trial design decision-making. Following the discussion of the principles of PRECIS-2 and the ten proposed domains the participants’ dialogue opened up into other areas that were not on the agenda but were relevant and were included in “Any other business”.

170
For instance, the topics of “Trial design changes over time” and “Future monitoring of PRECIS-2 use” were helpful to the overall project but not specifically on the agenda. This demonstrated that participants were embracing the overall topic of how PRECIS-2 could be used in trial design and there was “brainstorming” so participants felt able to propose ideas for PRECIS-2 through free flowing discussion between participants, building on each other’s suggestions [142]. There did not appear to be blocking of ideas from one participant or particularly dominant personalities that preventing any of the participants from freely expressing their opinions [143]. With regard to using scoring systems the participants were unsure if using scores on the spokes would be the clearest way of showing the domain scores and suggested testing out this as well as user testing of polar or coxcomb diagrams.

The participants in the groups also came up with a project that could be pursued post-PhD to keep track of how PRECIS and PRECIS-2 are used. The group suggested that it may be useful to consider dissemination of the trial design tool and which journals had published methodological articles on PRECIS. Specifically, were articles published in more general medical journals or specialist journals? So far, (Table 25), the journals that have published methodological articles include: Trials (2), J Clin Epidemiol (2), Int J Tuberc Lung Dis (1), PLoS ONE (1), BMC Med Res Methodol (1), Health Serv Res (1), Diaglogues Clin Neurosci (1), Eval Health Prof (1), Implement Sc (1), Transl Behav Med (1), J Public Health (Oxf) (1), American Journal of Preventive Medicine (1) and The Journal of the American Board of Family Medicine (1).
### Table 25 Journals that have published methodological work on PRECIS

<table>
<thead>
<tr>
<th>Reference</th>
<th>Reference</th>
<th>Journal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>Reference</td>
<td>Journal</td>
</tr>
<tr>
<td>-----------</td>
<td>-----------</td>
<td>---------</td>
</tr>
<tr>
<td><strong>Harden (2013)</strong></td>
<td>Harden SM, Burke SM, Haile AM, Estabrooks PA: Generalizing the Findings From Group Dynamics-Based Physical Activity Research to Practice Settings: What Do We Know? Eval Health Prof 2013.</td>
<td>Eval Health Prof</td>
</tr>
</tbody>
</table>
Study strengths and limitations of the third brainstorming meeting
There were a small number of participants at the meeting, there was only one newcomer to PRECIS (RL) but she was aware of the PRECIS tool. The small number of participants with only one researcher (TL) who did not know one of the participants (RL) meant that there was an open informal discussion with free flow of questions and views. Additional participants would have brought an extra dimension to the discussion and it is possible that by not having more participants who did not have a vested interest in the outcome of the meeting, we could have had a richer output. The other participants who were unable to attend had responded with apologies stating they wished they could attend the first brainstorming meeting (AA, BG, JC, MJ). The meeting could have been rescheduled perhaps (and it would have been had this been the only form of consultation) but as it was one of several ways in which we sought the input of others, the meeting went ahead, smaller than planned but still useful. The PRECIS tool depends on the interaction between health care professionals and trialists and it is important to have both. The creation of PRECIS-2, however, is more abstract and there is sometimes a disconnect between what fellow researchers thought we were trying to do and the real world, with the result it was difficult to recruit locally to assist with the PRECIS project.

In analysing this third brainstorming meeting, we did not record the meeting in April 2013. We therefore did not use a transcript as the basis for determining the content of the meeting. Analysis was simply constructed from the notes taken based on the direction and content of the discussion of the four participants. It is unlikely that key messages were missed as three of the steering group (KL, ST, FS) were involved in the PRECIS project this is unlikely and follow up questions to FS and RL, post meeting, were responded to by these participants.

Conclusion
This third and final brainstorming session led onto user testing of the ideas presented to this group of participants (Chapter 5) – in particular scoring presentation was added to User testing. All the domains discussed and changes worked through brainstorming following the Delphi were kept for testing. This
brainstorming session also suggested further work that could be considered to monitor PRECIS-2 use and determine if the tool was percolating down to evidence users and health care practitioners and was being published in general medicine applied healthcare journals.

<table>
<thead>
<tr>
<th>Summary of Brainstorming chapter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Three separate brainstorming meetings were undertaken at different times to inform the next stage of work to improve the original PRECIS tool: (1) at the beginning of the project prior to the Delphi process (Dundee), (2) after the Delphi (Toronto) and (3) prior to User testing (Dundee).</td>
</tr>
</tbody>
</table>

PRECIS ideas for improvement included design, scales, and additional domains developed from systematic review of published work using PRECIS and the results from the Delphi. Ideas for PRECIS-2 models for User Testing were also tested out and there was positive feedback from the training package under development to introduce new users to PRECIS-2.

**Conclusion**

The brainstorming meetings were successful in informing the next stages in the improvement of the original PRECIS tool and the creation of PRECIS-2 models for user testing.
Chapter 6: User testing

Introduction

User testing of candidates for PRECIS-2 occurred after the Delphi survey of authors citing PRECIS and post brainstorming. User testing is based on product design methodology and is a useful methodology for testing out use of “products” or tools on a one to one basis through spontaneous feedback. We used this phase to refine the PRECIS-2 model prior to validity and reliability testing using experienced trialists and methodologists and a selection of protocols.

Aims and Objectives

Aims

To produce an improved version of the PRECIS tool called PRECIS-2 that can be tested for validity and reliability.

Objectives

1. Are the instructions for using PRECIS-2 clear for users?
2. Which domains will we keep, remove or add to PRECIS?
3. What wording will we use for domain descriptions to ensure that they are easily understood?
4. Have we included all restrictions that would shift a domain from pragmatic to explanatory?
5. What scoring system will we use for a modified PRECIS tool?
6. What format of the PRECIS-2 wheel will be used, spydergram (original) or polar chart?
7. How shall we present PRECIS-2? Wheel or table or both?
8. Is the PRECIS-2 tool intuitive to users when used to assist in discussing trial design?
Methods

Participants

A total of 18 people indicated in the Delphi survey that they would be interested in participating in user testing. Invitations were sent on 14th May with follow up 21st June 2013. In addition KL purposively selected testers so that there was a balanced international input from a broad selection of potential users of the PRECIS-2 tool. Through the Delphi we had received input from international trialists so KL was keen to recruit further input from local and UK trialists. ST, FS made suggestions for face to face meetings (as well as one trialist DW from Nottingham university ST thought might be interested in PRECIS-2). These trialists were invited to contribute to user testing in stages to ensure we had adequate numbers of participants for user testing PRECIS-2 – two were contacted on 7th May and two on 14th May, four on 21st June, two on 30th July and one on 1st August 2013.

Ideally in user testing 15 to 20 people are used to ensure the sample size is adequate. Twenty is often needed to gain adequate information for a “product” to test and improve the design to make it “user friendly”.

Methodology for user testing

To determine the final PRECIS-2 model that would be used in validity and reliability testing we wanted to test out the ideas collected so far through user testing. User-testing candidate PRECIS-2 models on a one-to-one basis involved face to face interviews or using Skype and sharing the PRECIS-2 user testing document [98, 129]. The user (e.g. a trialist) was presented with PRECIS-2, including the instructions for using the tool and a series of questions and tasks applying the tool to a specific trial. The PRECIS-2 domain descriptions and restrictions to classify a trial from very pragmatic to very explanatory were also user tested. With participants’ permission, user-tests were audio recorded and KL also took notes.
All recordings were deleted once transcribed. Both ST and KL had experience of this methodology from past and current EC FP7 projects (e.g. http://www.decide-collaboration.eu/).

Prior to user testing the materials were prepared: an information sheet for participants (Appendix Chapter 6, Box 6.1), a consent form for participants ((Appendix Chapter 6, Box 6.2) and the topic guide for PRECIS-2 ((Appendix Chapter 6, Box 6.3). The topic guide contained themes highlighted in the Delphi and brainstorming. This was modified after each user testing with a participant or in groups if several interviews were conducted in a short space of time. Trial examples in primary care ([124], surgery [120] were initially used to ask participants questions about different PRECIS-2 domains and assist understanding.

The process of user testing was pilot tested. As it was successful and resulted in no changes to the process, this pilot was included in the overall user testing. KL, ST reviewed notes and transcriptions together after each test, or after several tests, looking for barriers and facilitators to the use of PRECIS 2, categorised according to the severity of the problem: high (critical errors such as incorrect interpretation or high degree of uncertainty or dissatisfaction), medium (much frustration or unnecessarily slow use), and low (minor or cosmetic problems). Notes were taken on changes since last user testing; the rationale for changes; the domains discussed to ensure all domains and restrictions were tested; scoring preferences e.g. 1 to 5, 1 to 7; design of PRECIS-2 using a coxogram or spydergram (i.e. as the original tool); advice on using PRECIS-2; difficulty in using PRECIS-2 and feedback on improving user testing.

The version of PRECIS 2 that finally emerged from iterative user-testing was then tested for validity and reliability.
Results

Participants

User-testing of PRECIS-2 candidates involved an international group of 19 researchers and trialists (Australia, Brazil, Canada, Germany, Greece, Italy, Norway, UK, USA) from 7th May until 3rd September 2013. Twelve of the user testers were involved in the PRECIS-2 Delphi survey (Table 26) and were already aware of PRECIS. We also invited comment from others who were unfamiliar with PRECIS or PRECIS-2 (Table 27). Early career researchers as well as experienced researchers gave feedback. KL also tested PRECIS-2 with six researchers in Dundee and Aberdeen starting with pilot testing for the User testing in Aberdeen. Nine out of the 29 invited (31%) did not participate in user testing having agreed to do so, six were too busy and testing was also over the summer holiday period so this might have made it harder (Table 6.3). One of the invitations was sent to a visiting German GP (IG) KL decided to involve this participant in the pilot for the Validity and Reliability testing (see chapter 8).

Testing took from 30 minutes to an hour based on the amount of time available to the participants. Twelve participants used Skype and sharing the document for user testing and five were involved in face to face testing. Two participants with faulty Skype connections had the user testing document sent to them and user testing was undertaken on the telephone in the USA and Greece. Two additional participants started on Skype but telephone was also used to finish off testing in Sheffield and Leeds.
Table 26 Twelve Participants in User testing who had contributed in the Delphi on PRECIS-2

<table>
<thead>
<tr>
<th>No</th>
<th>Name</th>
<th>Country</th>
<th>Research Area</th>
<th>Further phases</th>
<th>No trials involved in</th>
<th>Answer to question 1 in the Delphi “Have you used PRECIS tool to design a clinical trial?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Franz Portzsolt</td>
<td>Germany</td>
<td>Haematologist/oncologist trialist, 40 years in trials - 20 to 50 trials</td>
<td>2\textsuperscript{nd} round, User testing</td>
<td>40 years in trials: 20-50</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>Peter Hall</td>
<td>UK</td>
<td>Oncologist, design and running of oncology clinical trials, economic analysis</td>
<td>User testing</td>
<td>4 trials total: 2 trials health economics, 1 Phase 2, 1 Phase 3</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>Gianni Virgili</td>
<td>Italy</td>
<td>Ophthalmologist, systematic reviewer, hospital care management</td>
<td>User testing</td>
<td>Systematic reviewer</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>Vilemine Carayanni</td>
<td>Greece</td>
<td>Design trials considering economic evaluation</td>
<td>User testing</td>
<td>2 – economic evaluation</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>Christian Gold</td>
<td>Norway</td>
<td>Music Therapist, trialist</td>
<td>User testing</td>
<td>About 8 trials some quasi RCT</td>
<td>Yes - I have used it as an idea, not very systematically. But it has been helpful both in guiding decisions and in justifying them.</td>
</tr>
<tr>
<td>No</td>
<td>Name</td>
<td>Country</td>
<td>Research Area</td>
<td>Further phases</td>
<td>No trials involved in</td>
<td>Answer to question 1 in the Delphi “Have you used PRECIS tool to design a clinical trial?</td>
</tr>
<tr>
<td>----</td>
<td>--------------------</td>
<td>----------</td>
<td>---------------------------------------------------</td>
<td>---------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>6</td>
<td>Eric Brass</td>
<td>USA</td>
<td>MD, Trialist, Director Centre for Clinical Pharmacology</td>
<td>1(^{st}) and 2(^{nd}) round, user testing, Theta Brainstorming, validity testing</td>
<td>12 if consider large clinical trials</td>
<td>No</td>
</tr>
<tr>
<td>7</td>
<td>Gisele Huf</td>
<td>Brazil</td>
<td>Medical doctor, trialist</td>
<td>1(^{st}) and 2(^{nd}) round, user testing</td>
<td>5 trials</td>
<td>No</td>
</tr>
<tr>
<td>8</td>
<td>Louis D Fiore</td>
<td>USA</td>
<td>MD, internal medicine, epidemiologist, trialist</td>
<td>1(^{st}) and 2(^{nd}) round, User testing</td>
<td>About 50 trials</td>
<td>No</td>
</tr>
<tr>
<td>9</td>
<td>Clare Relton</td>
<td>UK</td>
<td>MD, trialist</td>
<td>1(^{st}) and 2(^{nd}) round, user testing</td>
<td>4/5 homeopathy trials</td>
<td>No</td>
</tr>
<tr>
<td>10</td>
<td>Jodi Segal</td>
<td>USA</td>
<td>MD, Trialist, lecturer</td>
<td>2(^{nd}) round, User testing, validity testing</td>
<td>No trials, 9 observational studies, systematic reviews</td>
<td>No</td>
</tr>
<tr>
<td>11</td>
<td>Rodger Kessler</td>
<td>USA</td>
<td>MD, psychologist</td>
<td>1(^{st}) round, user testing</td>
<td>10-15 trials</td>
<td>Anonymous</td>
</tr>
<tr>
<td>12</td>
<td>Grace Thompson*</td>
<td>Australia</td>
<td>Early career researcher, Lecturer, Music Therapy</td>
<td>1(^{st}), 2(^{nd}) round, user testing</td>
<td>2 trials</td>
<td>Anonymous first round</td>
</tr>
</tbody>
</table>

* Grace Thompson was invited 21st June - the other participants were all invited 14th May with reminders 21st June.
<table>
<thead>
<tr>
<th>No according to time invited</th>
<th>Name</th>
<th>Research area</th>
<th>No trials involved in</th>
<th>Heard of PRECIS</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>Graeme MacLennan</td>
<td>Senior Statistician</td>
<td>10</td>
<td>Yes, originally at PRACTIHC (The Trial Protocol Tool) meeting in South Africa over 10 years ago.</td>
<td>ST/HSRU- U of Aberdeen</td>
</tr>
<tr>
<td>23</td>
<td>Diane Witham</td>
<td>Research manager Trialist</td>
<td>20 +</td>
<td>Used PRECIS in trial published in TRAILS journal with Daniel Bratton. Tend to use informally now in head but don’t use formally with trial team.</td>
<td>ST/Nottingham trials centre</td>
</tr>
<tr>
<td>24</td>
<td>Miles Witham</td>
<td>MD Cardiovascular Diabetes medicine, Elderly care - trialist</td>
<td>19</td>
<td>Read paper -don’t use tool but concepts underlying designing trials familiar.</td>
<td>ST/Uo Dundee</td>
</tr>
<tr>
<td>26</td>
<td>Vishnu Modhok</td>
<td>GP – trialist dermatology</td>
<td>2</td>
<td>Presentation by KL</td>
<td>KL asked - U o Dundee</td>
</tr>
<tr>
<td>27</td>
<td>Fiona Williams</td>
<td>Developmental epidemiology, MPH course director</td>
<td>1</td>
<td>No</td>
<td>KL asked directly /U o Dundee</td>
</tr>
<tr>
<td>28</td>
<td>Blair Smith</td>
<td>GP – Pain specialist - Trialist</td>
<td>5 (involved in more but not specifically designed)</td>
<td>Presentation by KL</td>
<td>KL asked directly /U o Dundee</td>
</tr>
</tbody>
</table>
Difficulty in using PRECIS-2

Most of the users (11) found PRECIS-2 relatively easy to use (Figure 12) with scores of “0 to 3” for difficulty on a scale of 0=very easy to 10=very difficult. Many believed, however, that practitioners might find it harder to use if they seriously thought about the domains and restrictions when designing the trial. This was emphasised by two of the participants in user testing who scored difficulty as “5” – “If PRECIS not hard to use then useless, creating useful resource to think critically hard to use as a tool.”

Three who found it difficult particularly emphasised that “difficult to use PRECIS alone...qualitative items hard if systematic reviewer. Don’t have the support of a professional team designing the trial.”

And a junior researcher who used PRECIS in thesis stated “thought provoking and challenging, should be challenging and make you think about what you are doing in trial design. If too easy then not giving enough thought, easy as can see where it is going and understand where it is going. As tool makes sense, clear, how difficult it is to apply it is challenging.” Three others said they were unable to answer, they would not use PRECIS-2.

![Figure 12 Difficulty in using PRECIS-2 for User testers](image-url)
Advice on how to use PRECIS-2

The advice on “how to use PRECIS-2” developed from being a six step iterative process to a four step iterative process half way through user testing (Model 7 out of 13 models). The words used were critical and it was clear that the definition of pragmatic and explanatory mean different things to the different users. It is impossible to “please everybody” but the point of user testing was to create a tool that could be easily understood by the majority of users. One of the users, PH, already suggested that there could be a PRECIS-3 “Continuum doesn’t necessarily help you think about applicability. Is this a tool to help you think about reimbursement decisions or efficacy and safety but maybe this is PRECIS-3.”
<table>
<thead>
<tr>
<th>Key barriers/facilitators</th>
<th>Barrier/Facilitator Severity rating</th>
<th>Action taken</th>
</tr>
</thead>
<tbody>
<tr>
<td>The applicability scale (“1” very low applicability to “5” very applicable) ruins the paradigm of the explanatory/pragmatic model.</td>
<td>High * critical errors such as incorrect interpretation or high degree of uncertainty or dissatisfaction</td>
<td>Explanatory to pragmatic scale used</td>
</tr>
<tr>
<td>Lack of clarity in definition for pragmatic and explanatory. Early version explanatory “trial aiming to maximise the intervention’s chance of demonstrating efficacy and safety in ideal care” and pragmatic “aiming to maximise the intervention’s applicability in routine or standardised care”</td>
<td>Medium * much frustration or unnecessarily slow use</td>
<td>Used definition of pragmatic and explanatory in original 2009 PRECIS paper: “explanatory: interventions works under ideal conditions” and “pragmatic: intervention works under usual conditions”</td>
</tr>
<tr>
<td>Current information on how to use PRECIS-2 and each of the nine PRECIS-2 domains too long</td>
<td>Medium</td>
<td>Created 3 page information sheet</td>
</tr>
<tr>
<td>Lack of clarity and consistency in words patients, care recipients and provider in PRECIS-2 domains</td>
<td>Medium</td>
<td>Use word participants to include health care professionals, patients and other recipients of an intervention and/or providers of the intervention.</td>
</tr>
<tr>
<td>Difficulty seeing at a glance difference between delivery of intervention and adherence or compliance of patient or participant to intervention</td>
<td>Medium</td>
<td>PRECIS-2 domains now: Flexibility (Delivery), Flexibility (Adherence).</td>
</tr>
<tr>
<td>Organisation – what is needed to deliver the comparator?</td>
<td>Medium</td>
<td>Removed to reduce complexity – if left would need to have comparator domains for three other domains: Follow up, Flexibility Delivery and Flexibility Adherence.</td>
</tr>
<tr>
<td>Difficulty reading labels around PRECIS-2 wheel so some upside down and reader needed to move head or paper if printed copy.</td>
<td>Low * minor or cosmetic problems</td>
<td>Labels around PRECIS-2 wheel now all vertical</td>
</tr>
<tr>
<td>Benefit of PRECIS-2 table in addition to the PRECIS-2 wheel</td>
<td>Positive feedback</td>
<td>Use both table for rationale and wheel for easy visualisation or how pragmatic/explanatory a trial is</td>
</tr>
<tr>
<td>Cognitive approach on the PRECIS-2 wheel labels tested in user testing produced range of responses, not just “yes” and “no”.</td>
<td>Positive feedback</td>
<td>Information sheet that was developed also used questioning approach to frame domain description – like PRECIS-2 wheel.</td>
</tr>
<tr>
<td>Use of colour blue in font for PRECIS-2 domain names in wheel and as heading for domain information as “trusting colour”</td>
<td>Positive feedback</td>
<td>Blue used on PRECIS-2 website, in toolkit and in PRECIS-2 wheel headings.</td>
</tr>
</tbody>
</table>
PRECIS-2 went through 13 iterations during user-testing so changes were not made after each user testing of the 19 participants. Table 28 summarises the key barriers and facilitators identified, along with their severity. It was decided to keep the appearance of the PRECIS-2 wheel similar to the original PRECIS for the sake of continuity but provide tabular support for those who wanted it. PRECIS-2 domains could be scored on a scale of 1 to 5 with “1” very explanatory, “2” rather explanatory, “3” equally explanatory/pragmatic “4” rather pragmatic and “5” very pragmatic.

Following user testing with the 19 participants the steering group (KL, MZ, FS, ST) and a visiting GP (IG) discussed PRECIS-2 prior to the beginning of validity and reliability testing. The key change that was decided was to create a three page information sheet for validity and reliability testing, as testers needed to quickly grasp the PRECIS-2 tool to rate trial protocols and it would not be possible to give them the full elaboration paper.

This PRECIS-2 version would use a more questioning approach to frame the simple domain description, just like the PRECIS-2 wheel labels tested in user testing. This would enable a more cognitive approach which had been pointed out by one of the user testers (MW). It was important that questions posed in the domain descriptions simply could not be answered with “Yes” or “No” but would enable a range of responses. It was decided that each PRECIS-2 domain would have only one sentence for each domain describing how to score a pragmatic trial that is identical to usual care to a more explanatory trial:

- For Eligibility criteria: we could ask "to what degree are the participants included in the trial similar to those who would receive this intervention if it was part of usual care?"

- For Recruitment we could ask: "how much extra effort will be made to recruit participants, over and above those in the usual care setting in which the intervention might later be deployed?"

In the final version this became “how much extra effort will be made to recruit participants, over and above those in the usual care setting to engage with patients?”
There was an issue with easily and quickly distinguishing between two domains, the flexibility difference between delivering care and those who are RECEIVING the intervention: patients, health care providers. The label was changed to Flexibility (Delivery), Flexibility (Adherence).

The steering group also decided to use the word participants instead of patients in describing domains with a Note in the toolkit: “Participants” include patients or other individual recipients of an intervention, and/or providers of the intervention. This may include individual participants and/or one or more levels of clusters. For example, in a trial of a continuing education intervention, participants may be health professionals and trained instructors and the trial may be randomised into clusters at the level of the instructor.

**Discussion**

User testing gave insight into how PRECIS-2 users would use the tool. User testing changes occurred if: several participants agreed on an issue; it was clear that a change was desirable in improving functionality; and there was a strong belief that feedback improved the tool. One of the most positive aspects of user testing was that all the participants in user testing grasped the tool on first sight and the first impressions of all of the participants picked up that this was a multi-dimension tool with a scale for scoring decision making, some previous users of the tool noticed now outer circle not just “spokes” which had been one of the design points made during the Delphi phase. Most of the participants found the domain labels helpful, in particular for the first time user of PRECIS-2. For others, they acted as a reminder so they did not have to “cross reference” and go back to the domain description. A psychologist particularly liked the colloquial language for the domains and several doctor/trialists liked the questioning format for the domains: “Helpful, formalize what can take 1 to 2 years to learn. So helpful, as can’t remember domain description with 1-5 scoring system.”

The scale for evaluating user testing is based on DISsatisfaction: Low (minor or cosmetic problems), Medium (much frustration or unnecessarily slow use) and High (critical errors such as incorrect
interpretation or high degree of uncertainty or dissatisfaction). Most of the issues users had were with the wording of advice for using the tool, domains or restrictions missed out. KL was demonstrating a new or different version of a tool that some participants had seen before but most of the participants quickly grasped how the tool worked; were not frustrated by problems and interpreted the tool correctly. If the scale had been for satisfaction, most would have been highly satisfied which was the ultimate aim of user testing. More important suggestions related to scoring. Scoring PRECIS-2 using a scale of “1” to “5” was confirmed after user testing with three participants. However, the fourth participant in user testing (an American trialist who was a practising internist and professor who believed trials should be embedded in clinical practice) noticed a critical error in the Applicability scale for creating trials “fit for purpose”. This PRECIS-2 model assessed applicability using a scale of “1” very low applicability to “5” very applicable instead of “1” very explanatory to “5” very pragmatic. This user tester stated: “The applicability model screws up entirely explanatory/pragmatic model. Ruins paradigm.” KL had been struggling with the using the applicability scale and making it work in practice but it was only through user testing that the issue had been pinpointed.

Some participants made minor suggestions that were not considered by KL and ST as candidates for PRECIS-2 changes. The intention was to make changes that addressed all barriers that stopped people engaging with the tool but to be more selective with regards to minor barriers. An important factor in this decision making process was to enable previous users of the original PRECIS tool to recognise the modified PRECIS tool and therefore increase acceptance and uptake of the new PRECIS-2 tool.

Sometimes changes were made but not immediately after user testing had indicated that they would improve the tool. This was often due to pre-arranged testing and changes could not be made or discussed (with ST) prior to the next appointment. However, while changes for instance to “Advice on using PRECIS-2” were not changed and further testing emphasised that changes were necessary, different domains were tested with six users (9th to 14th user tests) collating further information to improve the tool.
Three of the users said that they would not use PRECIS-2 and although they gave feedback were unable to score the difficulty in using “as can’t answer, wouldn’t use”. The reasons for this may vary but could be that one participant did not fully understand the pragmatic/explanatory continuum (this participant was not a native English speaker); another thought all trials should be of pragmatic design and did not like the PRECIS-2 wheel and process; and the third user tester believed trial design was self-explanatory and influenced by other factors including predetermined specifications in grants so the pragmatic/explanatory continuum was not relevant. The other participants in user testing had a range of experience in designing clinical trials or were involved in systematic reviews (1) and found the tool relatively easy to use. There is no reason to believe that these participants were simply stating what the researcher KL wanted them to say as the user testing produced insightful suggestions and considered answers, and they had nothing to gain by stating that the tool was easy to use when it was not. However, it was clear that many were aware that using PRECIS-2 (as PRECIS) would get easier with more practice and if working in a team to discuss the domains and carefully evaluate exactly what is intended by trialists designing a trial.

User testers’ input was based on their diverse experience in trial research, systematic reviewing and practical clinical experience which included management. They provided useful input from the perspective of designing trials to meet targets of grant funders and including participants who were both patients – health care recipients and health care providers. Their input also highlighted issues of informed consent, with one participant recommending informed consent AFTER randomisation using Zelen designs to assist making trial results more applicable and participants in trials more comparable to usual “real world patients” [144, 145]. Another participant suggested PRECIS-2 would work well as part of a package of methods including the STEPS business model for recruitment and marketing the trial to funders and participants - both recipients of the intervention and those delivering the trial intervention [146].
Some users were involved who would consider using PRECIS-2 for systematic reviewing and they highlighted the issue that it is hard to score the PRECIS-2 domains without clinical expertise of the trial area. There were suggestions to pilot test the PRECIS-2 methodology and train users in applying PRECIS-2. In addition, issues of using PRECIS-2 independently, without discussing scores, were also mentioned; ensuring the individuals in a team discussed scores to reach consensus was important, as inaccurate scoring could be important in systematic reviewing and not just in trial design.

**Study strengths and limitations**

One of the strengths of this user testing study is that KL was involved in all the sessions so there was consistency in testing and continuity in developing the tool, ensuring that there was less likelihood of missing out any issues about domains or restrictions in moving from more pragmatic to more explanatory. As sessions were also based on availability of the participants, the testing was fitted in around the participant’s schedule and did not thus pose the same logistical problems of gathering a group of trialists together at the same time and place to discuss the development of PRECIS. Using Skype enabled testing of PRECIS-2 across time zones and prevented the expense of travelling for a short time period which would not have been feasible. One of the testers, a psychologist, was very positive about user testing stating that “Organised, a lot of information to cover within an hour.”

Some of the user testing sessions were particularly insightful and valuable; LF for instance in the fourth user testing immediately and completely understood the “fit for purpose” in designing trials which was refreshing but then spotted the flaw in using a scale of applicability for PRECIS-2 instead of explanatory/pragmatic scale to assess applicability – a trial is or it is not “fit for purpose” so sliding scale for applicability impossible to measure. This user testing emphasised the importance of not working in isolation but getting input from different people and that user testing works both ways: it is a useful experience for both parties involved.
There were, however, challenges. One of the difficulties with user testing is not being able to discuss a question; KL was trying to find out what the user thought about the proposed PRECIS-2 tool. Sometimes KL rephrased questions but some of the participants wanted to discuss particular aspects and find out what KL thought or indeed what KL thought about their answers. Sometimes it was very hard not to engage in discussion but usually reiterating “My role is to ask questions. But, since it is your opinion we are interested in, I will be otherwise saying as little as possible.” – a statement taken from the introduction to user testing. Sometimes at the end of user testing there was some discussion off the record about the user testing content and development of PRECIS-2 in which KL gave her opinion. In the Eighth user testing it became clear that incorporating the restrictions in how to make a trial more explanatory into the user testing, would assist in developing the tool and test use of PRECIS-2. It had been difficult not to engage in answering specific questions on how to score a domain. For one of the participants, however, the session was used more as a teaching opportunity to put their views on pragmatism to KL and the statements did lead to a useful discussion about PRECIS-2 development. In another session there was opposition to the format of the testing and there was not positive feedback about the tool. However, this demonstrated that a range of participants had been invited to comment on the tool – and as in the real world – not everyone will want to use PRECIS-2 and be willing to support developing the concept. Overall though, Skype enabled a friendly contact with participants that KL had not met before. As the structure of the user testing was already decided prior to the meeting, this prevented any hesitancy or shyness from KL in talking to “high flyers” who had published research citing PRECIS-2 in high calibre journals. It did mean that it was a learning opportunity for KL as views and opinions were imparted that KL was not familiar with.

Language was sometimes an issue. In some of the user testing sessions, KL talked to the user on the telephone (Brazil, Greece, USA) while sharing the document using Skype. If there were sufficient delays with sharing the user testing document, sometimes the document was sent to the user with strict instructions that it was a confidential draft document and not to be shared. This occurred twice. It was
clear that there were issues with discussing the tool in English which was not the native language of the participant in user testing. KL believed, however, this was an important part of user testing to ensure the language used was as simple as possible to help everyone who wanted to use the tool to understand it. Plain language would also assist translation.

There was one data collection error, with data for one user testing session being lost and unfortunately inadequate notes had been taken to save the qualitative feedback on domains. Future testing ensured the i-phone used for recording was charged or charging when recording to ensure no loss of data.

There were 13 iterations of the PRECIS-2 tool so changes were not made after each user testing of the 19 participants. User testing changes occurred if several participants agreed on an issue, it was clear that a change was desirable in improving functionality and there was a strong belief that feedback improved the tool. User testing, however, while largely positive did not produce the final version of PRECIS-2 that was tested in reliability and validity testing. While there was progress in developing the tool the final adjustments were made following user testing by the Steering group prior to use of the tool during reliability and validity testing.

Conclusions

PRECIS-2 went through 13 iterations during user-testing. We decided to keep the appearance of the PRECIS-2 wheel similar to the original PRECIS for the sake of continuity but provide tabular support for those who wanted it. Overall there was positive feedback from a diverse group of people some of whom were familiar with the original PRECIS tool but others not. User testing also assisted with resolving minor issues or cosmetic points which did not have serious consequences. User testing did resolve a big problem of using a scale of applicability for PRECIS-2 instead of a scale indicating how explanatory or pragmatic a trial was.
Summary
After 13 iterations, a PRECIS-2 model with nine domains (e.g. Eligibility, Recruitment, Flexibility Delivery) was ready to be tested for validity and reliability.

Conclusions
User testing provided helpful insight on PRECIS-2 and assisted in keeping the tool as simple and easy to use as possible for trialists and other users for instance systematic reviewers.
Chapter 7A: Methodology and pilot for PRECIS-2 validity and reliability testing

The only relevant test of the validity of a hypothesis is comparison of prediction with experience.

Milton Friedman

Introduction

In evidence based medicine (EBM), gold standards are important to assist with diagnosis, determine treatment and prognosis. For example in diagnosing a brain tumour, a blood test, is an essential part of this pathway to assist with disease diagnosis. This laboratory test is only an accurate predictor of disease if the test has been validated and proven that an abnormal blood sample is an indicator for a tumour. Magnetic Resonance Imaging (MRI) has been shown to be the gold standard for diagnosing a brain tumour.

In health research there are a number of tools that have become widely accepted as being useful and have face validity. Perhaps because an influential person develops and promotes the tool, encouraging others to use the tool. The tools themselves though have not been tested for construct or discriminant validity or rigorously tested for reliability but are accepted as part of the tool kit in research methodology. For instance the GRADE approach (Grading the quality of evidence and the strength of recommendations) (http://www.gradeworkinggroup.org/index.htm)
has been adopted by many organisations including the World Health Organisation, the British Medical Journal and NHS Quality Improvement Scotland. The Normalisation Process Theory (NPT) [http://www.normalizationprocess.org/] is another increasingly accepted method for implementing and evaluating complex interventions but it has no validity and reliability work. Another tool, used in the thesis, that has no formal validity that also requires judgements to be made is the Risk of Bias tool [46], essential in the assessment of internal validity of included RCTs in the systematic reviews of interventions in The Cochrane Library.

We believed that testing the validity and reliability of PRECIS-2 was important on several levels. Firstly we wanted PRECIS-2 to have face validity through involving a large number of participants and potential users of the tool in creating and developing the PRECIS-2 tool. This was to ensure that the PRECIS-2 tool would be accepted as measuring how pragmatic trials were on a continuum of explanatory to pragmatic - as intended. We also wanted to assess the construct validity of PRECIS-2 or discriminant validity and reassure ourselves that the tool does measure what it is intended to measure i.e. PRECIS-2 has 9 domains that were considered to be important in measuring pragmatism in RCTs. Finally, we also wanted to assess the inter-rater reliability of PRECIS-2 and evaluate if different raters could use the tool with the same information and get similar scores.
There were no published validity and reliability studies on the original PRECIS tool. Prior to PRECIS joint publication [48, 49], however, work was undertaken on face validity and inter-rater reliability. As there were problems validating the trial design tool, the published article became more conceptual and encouraged others to develop PRECIS. The published version of PRECIS in 2009 had 10 domains but unpublished work prior to this publication indicated that the developers added and deleted domains.

The development process for validating PRECIS information presented here comes from e-mails exchanged between collaborators. In addition, to the project website for a group developing tools to support the conduct of pragmatic trials, which had a presentation on “Spokes” (the PRECIS wheel) by Dave Sackett on behalf of over 23 collaborators [65]. This has now been suspended.

Initially, PRECIS had had eight domains: Participant eligibility criteria, Intervention flexibility, Practitioner expertise (intervention), Follow-up intensity, Follow-up duration, Participant compliance, Practitioner adherence and Primary analysis. This early work on developing PRECIS indicated there were problems with validating the tool. For example, seven raters looked at two trials and in both there was disagreement on almost all domains as to where to place the trials on the pragmatic-explanatory continuum. For one domain, Intervention flexibility, there were scores at the extremes, i.e. the trials were considered both highly pragmatic and highly explanatory. This work used a 5-point scoring system and the PRECIS collaborators concluded that their definitions for how to use this system were inadequate. They then considered switching to a 3-point scale. Further work with up to 12 raters looking at three trials (called ICD [147], CRASH [148] and BREECH [149]) found a great deal of variability. A histogram showing the variability across the pre-publication eight domains of PRECIS is shown in Figure 13; ideally each domain (Q1, Q2 etc.) would have one column. They concluded that there was little agreement. On a 1-10 scale of difficulty of use, where 1 was ‘Very easy’ and 10 was ‘Very difficult’, 25 of 32 raters scored PRECIS as 5 or above. By consensus, PRECIS was modified in two ways: one domain was dropped Follow-up duration and three were added Primary Trial Outcome and Practitioner Expertise Comparison Intervention and Comparison intervention flexibility giving the
current ten and the scoring system was removed. The paper describing PRECIS was made more conceptual and described as a work in progress [48, 49].

### Figure 13 Scoring counts for up to 12 raters using the eight domain pre-publication version of PRECIS with the ICD, CRASH and BREECH trials

At the time of writing (August 2014) twelve independent teams have since published studies looking at the properties of the 10-domain PRECIS tool and all have identified problems. Although it is simple to add a 0-4 [73, 150], 0-5 [64], 1-5 [39, 54, 55, 67, 70], 0-10 [58], 1-20 [40] or similar scoring system, whether this is sensible, or what the difference between, say, a score of 4 and a score of 3 really means remains unclear. Koppenhaal and colleagues thought their scoring on a 1-5 or a 0-100% scale seemed arbitrary, though they eventually favoured 1-5 [39]. Several authors calculated average domain scores together with variance. Taking data from Koppenhaal et al and expressing domain variance as a proportion of that domain’s average score, gave a median variance of 34% across all domains [39]. Riddle et al measured up to 42% variance across all domains [53]. Witt in a systematic review (using Likert scale of 1-10) had low inter-rater reliability after the first round of scoring Primary analysis for instance -0.12 Intra class coefficient (ICC) and but after a consensus meeting obtained an ICC of 0.77 and usually only one point difference for all domains [70]. Glasgow et al [56] did three studies using

<table>
<thead>
<tr>
<th>Domain</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICD</td>
<td></td>
</tr>
<tr>
<td>Q1</td>
<td>1</td>
</tr>
<tr>
<td>Q2</td>
<td>2</td>
</tr>
<tr>
<td>Q3</td>
<td>3</td>
</tr>
<tr>
<td>Q4</td>
<td></td>
</tr>
<tr>
<td>Q5</td>
<td></td>
</tr>
<tr>
<td>Q6</td>
<td>4</td>
</tr>
<tr>
<td>Q7</td>
<td></td>
</tr>
<tr>
<td>Q8</td>
<td>7</td>
</tr>
<tr>
<td>Q9</td>
<td></td>
</tr>
<tr>
<td>Q10</td>
<td>10</td>
</tr>
<tr>
<td>Q11</td>
<td></td>
</tr>
<tr>
<td>Q12</td>
<td>12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CRASH</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>1</td>
</tr>
<tr>
<td>Q2</td>
<td>2</td>
</tr>
<tr>
<td>Q3</td>
<td>3</td>
</tr>
<tr>
<td>Q4</td>
<td></td>
</tr>
<tr>
<td>Q5</td>
<td></td>
</tr>
<tr>
<td>Q6</td>
<td>4</td>
</tr>
<tr>
<td>Q7</td>
<td></td>
</tr>
<tr>
<td>Q8</td>
<td>8</td>
</tr>
<tr>
<td>Q9</td>
<td></td>
</tr>
<tr>
<td>Q10</td>
<td>10</td>
</tr>
<tr>
<td>Q11</td>
<td></td>
</tr>
<tr>
<td>Q12</td>
<td>12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>BREECH</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>1</td>
</tr>
<tr>
<td>Q2</td>
<td>2</td>
</tr>
<tr>
<td>Q3</td>
<td>3</td>
</tr>
<tr>
<td>Q4</td>
<td></td>
</tr>
<tr>
<td>Q5</td>
<td></td>
</tr>
<tr>
<td>Q6</td>
<td>6</td>
</tr>
<tr>
<td>Q7</td>
<td></td>
</tr>
<tr>
<td>Q8</td>
<td>8</td>
</tr>
<tr>
<td>Q9</td>
<td></td>
</tr>
<tr>
<td>Q10</td>
<td>10</td>
</tr>
<tr>
<td>Q11</td>
<td></td>
</tr>
<tr>
<td>Q12</td>
<td>12</td>
</tr>
</tbody>
</table>
PRECIS and obtained median variances of 37%, 68% and 47%. In The MOHR study there was either complete agreement or 1 point difference with average variances of 9% [150] while Sanchez using two raters calculated weighted agreement scores for PRECIS ranged from 63.9 to 78.5 %, with a median of 73.9 % [67]. The last three groups had a clear training programme, building on from the original study by Glasgow [56] as the researchers gained experience using the PRECIS tool, with Sanchez pilot testing the rating process with a sample of four trials [67, 68].

Finally, 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) [53]. The most trials considered was 20 (when this PhD PRECIS validity and reliability testing work started) but only by two raters [39] though a subsequent study considered 113 trials but again with only two raters [67]. Overall, when this PRECIS improvement and validation PhD project started in October 2012, there appeared to be clear scope for improvement as there was so much variation limiting our confidence in the validity and reliability work that had been undertaken. Therefore it was important to ensure we had an adequate sample size (see Sample size calculations Chapter 7B page 221).

**Aims and Objectives**

**Aims**

To test the validity and inter-rater reliability of the modified PRECIS-2 tool.

**Objectives**

The aim of the validity and reliability testing was achieved through a number of objectives.

1. Recruit sufficient experienced trialists, methodologists to assist with validity and reliability testing to have adequate sample size

2. Select a broad spectrum package of trials or protocols for validity and reliability testing
3. Make it as easy as possible for individual participants to assist with validity and reliability testing

4. Analyse the inter-rater variability using the Intra-Class Coefficient (ICC) for all nine PRECIS-2 domains

5. Analyse the discriminant validity of PRECIS-2 to determine pragmatism

6. Assess the face validity of the PRECIS-2 tool

**Methods**

**Test materials**

KL and ST selected protocols from 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 Initiative in 2007 [151]. We decided to use protocols as we felt they were a better reflection of trial information that was closest to the information trialists would consider when designing a trial and give more design details than final publications. In addition, reporting issues by authors that have not adhered to CONSORT guidelines for final trial reports could have also been an issue with missing information [152].

In selecting the protocols for validity and reliability testing, ST and KL screened the 150 SPIRIT protocols by looking at the abstracts. In addition, all trial protocols longer than 60 pages were automatically excluded to reduce burden on raters. (We believed that using several 100 page trial protocols would mean that participants would rate fewer trials using PRECIS-2 and our assumption turned out to be correct (Recruitment rate for the validity and reliability study p235). To ensure a broad range of trials were selected, five categories were noted: country, number of centres and type, length of protocol (number of pages) and number of participants (see Appendix Table 7.1). Reasons for exclusion are also given (see Appendix Table 7.1. The trials were narrowed down to 34 by ST and KL before 15 were
chosen by KL which represented a range of trials including factorial and cluster randomised trials which had not been considered by the original tool. It became clear that the SPIRIT database of trials seemed to have more pragmatic trials and that the shorter protocols are less likely to be explanatory. However, the final purposively selected 15 protocols (Table 29) included a range of trials with different publications, different countries and different authors with different interventions.

Trials were selected to exemplify the different trial types and demonstrate particular aspects of trials that would stimulate discussion to develop PRECIS 2. Some were drug trials, others physiotherapy or educational programmes in a variety of settings, countries and published in different journals with different designs, e.g. cluster randomised and factorial analysis design. Five of these trial protocols were then used in the pilot study to test suitability and methods prior to the main validity and reliability study.
<table>
<thead>
<tr>
<th>Trial</th>
<th>Intervention</th>
<th>Comparator</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. <strong>Tripathy P 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.</strong> <em>Trials</em> 2011, 12:182.</td>
<td>Improve maternal and newborn health in a participatory learning and action cycle where ASHAs support community women's groups through a four-phase process in which they identify and prioritise local maternal and newborn health problems, implement strategies to address these and evaluate the result</td>
<td>In common with intervention: 1. Carry out at least one village health committee meeting about rights and entitlements in each village during the study period. 2. Organise meetings with government officials and hospital management committees to inform the provision of appropriate care for mothers and newborns in facilities in the study districts. 3. Carry out at least one meeting with ASHAs to strengthen their job motivation and help them further enhance their work performance.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cluster (5000 population) Total - 30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Prospective, multicentre, randomized, parallel-group, superiority trial</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Randomised controlled trial</td>
</tr>
<tr>
<td>Trial</td>
<td>Intervention</td>
<td>Comparator</td>
</tr>
<tr>
<td>-------</td>
<td>--------------</td>
<td>------------</td>
</tr>
<tr>
<td>6. Abbott JH, et al: Exercise therapy, manual therapy, or both, for osteoarthritis of the hip or knee: a factorial randomised controlled trial protocol. Trials 2009, 10:11.</td>
<td>Three interventions: (1) a supervised multi-modal exercise therapy programme (2) an individualised manual therapy programme (3) both exercise therapy and manual Therapy</td>
<td>No physiotherapy, usual medical care and other health providers will be “usual care”</td>
</tr>
<tr>
<td>Trial</td>
<td>Intervention</td>
<td>Comparator</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>7. Fluid lavage of open wounds (FLOW): design and rationale for a large, multicenter collaborative 2 x 3 factorial trial of irrigating pressures and solutions in patients with open fractures. <em>BMC Musculoskelet Disord</em> 2010, 11:85.</td>
<td>5 interventions: Soap - 80 mL of the clear liquid soap (Castile Soap, Triad Medical Inc. Franklin, Wisconsin - 17% concentration in de-ionized water preserved in 90 mL bottles) with a sterile syringe into a 3L bag of normal saline. 1. soap + low pressure 2. soap + gravity flow pressure 3. soap + high pressure 4. saline + low pressure 5. saline + high pressure</td>
<td>Normal saline group (control) will receive sterile normal saline provided in 3L bags with gravity flow pressure</td>
</tr>
<tr>
<td>8. Harvey LA, Dunlop SA, Churilov L, Hsueh YS, Galea MP: Early intensive hand rehabilitation after spinal cord injury (&quot;Hands On&quot;): a protocol for a randomised controlled trial. <em>Trials</em> 2011, 12:14.</td>
<td>Usual care plus early intensive hand rehabilitation</td>
<td>Usual care - so will NOT receive any electrical stimulation to the target hand or upper limb nor will they be exposed to the instrumented exercise workstation.</td>
</tr>
<tr>
<td>9. Dalum HS, Korsbek L, Mikkelsen JH, Thomsen K, Kistrup K, Olander M, Hansen JL, Nordentoft M, Epol LF: Illness management and recovery (IMR) in Danish community mental health centres. <em>Trials</em> 2011, 12:195.</td>
<td>Illness Management and Recovery (IMR) programme - a curriculum-based psychosocial intervention designed as structured programme with a recovery-oriented approach. The aim of IMR is to rehabilitate people with severe mental illnesses by helping them acquire knowledge and skills in managing their illness and achieve personal recovery goals.</td>
<td>Usual care - This means individual adapted interdisciplinary treatment including medication, individual support, occupational therapy, psycho-education and group therapy.</td>
</tr>
<tr>
<td>Trial</td>
<td>Intervention</td>
<td>Comparator</td>
</tr>
<tr>
<td>-------</td>
<td>--------------</td>
<td>------------</td>
</tr>
<tr>
<td>10. Sandercock P, et al: Third international stroke trial (IST-3) of thrombolysis for acute ischaemic stroke. <em>Trials</em> 2008, 9:37.</td>
<td>'Immediate rt-PA' - recombinant tissue-type plasminogen activator (Alteplase, Boehringer Ingelheim; or Activase, Genentech) in a total dose of 0.9 mg per kg of body weight up to a maximum of 90 mg. Ten per cent of the dose is given as an intravenous bolus delivered over one minute followed by the rest of the infusion over the next 60 minutes.</td>
<td>Control must avoid treatment with rt-PA and should receive stroke care in the same clinical environment as those allocated 'immediate rt-PA'.</td>
</tr>
<tr>
<td>Trial</td>
<td>Intervention</td>
<td>Comparator</td>
</tr>
<tr>
<td>-------</td>
<td>--------------</td>
<td>------------</td>
</tr>
<tr>
<td>15. Vass CD, Sahota O, Drummond A, Kendrick D, Gladman J, Sach T, Avis M, Grainge M: REFINE (Reducing Falls in In-patient Elderly)—a randomised controlled trial, <em>Trials</em> 2009, 10:83.</td>
<td>Bedside chair and bed pressure sensors, incorporating a radio-paging alerting mode to alert staff to patients rising from their bed or chair</td>
<td>Those without equipment sensors – in pilot nurses detected “dummy equipment so control group NO sensors.</td>
</tr>
</tbody>
</table>
Pilot study

Pilot study aims
To test the materials and process for validity and reliability testing.

Pilot study objectives
The aims of the pilot were achieved through a number of objectives.

1) Test if the instructions for using the PRECIS-2 tool are clear to users.
2) Check if the domain instructions are clear to enable users to score pragmatism on a score of 1 (very explanatory) to 5 (very pragmatic).
3) Decide if we need to shorten the information document on “How to use PRECIS-2?”.
   • If yes, decide what should be included in the crib sheet.
4) Decide if the trial protocols selected are suitable for the validity and reliability testing.
5) Measure approximately how long it takes raters to score pragmatism using PRECIS-2.

Pilot study participants
For the pilot study we invited a visiting researcher IG (an academic general practitioner in Germany involved in trials), three supervisors (FS, ST, MZ) and the PhD candidate KL.

Pilot study methods
The package consisted of three items: five pdf trial protocols purposively selected from the 15 protocols selected from the SPIRIT database for validity and reliability testing (Table 29), PRECIS-2 information draft document “Guide to using PRECIS-2” with a table for raters to mark their scores for individual trials. We used the full description of how to use the PRECIS-2 tool which included: how to use the tool Figure 15, the ten domains in PRECIS-2 with domain descriptions (Table 30), scoring using a Likert scale of 1-5 (Figure 14), examples of fully pragmatic trials and detailed information on how to reduce pragmatism for more explanatory trials. This document was part of a larger document (39
Table 30 Ten PRECIS-2 domains for pilot of validity and reliability testing

<table>
<thead>
<tr>
<th>Domain</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eligibility</td>
<td>The similarity between the criteria used to select patients for the trial and the criteria that would be used if the intervention was in usual care.</td>
</tr>
<tr>
<td>Recruitment</td>
<td>The similarity between the recruitment path for patients in the trial and the path that would be used if the intervention was in usual care.</td>
</tr>
<tr>
<td>Setting</td>
<td>The similarity between the setting of the trial and the usual care setting.</td>
</tr>
<tr>
<td>Organisation (intervention)</td>
<td>The similarity between the resources, provider expertise and the organisation of care delivery in the intervention arm of the trial and those available in usual care.</td>
</tr>
<tr>
<td>Organisation (comparator)</td>
<td>The similarity between the resources, provider expertise and the organisation of care in the comparison arm of the trial and those available in usual care.</td>
</tr>
<tr>
<td>Flexibility (providers)</td>
<td>The similarity between the flexibility in how providers can deliver the intervention and comparator and the flexibility likely in usual care.</td>
</tr>
<tr>
<td>Flexibility (patients)</td>
<td>The similarity between the flexibility in how patients must adhere to their allocated treatment and the flexibility likely in usual care.</td>
</tr>
<tr>
<td>Follow-up</td>
<td>The similarity between the intensity of measurement and follow-up of patients in the trial and the likely follow-up in usual care.</td>
</tr>
<tr>
<td>Primary outcome</td>
<td>The direct relevance of the trial’s primary outcome to patients and providers.</td>
</tr>
<tr>
<td>Primary analysis</td>
<td>The inclusion of all patients post-randomisation, in the analysis of the primary outcome.</td>
</tr>
</tbody>
</table>

Note: this table has changed - it is specific to this stage of the PhD project before the final version –
We think there are four steps to using PRECIS-2, which may be iterative depending on what you discover after going through the steps.

**Step 1: Why are you doing your trial?**

Your first step is to be clear about why you are doing your trial. Are you:

Aiming to take an explanatory approach and answer the question:

‘*Can this intervention work under ideal conditions?*’

Aiming to take a pragmatic approach and answer the question:

‘*Does this intervention work under usual conditions?*’

Both have their place but trialists should be clear which path they are on. This is not a dichotomy; there will be a continuum from pragmatic to explanatory. (Explanatory trials have also been called efficacy trials while pragmatic trials are also known as practical or effectiveness trials but we prefer to use the terms explanatory and pragmatic.)

**Step 2: Consider your trial design choices for each of the PRECIS-2 domains**

This step is explained in more detail for each domain below.

**Step 3: Score 1 to 5 for these choices** by making a mark on the PRECIS-2 wheel according to how pragmatic or explanatory each is in relation to the extremes of each domain.

**Step 4: Review your PRECIS-2 wheel**

Review your design choices (Step 2) on the PRECIS-2 wheel to see whether they will produce a trial that will support the aim identified in Step 1. Go back to Step 2 and modify your choices if required.

The five protocols selected were purposively very different to test out the process and determine the likelihood of validity and reliability of PRECIS-2: a physiotherapy trial in an army setting in Madigan,
USA [154]; a cluster randomised trial in rural India [155]; a trial on tonsillectomy in Northern England [156]; an ophthalmology surgical trial to treat glaucoma [157] and an international trial using thrombolysis to treat acute stroke patients [158].

The package for the pilot was sent out to the participants on the 5th September 2013 in preparation for the meeting in Dundee on the 9th September 2013. MZ would join the discussion by conference call from Toronto. The meeting (minimum time 1 hour) would discuss scores and changes that needed to be made to the package for the PRECIS-2 validity and reliability study.

**Pilot study results**
Due to time issues, only four participants (IG, KL, ST, FS) had scored three of the trial protocols: the physiotherapy versus corticosteroid trial [154]; the cluster trial in India [155] and the tonsillectomy trial [156] so only these three trials were used in the pilot. MZ had been unable to assist in the pilot using PRECIS-2 to score the trials but was part of the evaluation of the pilot PhD project. Raters in the pilot did confirm that it had taken approximately twenty minutes to score trials using PRECIS-2 and in some cases less time. ST had used a PRECIS-2 wheel but others had used the table to score the trials.

There was noticeable variability in scoring for the three protocols for the pilot.

The current document of 24 pages was too long and it was clear that a 3 page crib sheet would need to be prepared to make validity and reliability testing as easy as possible for raters. In addition a PRECIS-2 wheel had not been included in the pilot package though one been used by one of the raters (ST) so this was also included in the full validity and reliability study.
The first protocol involved two treatments - corticosteroid and physiotherapy treatment – which were both standard practice for treatment of shoulder injuries [154]. In the pilot, KL during discussion wrote down two scores for the Organisation domain for both interventions (See Table 31 – line 5 Organisation - comparison. In this case both treatments could be considered as the Intervention. As the organisation was similar for both interventions, then this extra domain was not inserted into the PRECIS-2 wheel, it was simply discussed to determine whether the steering group thought it should be included as the tenth domain.

There was disagreement in almost all of the domains. The Setting was considered to be explanatory by one of the raters (KL). As the setting for this trial was army only and organisation atypical to USA in general, and the majority of the population seeking treatment for shoulder injuries were out with the army, KL believed this had an impact on Eligibility and Organisation. The majority (three out of four raters) believed that if the results would be used in purely army setting then the trial would be in the “usual” setting and the results very applicable otherwise the participants were more likely to be fitter and not comparable to the general population.
The Eligibility criteria also included a lot of exclusion criteria as well as being military personnel only and KL rated this lower as well. In Organisation, it was not usual care as trial participants had shoulder treatment with standardised training for both physiotherapists who had trained within the same time period so not just any physiotherapist in the unit. As Primary Outcome included Shoulder Pain and Disability Index (SPADI) and the secondary outcomes were acronyms, GRC (Global Rating of Change questionnaire to measure overall changes in quality of life) and the NPRS (Numeric Pain-Rating Scale) they were quite probably not usual outcomes for physiotherapists; KL thought explanatory as probably not the usual primary outcome from the perspective of the physiotherapists and a surrogate outcome from the perspective of the patients, MZ agreed. In the NHS in the UK, physiotherapists in practice would usually use a generic pain scale (i.e. 0 to 10) which may or may not be combined with a muscle strength grading scale (0 to 5). These would be more pragmatic outcome measures for usual care.

Considering the Follow-up domain; this was 1, 3, 6 and 12 months. ST and IG thought this was rather pragmatic with similar frequency but each visit would be longer than usual as three assessments using Shoulder Pain and Disability Index (SPADI) for movement, Global Rating of Change (GRC) for quality of life and the Numeric Pain Rating Scale (NPRS) for pain. KL thought this domain should have a PRECIS-2 score of 3 and was equally pragmatic and explanatory as more than usual frequency of follow up with longer time taken than usual for appointments. Usual follow up would be limited to treatment sessions, with patient coming back if still in pain and checking medical records for further treatment at one year would be the most pragmatic follow up assessment.
Table 32 ASHA protocol no. 1 [154](from 15 selected for validity and reliability testing)

<table>
<thead>
<tr>
<th>Domain</th>
<th>KL</th>
<th>FS</th>
<th>ST</th>
<th>IG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eligibility Criteria</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Recruitment Path</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>?</td>
</tr>
<tr>
<td>Setting</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Organisation intervention</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3?</td>
</tr>
<tr>
<td>Organisation comparison</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Flex of experimental intervention – Providers</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Flex of experimental intervention – Patients</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Follow up</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>?</td>
</tr>
<tr>
<td>Outcome</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Analysis</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>4</td>
</tr>
</tbody>
</table>

In the second protocol (see Table 32 for scores) problems arose in interpreting the published information for the Accredited Social Health Activists (ASHA) study in India [155]. Was the cluster typical of rural India with barriers separating clusters involved in intervention and control? And was the population like the participants in usual care. It included men, women of all ages, however, this trial could be considered to be more explanatory as there was randomisation at the cluster level (not individual patient level) and the trial clusters had to have ten ASHAS to be involved in a trial on education. It was not clear that this was usual.
Table 33 NESSTAC – protocol no.11 [155](from 15 selected for validity and reliability study)

<table>
<thead>
<tr>
<th>Domain</th>
<th>KL</th>
<th>FS</th>
<th>ST</th>
<th>IG</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Eligibility Criteria</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>2 Recruitment Path</td>
<td></td>
<td>4</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>3 Setting</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>4 Organisation intervention</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>5 Organisation comparison</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>6 Flex of experimental intervention – Providers</td>
<td>5</td>
<td>2</td>
<td>4.5</td>
<td>5</td>
</tr>
<tr>
<td>7 Flex of experimental intervention – Patients</td>
<td>5</td>
<td>1</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>8 Follow up</td>
<td>2</td>
<td>2</td>
<td>4.5</td>
<td>2</td>
</tr>
<tr>
<td>9 Outcome</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>10 Analysis</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4</td>
</tr>
</tbody>
</table>

Rating this third protocol NESSTAC – protocol 11 [156] (see Table 33 for scores) indicated that adequately published information is key to using PRECIS-2 to rate trials on the pragmatic to explanatory continuum, as one of the pilot raters (FS) had insight into this trial as he had been involved. Thus FS had different scores for Flexibility (provider). He thought that the surgeon has to undertake only one or two operations prior to being part of the trial so more explanatory but as there were no specific details in the protocol publication on expertise necessary to operate, KL, ST and IG thought that the trial was very pragmatic. There were similar scores for Flexibility of experimental intervention (patients) for FS giving a score of “1 – very explanatory” for no choice and ST, KL and IG giving “5 – very pragmatic” believing there was full flexibility for the patient to either chose to have the operation or to choose not to have it. This protocol indicated that we might need to give further advice on surgical trials and advise raters to state Not Applicable (N/A) but we decided to test this out as part of the validity and reliability study.
Pilot study discussion

SPIRIT trials

The GP visiting from Germany suggested that the selected trials should have been more diverse, and were all more pragmatic. Although one trial protocol had been scored as more explanatory by KL [154], the other two [155, 156] were indeed more pragmatic. Unfortunately the other two that had only been assessed by KL and IG had included another that was more explanatory. The steering group decided to accept the 15 protocols selected for the validity and reliability testing as overall the 15 trial protocols were a diverse mix of trials on the continuum of pragmatic to explanatory.

Three page crib sheet

It was decided to include the four steps to use PRECIS-2, the Likert scoring scale from 5 (very pragmatic) to 1 (very explanatory), the domain names, and short one line descriptions with a further one line sentence summarising all the ways of reducing the score from a very pragmatic trial to a very explanatory trial. We had to ensure consistency in the use of labels for PRECIS-2 wheel and the crib sheet and the longer elaboration paper for PRECIS-2.

One key decision for the creation of the 3 page crib sheet was to use a more questioning approach to frame the simple domain descriptions and enable a cognitive approach. We had used this technique successfully in user testing for the domain labels on the PRECIS-2 wheel which were also questions (Figure 16). Like the domains descriptions, we had to make sure the questions for each domain could not simply be answered with “Yes” or “No” but a range of responses, indicating the degree of pragmatism. This was a point made during user testing by an academic doctor working in trials involving care of the elderly. Following the pilot there was consensus in doing this to make it easier to use the PRECIS-2 tool for domain descriptions as well as the labels for domain names on the PRECIS-2 wheel.
Figure 16 Wheel for PRECIS-2 following user testing and used in Pilot testing for Validity and reliability testing

Note: this PRECIS-2 wheel has changed - it is specific to this stage of the PhD project before the final version – Figure 41

For example, for the Eligibility domain we could ask "to what degree are the participants included in the trial similar to those who would receive this intervention if it was part of usual care?" and the PRECIS-2 wheel label was “who gets into the trial?”. For Recruitment we could ask: "how much extra effort will be made to recruit participants, over and above those in the usual care setting in which the intervention might later be deployed?" and the PRECIS-2 wheel label was “how do patients get into the trial?".

We decided to make it clear on the crib sheet that any questions raters had when using PRECIS-2 should be directed to KL or ST – their email address would be on the information sheet in addition to being
given in the invitation e-mail. In addition to the information sheet we would send a table and wheel for raters to use to allow raters to score trials in whichever way they found easiest.

Cluster trials

There was discussion amongst pilot raters on clusters in rural India after using PRECIS-2 to score this trial on using ASHAS to improve maternal and newborn health [155]. In the draft elaboration paper which was used for this pilot, there was no mention of clusters. The use of this trial in the pilot highlighted that this needs to be added to the elaboration paper and could be a useful trial example for PRECIS-2 scoring [153]. The recommendation from KL was to insert one brief sentence into the paper on PRECIS-2 selection of participants twice, clusters and individuals.

Domain changes

The difference between two of the domains was discussed: Flexibility (providers) and Flexibility (receivers). The difference between DELIVERING care and those who are RECEIVING the intervention: patients, public or health care providers is sometimes not obviously straightforward so we need to be sure these domains are easily distinguishable. There was a suggestion to change the names of these domains to Flexibility (delivery) and Flexibility (adherence) and to add note on the definition of participants. In user guide information for the pilot stated in the text – participants were patients, public or health professionals. The suggestion was to add text making this clearer so add a note to 3-page crib sheet “Participants” include patients or other individual recipients of an intervention, and/or providers of the intervention, in all arms of the trial. This may include individual participants and/or one or more levels of clusters.”

The most noticeable decision of the steering group following user testing was to have nine domains for PRECIS-2 instead of 10,
Table 34. The domain that was dropped was the *Organisation – what is needed to deliver the comparator?* If we did include the domain *Organisation (comparator)* then to be consistent there would have to be comparator domains for three other domains: *Follow up, Flexibility Delivery* and *Flexibility Adherence*. Doing this was considered to increase the complexity of the tool; we did not want the new tool to be more complicated to use than the old tool. More importantly, the proposed comparator was usual care.

**Scoring**

Following user testing we decided to use a Likert scale 1-5. There was discussion on the wording for 2 = rather explanatory and 4 = rather pragmatic. There was concern that the word “rather” may not be easily understood to non-native English speakers. During user testing the word “mostly” had been used as well as “rather”. After discussing whether or not we keep both “mostly/rather” the steering group decided best to just use the word “rather”. IG had felt the scoring required a judgment call “a gut feeling” and wanted more “concrete advice for it”. KL assured more examples would be included in PRECIS-2 elaboration paper and website to assist users.

**Conclusions: Pilot study key issues**

Following discussion of the pilot results, we decided that we would use the validity and reliability study to determine how raters dealt with the following three issues we had previewed in pilot testing.

1. Cluster trials – randomisation at organisational level
2. Intervention comparison of two standard care treatments; and
3. Surgical trials and how raters dealt with scoring the *Flexibility Adherence* domain.

We have used the feedback on these issues from participants in validity and reliability testing to add additional information to the guidance for users on PRECIS-2 [153] and the PRECIS-2 website www.PRECIS-2.org.
The pilot was the basis for the 3 page crib sheet (Box 1) and key changes in the way domains were described. Finally, an important change was the removal of one of the domains – *Organisation* *Comparison* – otherwise we would have to include extra domains for *Follow-up* and *Flexibility Adherence*. Thus (Table 34) outlines the nine domain descriptions for validity and reliability testing. We wanted to keep the tool simple to enable trialists to design trials fit for purpose and believed that the nine selected domains would enable them to do this. The comparator or control in most cases was usual care but if not then it was likely that there would be few differences in domain rating for trials with several intervention arms. The development process from the start with the original tool to the finish with the PRECIS-2 used in the validity and reliability testing are indicated in

**Table 34** The nine PRECIS-2 domains descriptions, tested in validity and reliability testing

<table>
<thead>
<tr>
<th>Domain</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eligibility</td>
<td>to what extent are the participants in the trial similar to those who would receive this intervention if it was part of usual care?</td>
</tr>
<tr>
<td>Recruitment</td>
<td>how much extra effort is made to recruit participants over and above what that would be used in the usual care setting to engage with patients?</td>
</tr>
<tr>
<td>Setting</td>
<td>how different is the setting of the trial to the usual care setting?</td>
</tr>
<tr>
<td>Organisation (intervention)</td>
<td>how different are the resources, provider expertise and the organisation of care delivery in the intervention arm of the trial and those available in usual care?</td>
</tr>
<tr>
<td>Flexibility (delivery)</td>
<td>how different is the flexibility in how the intervention is delivered and the flexibility likely in usual care?</td>
</tr>
<tr>
<td>Flexibility (adherence)</td>
<td>how different is the flexibility in how participants must adhere to the intervention and the flexibility likely in usual care?</td>
</tr>
<tr>
<td>Follow-up</td>
<td>how different is the intensity of measurement and follow-up of participants in the trial and the likely follow-up in usual care?</td>
</tr>
<tr>
<td>Primary outcome</td>
<td>to what extent is the trial’s primary outcome relevant to participants?</td>
</tr>
<tr>
<td>Primary analysis</td>
<td>to what extent are all data included in the analysis of the primary outcome?</td>
</tr>
</tbody>
</table>
Figure 17 Stages of development of PRECIS-2
The PRECIS tool: Designing trials that are fit for purpose

<table>
<thead>
<tr>
<th>Domain</th>
<th>Score</th>
<th>Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eligibility Criteria</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recruitment Path</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Setting</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Organisation intervention</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flex of experimental intervention</td>
<td>-5</td>
<td>Delivery</td>
</tr>
<tr>
<td>Flex of experimental intervention</td>
<td>6</td>
<td>Adherence</td>
</tr>
<tr>
<td>Follow up</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Outcome</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Analysis</td>
<td>9</td>
<td></td>
</tr>
</tbody>
</table>

The PRECIS-2 Domains

The NINE PRECIS-2 domains are:

**Eligibility** – to what extent are the participants in the trial similar to those who would receive this intervention if it was part of usual care? For example, score 5 for very pragmatic criteria essentially identical to those in usual care; score 1 for a very explanatory approach with lots of exclusions (e.g. those who don’t comply, respond to treatment, or are not at high risk for primary outcome, are children or elderly), or uses many selection tests not used in usual care.

**Recruitment** - how much extra effort is made to recruit participants over and above what that would be used in the usual care setting to engage with patients? For example, score 5 for very pragmatic recruitment through usual appointments or clinic; score 1 for a very explanatory approach with targeted invitation letters, advertising in newspapers, radio plus incentives and other routes that would not be used in usual care.

- **Setting** – how different is the setting of the trial and the usual care setting? For example, score 5 for a very pragmatic choice using identical settings to usual care; score 1, for a very explanatory approach with only a single centre, or only specialised trial or academic centres.

- **Organisation** – how different are the resources, provider expertise and the organisation of care delivery in the intervention arm of the trial and those available in usual care? For example, score 5 for a very pragmatic choice that uses identical organisation to usual care; score 1 for a very explanatory approach if the trial increases staff levels, gives additional training, require more than usual experience or certification and increase resources.

- **Flexibility (delivery)** – how different is the flexibility in how the intervention is delivered and the flexibility likely in usual care? For example, score 5 for a very pragmatic choice with identical flexibility to usual care; score 1 for a very explanatory approach if there is a
strict protocol, monitoring and measures to improve compliance, with specific advice on allowed co-interventions and complications.

**Flexibility (adherence)** - how different is the flexibility in how participants must adhere to the intervention and the flexibility likely in usual care? For example, score 5 for a very pragmatic choice involving no more than usual encouragement to adhere to the intervention; score 1 for a very explanatory approach that involves exclusion based on adherence, and measures to improve adherence if found wanting.

- **Follow-up** - how different is the intensity of measurement and follow-up of participants in the trial and the likely follow-up in usual care? For example, score 5 for a very pragmatic approach with no more than usual follow up; score 1 for a very explanatory approach with more frequent, longer visits, unscheduled visits triggered by primary outcome event or intervening event, and more extensive data collection.

- **Primary outcome** – to what extent is the trial’s primary outcome relevant to participants? For example, score 5 for a very pragmatic choice where the outcome is of obvious importance to participants; score 1 for a very explanatory approach using a surrogate, physiological outcome, central adjudication or use assessment expertise that is not available in usual care, or the outcome is measured at an earlier time than in usual care.

- **Primary analysis** – to what extent are all data included in the analysis of the primary outcome? For example, score 5 for a very pragmatic approach using intention to treat with all available data; score 1 for a very explanatory analysis that excludes ineligible post-randomisation participants, includes only completers or those following the treatment protocol.

Please note: if you have any queries regarding domains please contact Kirsty k.loudon@dundee.ac.uk or Shaun Treweek streweek@mac.com

Postal address to send PRECIS-2 scores:

Kirsty Loudon  
Division of Population Health Sciences  
University of Dundee  
Kirsty Semple Way  
Dundee DD2 4BF  
UK

THANK YOU!

Box 1 Three page information sheet prepared post pilot validity testing
Methods for the PRECIS-2 validity and reliability study

Selecting raters for the PRECIS-2 validity and reliability study

Having selected the 15 trial protocols and tested the methods and materials in the pilot study, consideration was given to the participants involved in the next stage of the study. To ensure adequate validity and reliability testing of PRECIS-2 it was important to have experienced and knowledgeable trialists so this was an essential prerequisite. We had access to trialists and methodologists through collaborations including the Cochrane Methodology Review Group, the CONSORT Group, the Scottish Clinical Trials Units, the DECIDE (Developing and Evaluating Communication Strategies to Support Informed Decisions and Practice Based on Evidence) project, as well as researchers who had been involved in the modified Delphi and early user testing and had stated they were interested in assisting with the validity and reliability testing. We also had a list of researchers who have worked with the original PRECIS tool, some of whom had contacted us directly regarding collaboration. This group also included members of the independent teams that had evaluated PRECIS, all of whom we have contacted previously; they had shared both data and enthusiasm for improving PRECIS. Invitations also were sent to three editors of medical/trial journals, one of whom had offered to help when learning that validity and reliability testing had just started. The invitations were thus sent to six different groups of people who comprised experienced academic doctors, trialists and researchers.

Sample size for validity and reliability testing

Having run the pilot we were keen to ascertain the sample size for the validity and reliability testing. Previous published and unpublished work on PRECIS did not test the validity and reliability of the tool due to inadequate sample sizes so this was important for our study. All methodological work had either an inadequate number of raters or involved too few trials to test the inter-rater reliability and face value of the PRECIS tool. We measured PRECIS-2’s inter-rater reliability using the Intraclass Correlation Coefficient, ICC in SPSS, version 15.5. The sample size calculations depend on the number of raters and the number of trials these raters assess. Previous work stated that rating a trial using PRECIS takes
between 10 and 30 minutes, which agreed with our pilot on PRECIS-2 of 20 minutes or less. As PRECIS-2 was similar the sample size was a compromise between precision and what we could reasonably ask raters to do. So considering reasonable workload for the validity and reliability testing and assuming the ICC would be in the region of 0.7 (with 15 raters looking at 10, 15 or 20 trials) this would lead to precision of +/- 0.20, +/- 0.17 and +/- 0.14, respectively. We thus aimed to give our 15 raters between 10 and 15 trials to rate. We were unable to consider scoring more than 15 trials as PRECIS-2 was not quicker to use than PRECIS and raters would have time constraints so be unable to complete more. There was a balance between what could reasonably be expected of individual raters. This assumption turned out to be correct, we were unable to get 15 raters to complete 15 trials (Recruitment rate for the validity and reliability study p235).

Rater invitations to validity and reliability testing

On September 24th, 2013, we invited 33 raters to participate by e-mail using personalised invitations, an additional two invitations were sent to two participants with similar background to ensure we had an adequate sample size.

If the invitation was sent to someone who had assisted with the Delphi and/or with user testing this was acknowledged in the introduction. Other sources of introduction, i.e. The Sackett Symposium at Niagara Falls September 2013 (http://fhs.mcmaster.ca/sackettsymposium/), Canada were also used as a means of starting the e-mail request to assist in the validity and reliability PhD project for journal editors who had attended the conference. If, however the invitation was sent to someone KL had not had contact with beforehand ST did the introduction which was then followed up by KL. We asked for a reply as soon as possible but were keen not to pressurise with deadlines in the invitation email. We clearly stated however that the validity and reliability testing would run over six weeks, requesting participants to rate up to 15 trial protocols in batches of five, with a new batch every two weeks. The anticipated deadline was 1st November.
As we realised this was a significant burden on time, we initially offered an incentive of £100 ($160) using funding from the Chief Scientist Office to recognise assistance. This was notionally to cover around four hours of work to each trialist/methodologist completing the PRECIS-2 ratings. At a later stage in the validity and reliability testing we offered £200 to encourage participation in the PhD project. We also let raters know that their help would be acknowledged in any publication on PRECIS-2.

### Statistical measure – inter-rater variability

We chose the intraclass correlation coefficient (ICC) as it was a relatively simple statistical measure to assess methodological variation and we wanted to find out if the raters were using the information for domain rating of PRECIS-2 consistently. The ICC is a measure that determines the reproducibility of a variable, which could easily be calculated from an ANOVA table (e.g. using SPSS). It is based on the assessment of both systematic deviation and random variation. We used SPSS and 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. they are a random sample of all raters and protocols.

Although the reliability estimates for mixed and random ICC models are numerically identical they can be interpreted differently. Using the random effects model we can generalise the results to other raters, which is not possible with the mixed effects model which assumes there will be no reason for higher or lower scores by a rater – “rater performance interaction” whereas with the random effects model this is taken into account. We also used Type C intraclass correlation coefficients using a consistency definition – this means that “the between-measure variance is excluded from the denominator variance”. This assumes that the raters have similar scores, so we are checking for consistency rather than absolute agreement. We did not need or expect perfect agreement in scores for trial protocol domains, we were simply looking for similar scores. Finally, we also took the “average measure” for all raters compared to the single measure ICC which only considers the ICC for one typical individual ‘average’ rater scoring a domain.
To determine the effect of missing data we undertook sensitivity analysis using different groups of raters and protocols scored using PRECIS-2 imputing missing data using a score of 3 - equally pragmatic/explanatory, as well as undertaking multiple sensitivity analysis (10 imputations) in which random missing values were inserted.

**Duplicate data entry check**

To check reliability and consistency of data entry of the data from the validators, there was independent data entry by KL and DO’F into Microsoft excel sheets. Excel sheets were compared electronically and then back up spot checks were undertaken to ensure scores for a particular domain in a particular trial by different raters were accurately entered.

**Lack of information in trial protocols**

We were keen to ascertain what testers would do if there is no information available to score; would raters “guess” or use another system perhaps miss out score or score 3 for pragmatic/explanatory midpoint or indeed score “1” and assume explanatory if no details given. As decided in our pilot we thus gave no advice to participants if there was inadequate information to determine how they dealt with uncertainty.

**Discriminant validity (Method 1)**

We determined if PRECIS-2 could accurately discriminate trials of varying pragmatism as judged by the subjective rating by calculating odds (discriminant validity)[52]. We used two methods to consider if PRECIS-2 could discriminate between pragmatic and explanatory trial. As mentioned previously, there was already concern about the burden of rating trials so to reduce the workload for raters assisting with PRECIS-2 validity and reliability testing KL did not ask them to assess how pragmatic the trial protocol was prior to using PRECIS-2. This prediction turned out to be accurate so this was a well justified precaution.
Instead KL determined if PRECIS-2 could accurately discriminate trials of varying pragmatism as judged by the subjective rating by calculating AUROC odds (discriminant validity [159]). Two raters (KL, ST) using binary scores of more pragmatic = 1, more explanatory = 0, rated the overall pragmatism for the 15 trials. This was done by making a judgement of pragmatism based on reading the trial publication. As KL and ST were very familiar with the PRECIS- the issues covered by the PRECIS-2 domains were inevitably (if subconsciously) used for determining the decision whether or not a trial was pragmatic or explanatory. We also used the median values of the pragmatism rating of the nine individual domains for each of the 15 trials assessed by all the raters in the discriminant validity calculations, so there were 15 values for each domain. The median values were calculated from up to 18 raters.

As well as giving subjective global ratings of pragmatism of the 15 trial protocols, “pragmatic” or “explanatory”, KL, ST also gave more specific ratings of trials on the pragmatic/explanatory continuum from 1 to 5: 1=very pragmatic; 3=neither one nor the other; 5=very explanatory using Ordinal logistic analysis.

Firstly, using SPSS, we tried to fit a binary logistic regression model of PRECIS-2. Considering each trial, we used the values of pragmatism as decided by KL and ST (Yes, No), and we used 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 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.) for the PRECIS-2 domains and the State variables as Pragmatism (Yes, No) we calculated how good each domain in the PRECIS-2 tool is at
predicting pragmatism (1) displaying a AUROC curve with diagonal reference line and Standard error and confidence interval.

**Discriminant validity (Method 2)**

In a follow up study to determine discriminant validity KL and ST independently scored each of the 15 trials by considering how many times a rater (up to 19) had selected an explanatory score of “1” very explanatory or “2” rather explanatory for the 9 domains. So for example in a trial, one rater could have selected one explanatory score, another none and another rater three explanatory scores for the nine domains of a trial. We added all of these scores up, e.g. 1 + 0 + 3 etc. and then divided by the number of raters (up to 19). These scores were then compared with the global pragmatism ratings for the explanatory or pragmatic trials that KL and ST had agreed on. We also considered if any particular domains could predict pragmatism or if a trial would be more explanatory.

**Validation and reliability logistics**

Following agreement to assist in the validity and reliability testing PhD project, our plan was that all participants would be sent: a concise PRECIS 2 training package which comprised of a 3-page crib sheet (this could also be downloaded on the PRECIS-2 website that had been developed simultaneously), a PRECIS-2 wheel and table that could be used for scoring and between 5 and 15 trial protocols. The first five [154, 155, 157, 160, 161] were sent and then another five protocols [158, 162-165] once the PRECIS scores for the first batch had been received. The protocols were sent in batches of five, two weeks was suggested as the maximum time for rating trial designs.

<table>
<thead>
<tr>
<th>SUMMARY</th>
</tr>
</thead>
<tbody>
<tr>
<td>The materials and methodology for the full validity and reliability study were pilot tested using four raters (three members of the steering group and one visiting primary care trialist from Germany) with input from another steering group member. PRECIS-2 was finalised as a 9 domain tool: Eligibility, Recruitment, Setting, Flexibility (Delivery), Flexibility (adherence), Follow up, Primary Outcome and Primary Analysis. A 3-page crib sheet was created for the main validity and reliability testing. KL invited 33 experienced academic doctors, trialists and researchers to participate in validating PRECIS-2 using up to 15 trial protocols.</td>
</tr>
</tbody>
</table>
Chapter 7B: Results of the main study for validity and reliability testing

Introduction

Following the pilot of the SPIRIT trial protocols to test the materials and decide the best way to undertake the validity and reliability testing, sample size (as described in chapter 7a) was then calculated, based on the desired precision needed to validate PRECIS-2. The methodology for considering the discriminant validity of PRECIS-2 was determined using two methods (as described in chapter 7a) and electronic invitations to potential participants were then sent out.

Results

Participants

We initially invited 33 experienced trialists and methodologists (Table 35), to participate in Phase 2 inter-rater reliability and discriminant validity work and 30 responded - giving a response rate of 91%. Unfortunately not everyone was able to participate, and due to concerns about sample size we subsequently invited another two raters, firstly a trialist and former nurse who was experienced in CONSORT methodology and secondly another statistician who was familiar with PRECIS-2. Table 36 indicates the recruitment from the different groups of participants invited to assist in the PRECIS-2 validity and reliability testing. There was thus a 54% participation rate (19/35). The multiple input into PRECIS-2 development by some of the participants is outlined in Figure 18 indicating that six of the participants has also assisted in the modified Delphi, two had contributed through user testing and four had been present at the Toronto Brainstorming meeting.
<table>
<thead>
<tr>
<th>Name</th>
<th>Country</th>
<th>Research Area</th>
<th>Further phases (PRECIS-2 interest)</th>
<th>Participation</th>
<th>No of protocols completed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sabine N Van der Veer</td>
<td>The Netherlands</td>
<td>Medical informatics, trialist</td>
<td>Validity &amp; reliability testing</td>
<td>Yes</td>
<td>15</td>
</tr>
<tr>
<td>Vivian Andrea Welch</td>
<td>Canada</td>
<td>Clinical Epidemiology Methodologist Systematic reviewer</td>
<td>2nd round, brainstorming, validity &amp; reliability testing</td>
<td>Yes</td>
<td>4</td>
</tr>
<tr>
<td>Eric Brass</td>
<td>USA</td>
<td>MD, PhD, trialist, pharmacology, endocrinology</td>
<td>2nd round, user testing, validity &amp; reliability testing</td>
<td>Yes</td>
<td>10</td>
</tr>
<tr>
<td>Jerry A Krishnan</td>
<td>USA</td>
<td>MD, PhD,</td>
<td>1st and 2nd round, Toronto Brainstorming, Validity &amp; reliability testing</td>
<td>Yes</td>
<td>6</td>
</tr>
<tr>
<td>Jodi Segal</td>
<td>USA</td>
<td>MD, Trialist, lecturer, systematic reviewer/observational study design</td>
<td>2nd round, User testing, validity &amp; reliability testing</td>
<td>Yes</td>
<td>15</td>
</tr>
<tr>
<td>Bridget Gaglio</td>
<td>USA</td>
<td>MD, PhD, MPH, Kaiser Permanente</td>
<td>Validity &amp; reliability testing, (Methodological work on PRECIS)</td>
<td>Yes</td>
<td>15</td>
</tr>
<tr>
<td>Dan Riddle</td>
<td>USA</td>
<td>P.T., Ph.D., FAPTA Methodological work on PRECIS</td>
<td>Brainstorming, Validity &amp; reliability testing</td>
<td>Yes</td>
<td>15</td>
</tr>
<tr>
<td>Fernando Althabe</td>
<td>Argentina</td>
<td>MD - Obstetrician, Trialist</td>
<td>Validity &amp; reliability testing</td>
<td>Yes</td>
<td>15</td>
</tr>
<tr>
<td>Hans Hobbelen</td>
<td>The Netherlands</td>
<td>PhD, Physiotherapist, Trialist, Lecturer</td>
<td>Validity &amp; reliability testing</td>
<td>Yes</td>
<td>10</td>
</tr>
<tr>
<td>Marion Campbell</td>
<td>UK</td>
<td>Trialist, Statistician, Head of Department Health Services Research, University of Aberdeen</td>
<td>Validity &amp; reliability testing</td>
<td>Yes</td>
<td>15</td>
</tr>
<tr>
<td>Peter Tugwell</td>
<td>Canada</td>
<td>MD, Lecturer, Editor Journal of Clinical Epidemiology</td>
<td>Validity &amp; reliability testing</td>
<td>Yes</td>
<td>5</td>
</tr>
<tr>
<td>Dan Bratton</td>
<td>UK</td>
<td>Statistician Medical Research</td>
<td>Validity &amp; reliability testing</td>
<td>Yes</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Name</td>
<td>Country</td>
<td>Title</td>
<td>Responsibilities</td>
<td>Participated in</td>
</tr>
<tr>
<td>---</td>
<td>--------------------</td>
<td>---------</td>
<td>--------------------------------</td>
<td>-----------------------------------------------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>13</td>
<td>Claudia Witt</td>
<td>Germany</td>
<td>MD, Trialist, systematic reviewer</td>
<td>Validity &amp; reliability testing</td>
<td>Yes</td>
</tr>
<tr>
<td>14</td>
<td>Dan Steinfort</td>
<td>Australia</td>
<td>MD, PhD</td>
<td>Validity &amp; reliability testing</td>
<td>Yes</td>
</tr>
<tr>
<td>15</td>
<td>Sally Hopewell</td>
<td>UK</td>
<td>PhD, RGN, Methodologist</td>
<td>Validity &amp; reliability testing</td>
<td>Yes</td>
</tr>
<tr>
<td>16</td>
<td>Gordon Forbes</td>
<td>UK</td>
<td>Research assistant, Statistician</td>
<td>Validity &amp; reliability testing, methodological testing of PRECIS-2</td>
<td>Yes</td>
</tr>
<tr>
<td>17</td>
<td>Russ Glasgow</td>
<td>USA</td>
<td>Clinical psychologist</td>
<td>methodological work on PRECIS</td>
<td>Yes</td>
</tr>
<tr>
<td>18</td>
<td>Doug Altman</td>
<td>UK</td>
<td>Statistician, Trialist, Methodologist</td>
<td>Validity &amp; reliability testing</td>
<td>Yes</td>
</tr>
<tr>
<td>19</td>
<td>An-Wen Chan</td>
<td>Canada</td>
<td>MD DPhil FRCPC Trial methodology, SPIRIT guidance for protocol publication</td>
<td>Brainstorming, Validity &amp; reliability testing</td>
<td>Yes</td>
</tr>
<tr>
<td>20</td>
<td>Clive Adams</td>
<td>UK</td>
<td>Mental health, trialists, systematic reviewer</td>
<td>Methodological testing of PRECIS</td>
<td>No</td>
</tr>
<tr>
<td>21</td>
<td>Mark Spigt</td>
<td>UK</td>
<td>PhD researcher public health, general practice</td>
<td>Methodological testing of PRECIS</td>
<td>No</td>
</tr>
<tr>
<td>22</td>
<td>Stan Shapiro</td>
<td>USA</td>
<td>MD, Editor Clinical Trials</td>
<td>Brainstorming (web)</td>
<td>No</td>
</tr>
<tr>
<td>23</td>
<td>David Moher</td>
<td>Canada</td>
<td>CONSORT, SPIRIT,</td>
<td>Brainstorming (web)</td>
<td>No</td>
</tr>
<tr>
<td>24</td>
<td>Eric Johnson</td>
<td>USA</td>
<td>Researcher, PhD</td>
<td>Delphi, Brainstorming (web)</td>
<td>No</td>
</tr>
<tr>
<td>25</td>
<td>Peter Selby</td>
<td>Canada</td>
<td>Addictions - MD, trialist</td>
<td>1st round Delphi, Brainstorming, Methodological testing of PRECIS</td>
<td>No</td>
</tr>
<tr>
<td>26</td>
<td>John Powers</td>
<td>USA</td>
<td>MD, trialist</td>
<td>1st round Delphi, Brainstorming</td>
<td>No</td>
</tr>
<tr>
<td>27</td>
<td>Clare Relton</td>
<td>UK</td>
<td>MD, trialist</td>
<td>1st and 2nd round Delphi, user testing</td>
<td>No</td>
</tr>
<tr>
<td>28</td>
<td>Peter Hall</td>
<td>UK</td>
<td>Oncologist - MD, Trialist, Economic analysis</td>
<td>User testing</td>
<td>No</td>
</tr>
<tr>
<td>29</td>
<td>Christian Gold</td>
<td>Norway</td>
<td>Music Therapist, Trialist</td>
<td>Delphi, User testing,</td>
<td>No</td>
</tr>
<tr>
<td>30</td>
<td>Louis D Fiore</td>
<td>USA</td>
<td>Internal medicine - MD, Trialist, epidemiologist</td>
<td>1st and 2nd round Delphi, User testing</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Name</td>
<td>Country</td>
<td>Role/Title</td>
<td>Delphi Round</td>
<td>Other Details</td>
</tr>
<tr>
<td>----</td>
<td>------------------</td>
<td>---------</td>
<td>---------------------------------</td>
<td>--------------</td>
<td>--------------------------------------------------------------</td>
</tr>
<tr>
<td>31</td>
<td>Grace Thompson</td>
<td>Australia</td>
<td>Music Therapist, Tri alist, Researcher</td>
<td>2nd round Delphi, User testing</td>
<td>No</td>
</tr>
<tr>
<td>32</td>
<td>Kalipso Chalidou</td>
<td>UK</td>
<td>MD, PhD</td>
<td>Delphi, original PRECIS tool</td>
<td>No</td>
</tr>
<tr>
<td>33</td>
<td>Jean Raymond</td>
<td>Canada</td>
<td>Oncologist – MD, trialists</td>
<td>1st round Delphi</td>
<td>No</td>
</tr>
<tr>
<td>34</td>
<td>John Fletcher</td>
<td>Canada</td>
<td>MD, CMAJ</td>
<td>Published original PRECIS</td>
<td>No</td>
</tr>
<tr>
<td>35</td>
<td>Steven Goodman</td>
<td>USA</td>
<td>MD, Editor J Clinical Trials</td>
<td>Published articles by Dave Sackett on pragmatic/explanatory trials</td>
<td>No</td>
</tr>
</tbody>
</table>
Reasons for not participating included: eight invitees too busy (five trialists who had participated in the Delphi and initially stated they were interested in assisting with validity and reliability testing); lack of response from three Delphi participants and two invitations to experienced trialists. The response was very enthusiastic from those not able to assist due to work schedules, one stating “I look forward to seeing your final version of PRECIS-2 and using it in my teaching and research.”

Table 36 Different groups of 35 participants invited to participate in Validity and Reliability testing

<table>
<thead>
<tr>
<th>Group</th>
<th>No of participants invited</th>
<th>No of participants participated*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Trialists/Methodological interest</td>
<td>12</td>
<td>11</td>
</tr>
<tr>
<td>2 Toronto Brainstorming meeting</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>3 Brainstorming and Delphi</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>4 Delphi and user testing</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>5 Delphi only</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>6 Medical editors</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>7 Additional requests</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

*Not cumulative

Thus by 3rd December 2013, validity and reliability of PRECIS-2 had been tested by 19 international trialists from seven countries – USA (8), UK (3), Canada (3), The Netherlands (2), Argentina (1), Australia (1), Germany (1), who used PRECIS-2 to score a varied sample of 15 RCT protocols. Of these, 19 used PRECIS-2 to score up to 15 trials: seven of these raters scored 15, while 12 scored 10. There were no obvious differences between the two groups with regard to country, research area, or profession who completed 5, 10 or 15 trials (Table 35).

Exclusion criteria

Originally twenty two researchers who had been involved in the Modified Delphi stated that they would like to help with the validity and reliability testing (academic doctors, researchers). After work with some of this international group in user testing some of these participants who had offered to assist with validity and reliability testing were not selected. It was important in seeking to prevent bias that we had a group of raters who understood the work involved and would fairly assess the trial
protocols using PRECIS-2. It appeared that some individuals had extreme views on pragmatic and
explanatory trials and did not, for instance accept that there was a continuum of pragmatic-ness – to
them a trial was either pragmatic or explanatory, there was a dichotomy. Also for some of the
participants their English did not appear to be adequate to cope with the additional workload and so
it would be uncertain that they would be able to rigorously test the PRECIS-2 tool and understand the
process, protocols and information sheet.
Participants in Validity & Reliability testing (n = 19)
- Participants from Delphi (n = 4)
- Participants from User testing (n = 2)
- Toronto brainstorming (n = 4)
- PRECIS interest (n = 13)
*Not cumulative (n = 2 Delphi and Brainstorming)
(n = 2 Delphi and User testing)

Validity & Reliability testing

Figure 18 Flow diagram of participants in validity and reliability testing
Validity and reliability deadlines for participants

The deadlines were changed from 1st November to 12th November 2013 and then 28th November 2013 to allow adequate sample size for rating. However a rating of five trials by a very experienced trialist and statistician who had sent several e-mails confirming his interest in the validity and reliability testing was accepted 3rd December 2013.

Validity and reliability incentives for participants

Only one researcher who had previously undertaken methodological work on PRECIS stated that the sum of money offered in the invitation was inadequate for the work involved. Some of the initial raters declined payment so we were able to increase the value of the incentive to those who had yet to finish validity and reliability testing to £200. In the end we offered payments to 10 participants though two subsequently donated money to the PhD project and were happy just to be acknowledged in publications. We did not offer an incentive to those who had only rated five trials or less.

Recruitment rate for the validity and reliability study

Figure 19 indicates the recruitment rate to the validity and reliability testing, there was initial interest with raters saying yes they could help but later having to retract their offer. Eight people were too busy (Some raters said they could rate all 15 trials but later found that due to work commitments they were only able to do 5 or 10 trials.
Figure 19 Recruitment rate for PRECIS-2 validation

Time to rate batch of five trials

The shortest time to return ten trials was six days for the final rater invited to participate but he was undertaking methodological work on PRECIS-2 and a highly motivated statistician. From the original raters invited on the 24th September two American researchers, a doctor and trialist who taught using PRECIS and another research scientist who had undertaken methodological work using PRECIS took six days to rate 15 trials, alongside their usual work load. In the invitation e-mail we had suggested the time to rate trials was but one rater stated “Maybe you scared people by telling them it takes 1.5 hours to do 5 – it doesn’t.”

Results format for PRECIS-2 scores

It was suggested to raters that they could send their PRECIS-2 scores in a table or wheel format - 14 raters used a table which prompted the rationale for their decision (example Table 37 Example of scoring by one of the raters for the ASHA protocol using the Table (FA)Table 37) with only one rater using a PRECIS-2 wheel to score trials (example –Figure 20) which was then scanned in and sent via e-mail. The rationale for scoring was completed for every domain in every trial by some raters or occasionally if a rater felt justification for a decision was necessary. Two raters sent their scores using
e-mail with reasons (if they considered these to be important) along with additional comments so did not send completed tables as attachments. Another two raters created an excel sheet to submit their results.
Table 37 Example of scoring by one of the raters for the ASHA protocol using the Table (FA)

<table>
<thead>
<tr>
<th>Domain</th>
<th>Score</th>
<th>Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eligibility Criteria</td>
<td>5</td>
<td>Only exclusions are women who migrate out of the study area (estimated in 1.3%)</td>
</tr>
<tr>
<td>Recruitment Path</td>
<td>4</td>
<td>In this cluster trial there are 2 different recruitments: 1) for the intervention at the intervention clusters (ASHAs as facilitators and then women for the groups); and 2) women having births during the study period in all clusters. For the intervention recruitment, I give a 5 (very pragmatic) For data collection I give a 3, as a special strategy to identify all women having deliveries was necessary. It is clear that this is the only way to to the study. However this can only be done for the study.</td>
</tr>
<tr>
<td>Setting</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Organisation intervention</td>
<td>4</td>
<td>The intervention to women receives a 5. The training to Ashas receives a 3, because incentives are necessary to sustain their activities. The protocol clearly includes the incentives as part of the intervention, so a 4 can also be given.</td>
</tr>
<tr>
<td>Flex of experimental intervention – Delivery</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Flex of experimental intervention – Adherence</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Follow up</td>
<td>5</td>
<td>For outcome assessment, women are only visited after delivery</td>
</tr>
<tr>
<td>Outcome</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Analysis</td>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>
Figure 20  Example of scoring by one of the raters for the IST-3 protocol [157] using the PRECIS-2 wheel (JS)
Duplicate data entry
There was complete agreement between two independent data entry people for the validity and reliability study.

Range of scores
As described previously (see p74 Methods – statistical measure – Inter-rater reliability) we used the ICC to determine consistency in scoring each domain using the Average Measure to calculate the ICC for all raters scoring the domain. Scores of average ICC less than 0.5 indicate simply chance or less than chance that raters agree or disagree. ICC of 0.7 is generally accepted as indicating that there is good agreement with Land and Koch indicating 0.61-0.80 is substantial agreement (Figure 21 Agreement benchmarks for the evaluation of observed $k$ values after Landis and Koch 1977 [166]). Dunn suggests any rating of agreement is subjective [167] but these ratings have be acknowledged as standards and can be used to describe consistency in scoring.

<table>
<thead>
<tr>
<th>$k$-statistic</th>
<th>Strength of agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>Poor</td>
</tr>
<tr>
<td>0.01-0.20</td>
<td>Slight</td>
</tr>
<tr>
<td>0.21-0.40</td>
<td>Fair</td>
</tr>
<tr>
<td>0.41-0.60</td>
<td>Moderate</td>
</tr>
<tr>
<td>0.61-0.80</td>
<td>Substantial</td>
</tr>
<tr>
<td>0.81-1.00</td>
<td>Almost perfect</td>
</tr>
</tbody>
</table>

Figure 21 Agreement benchmarks for the evaluation of observed $k$ values

Inter-rater reliability
Results are presented below in Table 38 for each of the PRECIS-2 domains: Eligibility, Recruitment, Setting, Organisation, Flexibility (delivery), Flexibility (adherence), Follow-up, Primary outcome,
Primary analysis. As we were unable to gather PRECIS-2 scores from 19 raters for all 15 trials, results are presented in Table 7.13 for batches of 15, 10 and 5 trial protocols with different numbers of raters and missing values were either replaced by “3” – chosen as this indicated uncertainty and equally pragmatic/explanatory or randomly-generated values between “1” and “5”. As we had a complete set of five trials scored by 18 raters using PRECIS-2 these will be detailed below with qualitative data on reasons for not scoring a trial domain. Further information on the other sets is available in the Appendix for Chapter 7.

Reasons given for missing data
Some of the raters did not score particular domains in trials 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. Many of the imputed values are due to “lack of time” and whole trials not being scored.
The ICC PRECIS-2 scores using randomly-generated values between 1 and 5 for missing scores (Table 38 Overall results for Inter-rater reliability for 9 PRECIS-2 domains including sensitivity analysis) for the nine different domains indicated that all of the domains except for one (Flexibility experimental intervention adherence) were reasonable or good based on the best estimate of the ICC and confidence intervals. The Eligibility domain had the best consistency with an ICC of 0.84 with tight

<table>
<thead>
<tr>
<th>Domain</th>
<th>Number of trials, raters</th>
<th>No. Imputed values = 3 * (%)</th>
<th>Intraclass Correlation</th>
<th>95% Confidence Interval</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eligibility</td>
<td>15, 7</td>
<td>1 (0.74)</td>
<td>0.88</td>
<td>0.75 - 0.95</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>10, 12</td>
<td>2 (1.67)</td>
<td>0.89</td>
<td>0.76 - 0.97</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>5, 18</td>
<td>4 (4.4%)</td>
<td>0.94</td>
<td>0.81 - 0.99</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Recruitment</td>
<td>15, 7</td>
<td>1 (0.74)</td>
<td>0.59</td>
<td>0.18 - 0.84</td>
<td>0.009</td>
</tr>
<tr>
<td></td>
<td>10, 12</td>
<td>2 (1.67)</td>
<td>0.60</td>
<td>0.10 - 0.88</td>
<td>0.012</td>
</tr>
<tr>
<td></td>
<td>5, 18</td>
<td>4 (4.4)</td>
<td>0.83</td>
<td>0.50 - 0.98</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Setting</td>
<td>15, 7</td>
<td>1 (0.95)</td>
<td>0.80</td>
<td>0.60 - 0.92</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>10, 12</td>
<td>2 (1.67)</td>
<td>0.80</td>
<td>0.56 - 0.94</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>5, 18</td>
<td>5 (4.4)</td>
<td>0.92</td>
<td>0.76 - 0.99</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Organisation</td>
<td>15, 7</td>
<td>2 (1.99)</td>
<td>0.72</td>
<td>0.44 - 0.89</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>10, 12</td>
<td>9 (7.5)</td>
<td>0.83</td>
<td>0.61 - 0.95</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>5, 18</td>
<td>6 (5.55)</td>
<td>0.75</td>
<td>0.25 - 0.97</td>
<td>0.006</td>
</tr>
<tr>
<td>Flex Exp Int Deliv</td>
<td>15, 7</td>
<td>3 (3.33)</td>
<td>0.74</td>
<td>0.47 - 0.90</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>10, 12</td>
<td>6 (6.67)</td>
<td>0.85</td>
<td>0.67 - 0.96</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>5, 18</td>
<td>6 (6.67)</td>
<td>0.92</td>
<td>0.75 - 0.99</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Flex Exp Int Adher</td>
<td>15, 5</td>
<td>0</td>
<td>0.50</td>
<td>-0.06 - 0.81</td>
<td>0.035</td>
</tr>
<tr>
<td></td>
<td>15, 7</td>
<td>8 (7.62)</td>
<td>0.24</td>
<td>-0.54 - 0.70</td>
<td>0.218</td>
</tr>
<tr>
<td></td>
<td>10, 12</td>
<td>17 (15.74)</td>
<td>0.57</td>
<td>0.04 - 0.88</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>5, 18</td>
<td>18 (15.56)</td>
<td>0.72</td>
<td>0.16 - 0.97</td>
<td>0.01</td>
</tr>
<tr>
<td>Follow-up</td>
<td>15, 7</td>
<td>1 (1.11)</td>
<td>0.60</td>
<td>0.18 - 0.84</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>10, 12</td>
<td>8 (8.89)</td>
<td>0.80</td>
<td>0.55 - 0.94</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>5, 18</td>
<td>6 (6.67)</td>
<td>0.85</td>
<td>0.55 - 0.98</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Primary outcome</td>
<td>15, 7</td>
<td>0</td>
<td>0.44</td>
<td>-0.13 - 0.78</td>
<td>0.054</td>
</tr>
<tr>
<td></td>
<td>10, 12</td>
<td>1 (1.11)</td>
<td>0.66</td>
<td>0.24 - 0.900</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td>5, 18</td>
<td>3 (3.33)</td>
<td>0.84</td>
<td>0.54 - 0.98</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Primary analysis</td>
<td>15, 7</td>
<td>0</td>
<td>0.67</td>
<td>0.32 - 0.87</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td></td>
<td>10, 12</td>
<td>3 (3.33)</td>
<td>0.73</td>
<td>0.39 - 0.92</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>5, 18</td>
<td>5 (5.55)</td>
<td>0.83</td>
<td>0.50 - 0.98</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

*PRECIS-2 score = 3*
confidence intervals 95% CI (0.69 to 0.94). Recruitment had wide confidence intervals 95% CI (0.20 to 0.84) but estimate reasonable ICC 0.58 given that not all the information was in Recruitment to score this domain so some raters guessed. The ICC score for Recruitment using randomly generated values was also statistically significant with a value of P = 0.006. The Setting domain had the next best agreement ICC 0.79 with reasonably tight confidence intervals 95% CI (0.60 to 0.92). The Organisation domain was similar ICC 0.72 (95% CI 0.46 to 0.89) although there was more imputed data. The Flexibility - delivery ICC 0.80 (95% CI 0.62-0.92) and Follow-up ICC 0.71 (95% CI 0.44 to 0.88) were also reasonable although there was more imputed data for the 12 and 18 raters looking at 10 and 5 trials. Primary analysis, which was largely unchanged from the previous PRECIS tool, had a reasonable ICC although the confidence interval was quite large ICC 0.67 (95% CI 0.37 to 0.84). Flexibility–adherence ICC 0.54 (95% CI 0.12 to 0.82) was not statistically significant (0.014) and Primary outcome ICC 0.68 (95% CI 0.38 to 0.87) domains had negative values in the lower confidence intervals.

Table 39 Rater scoring using randomly generated values (1 to 5) of PRECIS-2 domains using Intraclass coefficient (ICC) for 15 trials, 19 raters

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

*approx. 93% of missing data is due to TRIALS NOT being scored at all by raters e.g. 87/94 missing values

Discriminant validity results (Method 1) Agreement for Pragmatism scores for ST and KL was 80% (12/15) (Table 40). However the three trials that had different scores i.e. “1” instead of “0”, ST had scored the trials as equally pragmatic/explanatory (i.e. scored as “3”) so it could have been possible to have 100%
agreement if ST had decided to score in the opposite direction. These trials were however still included in the analysis using the overall pragmatism ratings of KL.

Table 40: Pragmatism scoring results for KL and ST of 15 trials

<table>
<thead>
<tr>
<th>Trial</th>
<th>Name</th>
<th>Pragmatism ST</th>
<th>Pragmatism ST 1-5 scale</th>
<th>Pragmatism KL</th>
<th>KL 1-5 scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ASHA</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>CHAT</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>Cortico injections</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>EAGLE</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>ECSSIT</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>Exercise therapy</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>FLOW</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>HANDS</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>IMR</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>10</td>
<td>IST3</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>NESSTAC</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>PRO micro</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>13</td>
<td>PROUD</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>14</td>
<td>REFINE</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>15</td>
<td>TASC</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 41: Discriminant validity measured using Area Under the ROC curves (AUROC)

<table>
<thead>
<tr>
<th>Domains</th>
<th>AUROC</th>
<th>95% Confidence intervals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary Outcome</td>
<td>.750</td>
<td>.493-.1000</td>
</tr>
<tr>
<td>Follow-up</td>
<td>.732</td>
<td>.475-.989</td>
</tr>
<tr>
<td>Primary analysis</td>
<td>.723</td>
<td>.451-.995</td>
</tr>
<tr>
<td>Flexibility delivery</td>
<td>.714</td>
<td>.435-.994</td>
</tr>
<tr>
<td>Eligibility</td>
<td>.625</td>
<td>.332-.918</td>
</tr>
<tr>
<td>Recruitment</td>
<td>.616</td>
<td>.316-.916</td>
</tr>
<tr>
<td>Flexibility adherence</td>
<td>.598</td>
<td>.302-.894</td>
</tr>
<tr>
<td>Setting</td>
<td>.589</td>
<td>.264-.915</td>
</tr>
<tr>
<td>Organisation</td>
<td>.571</td>
<td>.272-.871</td>
</tr>
</tbody>
</table>

These AUROC values (Table 41) are a numerical summary of the ROC curves displayed later (Figures 6.8-6.16). These values have been placed in order of discriminative ability for determining pragmatism in a trial. A score of 1 would be the ideal score and indicate that the PRECIS-2 domain outcome was
perfect at discriminating between pragmatic and explanatory. Randomly determined pragmatism would be 0.5. The results for all PRECIS-2 domains are greater than 0.5 although some are close. *Primary outcome* is the single variable that is most likely to discriminate how pragmatic a trial is based on this data – AUROC 0.750. Then in order of discriminating pragmatism: *Follow-up* 0.732, *Primary analysis* 0.723, *Flexibility (delivery)* 0.714, *Eligibility* 0.625, *Recruitment* 0.616, *Flexibility adherence* 0.598, *Setting* 0.589, *Organisation* 0.571. So the higher scores indicate better at predicting pragmatism. While the probability for all domains for predicting pragmatism is greater than 0.5, the confidence intervals are not tight and not significant.

Considering the blue lines of the ROC curves (Figure 22, Figure 23, Figure 24, Figure 25, Figure 26, Figure 27, Figure 28, Figure 29, Figure 30) nearly all of the values are above the green diagonal line though *Setting* and *Recruitment* are below the green line in some places. A perfect ROC curve would be a smooth curve. However, in the ROC curves for the nine PRECIS-2 domains there are also diagonal segments in the ROC curves – this is “*the average of the two most extreme paths and tends to underestimate the plot for diagnostically accurate test*” (Zweig and Campbell, 1993, p. 566)... the distance and angle of this diagonal line depend on the numbers of ties within Pragmatic -1 and Explanatory- 0 groups.” [168]. It is important to remember that each domain in the Discriminant variability is composed of 15 trials with up to 19 eligibility score values. So for instance in the *Eligibility* curve there are four diagonal lines; these correspond to some raters deciding that a trial is pragmatic instead of explanatory, or vice versa, instead of *all* raters deciding a trial is pragmatic or explanatory (Table 42 Rater scores for pragmatism for the Eligibility domain for 15 trialsTable 42). So in trial 1 – total agreement in pragmatism 18 scores of “5” and “4” and also in trial 14 with six scores of “5” and one score of “4” so for these trials all raters agreed pragmatic. For the remaining 13 trials, the scores were not unanimous with some raters stating trials pragmatic and others explanatory. Note discriminant validity analysis is only pragmatic or explanatory.
All the domains had trials which contained scores indicating raters disagreed with the assessment of pragmatism so there are diagonals in the AUROC curves. The PRECIS-2 domain which had only two diagonals is the Primary Outcome domain which is why it looks more like a curve and has the highest AUROC score. Considering the actual rater scores for the Primary Outcome domain (Table 43), again in trial 1 – total agreement with 16 scores of “5” and two scores of “4” but as can be seen in the other trial scores there is far less disagreement in pragmatism rating for the these trials too.

<table>
<thead>
<tr>
<th>Trial number</th>
<th>Pragmatism “4” or “5”</th>
<th>Explanatory “1” or “2”</th>
<th>Equally explanatory/pragmatic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>12</td>
<td>6</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>6</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Trial number</th>
<th>Pragmatism “4” or “5”</th>
<th>Explanatory “1” or “2”</th>
<th>Equally explanatory/pragmatic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>14</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>11</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>9</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>6</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>9</td>
<td>10</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>12</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>6</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>5</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>5</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>14</td>
<td>5</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>15</td>
<td>5</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Figure 22 Eligibility

Figure 23 Recruitment

Figure 24 Setting

Figure 25 Organisation

Figure 26 Flexi. Delivery

Figure 27 Flex. Adherence
Figure 28 Follow up

Figure 29 Primary Outcome

Figure 30 Primary Analysis
Discriminant validity results (Method 2)

Pragmatic trials had less frequent use of “1” or “2” in their domain scores (“1” = very explanatory or “2” = rather explanatory) (Table 42, Table 43). The number of explanatory choices was calculated by summation of the number of times score of “1” or “2” for each domain for each of the 15 trials divided by the number of raters providing a score. The threshold below which a trial was classed as pragmatic was set at “1.4” because the lowest total score for an explanatory trial KL and ST agreed on was “1.5” and the highest total score was for a pragmatic trial that KL and ST agreed on was “1.4”. This can be seen in Table 44 with two blocks of blue indicating a match for global pragmatism and low average score for the number of domains scoring “1” or “2”. Two of the three where the colours do not match are where KL or ST chose different explanatory versus pragmatic descriptions of trials. Only trial 9 does not match a low score (threshold 1.4) with the global pragmatism score given by KL and ST.

Table 44 global pragmatism scores considering number of explanatory choices (out of 9) for a trial.

<table>
<thead>
<tr>
<th></th>
<th>&lt;=2 (overall)</th>
<th>Global pragmatism</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>0.83</td>
<td>Pragmatic</td>
</tr>
<tr>
<td>2.</td>
<td>0.50</td>
<td>Pragmatic</td>
</tr>
<tr>
<td>3.</td>
<td>2.56</td>
<td>Explanatory</td>
</tr>
<tr>
<td>4.</td>
<td>1.29</td>
<td>Pragmatic</td>
</tr>
<tr>
<td>5.</td>
<td>2.11</td>
<td>Explanatory</td>
</tr>
<tr>
<td>6.</td>
<td>2.00</td>
<td>Pragmatic</td>
</tr>
<tr>
<td>7.</td>
<td>1.50</td>
<td>Explanatory</td>
</tr>
<tr>
<td>8.</td>
<td>3.15</td>
<td>Explanatory</td>
</tr>
<tr>
<td>9.</td>
<td>2.85</td>
<td>Pragmatic</td>
</tr>
<tr>
<td>10.</td>
<td>1.36</td>
<td>Pragmatic</td>
</tr>
<tr>
<td>11.</td>
<td>2.29</td>
<td>Explanatory</td>
</tr>
<tr>
<td>12.</td>
<td>1.29</td>
<td>Pragmatic</td>
</tr>
<tr>
<td>13.</td>
<td>1.14</td>
<td>Explanatory</td>
</tr>
<tr>
<td>14.</td>
<td>0.57</td>
<td>Pragmatic</td>
</tr>
<tr>
<td>15.</td>
<td>3.43</td>
<td>Explanatory</td>
</tr>
</tbody>
</table>

Threshold 1.4
More explanatory trials generally have low scores in ALL domains (Table 45), but there does not appear to be any one particular domain particularly influencing an explanatory global rating. Generally speaking pragmatic trials have explanatory scores in fewer domains.

Table 45 Domains with low scores in pragmatic (white) and explanatory (blue) trials.

<table>
<thead>
<tr>
<th>Trial</th>
<th>Elig</th>
<th>Recruit</th>
<th>Setting</th>
<th>Org</th>
<th>Flex D</th>
<th>Flex Adh</th>
<th>Foll-up</th>
<th>Outcome</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3 E</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>5 E</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>6</td>
<td>2</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 **</td>
<td>11</td>
<td>11</td>
<td>6</td>
<td>4</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 E</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>6</td>
<td>2</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 **</td>
<td>8</td>
<td>1</td>
<td>4</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>9 *</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>10</td>
<td>9</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11 E</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14**</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15 E</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

* odd trial that did not agree with global pragmatism rating for ST/KL
** trials that KL and ST scored differently that were equally pragmatic/explanatory.

Discussion
Unlike with the development of the original PRECIS, this chapter formally tested PRECIS 2’s validity and reliability at the development stage. In addition, unlike other studies that have looked at PRECIS’ validity and reliability (Chapter 2, Table 3), our study was large enough to make meaningful statements about these properties with a reasonable sample size and power (see sample size information).

Ensuring adequate sample size for validity and reliability testing proved to be difficult as expected, which is why previous published studies have involved fewer raters. Many of the raters were keen to help but were over committed and had to drop out, though keen to see the results. By sending trial protocols in batches of five, at least we were able to include some independent rating of trial protocols rather than none for individual raters. However, our initial concerns about getting people to do 15 trials were justified; even though everything was done to make this process as easy as possible.
This study set out to improve PRECIS through wide, international collaboration. I am confident that the new tool addresses trial design domains that trialists and methodologists will recognise as relevant to the match between trial design and the intended use to which the results will be put. I have consulted over 80 individuals through the Delphi, Brainstorming, User testing and the Validity and Reliability testing that I believe that the PRECIS-2 tool has face validity. Nobody has disagreed with this statement.

Our quantitative reliability and validity work found that PRECIS-2 has generally good inter-rater reliability across the nine domains with 7/9 ICCs over 0.67 and reasonable discriminant validity with all nine domains showing better than chance relationships with subjective global ratings of pragmatism. It is important to note that the ideal way to use PRECIS-2 is not retrospective application to unfamiliar trials (as done here) but at the design stage of a user’s own trials. Used in this way, raters have all the relevant details and expertise to score the nine domains fully, which is not always the case when reading others’ trial protocols. That inter-rater reliability was still good, and discriminant validity reasonable albeit with much uncertainty, even when PRECIS-2 was used by researchers unconnected with the trial being scored is encouraging. This, and the very wide consultation, means PRECIS-2 has a much stronger foundation than the original PRECIS.

**Rater differences**

The group of raters included in the validity and reliability work had a range of experience (see Table 35), some had been involved in helping with the original studies that had used the original PRECIS tool (3 – Bratton, Riddle, Witt), and some had been involved in the different phases of the development of PRECIS-2 (9 raters) with some having had no involvement in the Delphi, Brainstorming or User testing (10 raters) (see Flow chart – Figure 6). There appeared to be no obvious difference in the five additional raters who used PRECIS-2 for 10 trials when compared to the seven raters who used PRECIS-2 for 15 trials based on involvement in this study, their research area or the country working in (see Table 35). The 12 raters who only rated 10 trials stated that time was the only limiting factor for not completing
a further five trials. Six of the raters came from USA, though only three had managed to do 10 or 15 trials – the others rated only five trials; five were from the UK – four completed 10 or 15 trials, three from Canada, two The Netherlands, one rater from Germany, one rater from Australia and finally there was one medical trialist from Uruguay.

The participants invited to participate were believed to be potential users of PRECIS-2. A broader sample of participants had been sampled during the Delphi and User testing where extreme views had been given indicating lack of interest or opposition to the PRECIS-2 tool. In the validity and reliability testing, however, I asked people to participate who I thought would use the tool. There did not appear to be much point consulting those who would not use the tool or were not sufficiently experienced in trial design to understand what the domains entailed.

**Trial differences**

As we had *not* managed to get a complete set of 15 raters who completed 15 trials and only 12 had done 10 with a total of 7 doing 15 we wanted to see if the trials in the third batch were significantly different to the first 10. Again, as in the raters, there did not appear to be any particular differences in the final batch of trials compared to the others. Neither did there seem to be particular problems with rating noted in comments from raters for this final bath, just considering the final five trials - three were scored as pragmatic by KL and two by ST. Although the 15 trials were on a continuum of pragmatism, the trials were generally more pragmatic (ST – 8; KL – 9 using an ordinal rating of pragmatic or explanatory) with fewer explanatory trials. By nature explanatory trial protocols tend to be longer and we had chosen to reduce the workload for raters by excluding trials over 60 pages. This was probably the right decision as we still did not manage to engage 15 raters to complete 15 trials. However, we are confident in the results, this is the largest study; 19 raters used PRECIS-2 to score five trials with seven raters doing all 15. Even in this worst case scenario of retrospective use of PRECIS-2 it performed well. It was also heartening that there was so much enthusiasm for this reliability and validity study of PRECIS-2, even by those who were unable to participate.
Domain descriptions

There were problems if raters did not have the content expertise to use PRECIS-2 to rate a trial, in particular judging usual care: “I had a very difficult time with trials that are far afield of my content area. Had to leave several cells blank because I simply don’t have any clue what “usual care” might look like.” Variation in scoring for “usual care” was also highlighted in Protocol 12 [156] which had two scores of “2” with raters saying not usual care outcome but two raters scoring “5” and two raters scoring “4” and one rater “4.5”. The primary outcome for this trial was incidence of HIV infection (measured by serological confirmation) i.e. a surrogate measure. Thus the difference in scores seemed “odd” and were certainly not in agreement with five out of seven of the raters that stated rather or very pragmatic outcome. The low scores saying “daily diary far removed from usual care” referred to random sample of 1000 women completed pictorial diaries while the most accurate outcome assessment would appear to be the comment “very tight outcome serologically confirmed”. It may be reading quickly not all of the raters have accurately assessed the outcome or been familiar with usual care to assess this primary outcome accurately.

Insufficient information

There was inconsistency if the rater could not find information to score a domain. Different methods had been used by raters including: guessing based on the information available for the other domains, using 3 = equally pragmatic/explanatory or using “1” very explanatory. One tester (BG) who used “1” and did not “guess” score had been involved in publishing using PRECIS for assessing three clinical trials on weight loss - Practice-Based Opportunities for Weight Reduction (POWER) Trials. A “0–4” scale had been used, with “0” being completely explanatory and “4” being completely pragmatic.

We had decided not to give advice on which was the most popular way to rate a trial protocol that did not have adequate information but to see what emerged from the data. However, as the purpose of PRECIS-2 is for assisting in trial design with the whole trial team present, this is only an issue for validity and reliability testing. For surgical trials, raters determined that there no adherence issues for patients so left blank, others wrote N/A, whereas others scored 1 very explanatory or 5 very pragmatic – full
range of compliance demonstrating that clearer guidance is needed for these types of trials using PRECIS-2. The information sheet for using PRECIS-2 has thus been edited and now states “During the design process, if there is uncertainty over how explanatory or pragmatic a domain is, then we suggest the score for this domain should be left blank. This will then highlight uncertainty and encourage discussion. If PRECIS-2 is used to consider how pragmatic included trials are in systematic reviews then a score of 3 may be chosen if there is inadequate information. This is different to the “3 = equally pragmatic/explanatory”.

The score of “3” was suggested for scoring trial domains included in systematic reviews (if insufficient information) to assist cumulative scoring of domains for an individual trial to give an overall indication of pragmatism. If there is no domain score this will artificially lower the overall trial score for 9 domains; “as 3” is the midway score it was believed this would be an objective score if inadequate information. It is also, of course, entirely possible that a trial is neither one approach nor the other for a particular domain, meaning a score if 3 is the only correct score to give.

It is worth noting there was also a problem using the original PRECIS tool for the domain “compliance”, the original Thorpe article states for the NASCET example “The experimental intervention in NASCET was offering a one-time operation. Because the 50% probability of operation was clearly stated in the original consent documents, patients who did not want surgery were unlikely to enter the trial (only 0.3% of admitted patients randomized to the operation refused it). This is a prophylactic strategy for achieving compliance and is thus, an explanatory approach.”

**Possible explanation for domain scoring difficulties**

*Flexibility – adherence.* There were four trial protocols [157, 161, 163, 169] from the 15 SPIRIT trials included in the sample for validity and reliability testing in which the patient had indirect input in treatment delivery and thus the domain *Flexibility of experimental intervention – adherence* was not relevant. The intervention was given or not.
1. Mothers have caesarean sections no compliance issues involved in CS [161]

2. Patients have no say in irrigation of wounds from open fractures [163]

3. Patients no input in suture choice and no compliance issues [169]

4. Patients allocated to laser treatment by surgeon or lens removal operation, not clear if adjunct treatment though eye drops afterwards involve compliance or given by surgeon, probably former, but no information [157]

Not Applicable (N/A) was the rationale for several raters scoring for these trials so this was not really missing data, but others had scored “5” or indeed “1” for very explanatory. As patients had consented to the surgical intervention and were immobile and/or unconscious, they were not involved in adhering to a treatment, it is unclear the rationale behind these scores. There were no comments from raters to help determine why one rater scored “5” and another “1”. These surgical trials demonstrated that clearer information on how to use PRECIS-2 was needed.

Even in a trial that was not surgical, for instance trial protocol 9 [165] “Illness management and recovery (IMR) in Danish community mental health centres” there was conflicting rationale for scoring decisions. One trialist wrote “Adherence strictly monitored – all attendees registered. Compliance monitored” giving a score of “1” whereas the two raters that scored “5” said “monitored but not influenced by additional efforts” and “highly flexible withdrawal”. A general comment about the Flexibility (adherence) domain was made by another rater: “although mentioned in most study protocols often not enough information is given to judge on this”. This latter comment would not be a problem for trialists designing the trial as they would decide on how much monitoring and improvement in compliance they wished to carry out. It is clear, however that the information giving guidance to raters has to be improved to take account of surgical trials as this domain may not be applicable. The information sheet for using PRECIS-2 has been edited and now states “In some trials e.g. surgical trials where patients are being operated on or Intensive Care Unit trials where patients are
being given IV drug therapy, this domain is not applicable as there is no compliance issue after consent has been given, so this score should be left blank.”

**Primary outcome** – many raters found it difficult to determine if the primary outcome was of interest to the patient, unless the area was known to the raters and their particular area of expertise it was tricky to determine if it was relevant or not. Patient Centred Outcomes Research Institute PCORI [www.pcori.org](http://www.pcori.org) help trialists in this area to determine more useful and relevant outcomes that can be used in clinical trials, as well as the Core Outcome Measures in Effectiveness Trials (COMET) website (http://www.comet-initiative.org). If PRECIS-2 was being used, as it was intended, to design a trial then the difficulties of evaluating the value of the primary outcome to patients and practitioners would not be problematic.

**Comparators and multiple interventions**

One of the raters, a medical practitioner and trialist in Europe had difficulty with the domain - *Flexibility of the experimental intervention delivery*. In trial protocol 4 which compares the intervention of “early clear lens extraction with intraocular lens implantation” against the comparator or control of “stepped approach of a combination of laser iridotomy surgery and medical treatment” [157]. It was clear from the rationale for scoring that this rater would have liked to score both of the delivery methods and not just one of the interventions. In this particular trial, KL, ST had thought that “usual care” was the comparator. The comment from the rater “It is still a problem if the trial compares two or more interventions (see my rating on Azuara-Blanco [157]), which are different in Flexibility adherence, this needs guidance. Furthermore information is needed on how to handle adjunct treatment to usual care versus usual care only. Is usual care only “no” intervention or also rated as an intervention which would be always end up with 5 and might be totally different than the intervention? Thus making it difficult to decide on a number between 1 and 5.”
**Sensitivity analyses**

We used both imputing the value “3” (equally pragmatic/explanatory) and randomly generated values of “1” to “5” if there were no scores and raters had stated unable to score (or indeed raters had no time to score all the protocols). We chose to leave in scores that a few raters had inserted for a few trials but had stated in the rationale for scoring that they were uncertain, the information was unclear or felt there was insufficient data. We felt that because this only occurred in a few trials it would not have a significant effect on the scores for the different domains.

The value of “3” was selected for missing values as we believed it represented an objective score for equally pragmatic/explanatory (and was similar to the advice we gave to PRECIS-2 users if there was uncertainty). Some may consider use of this score “3” to artificially increase the ICC as correlation between missing values will become high. However, we believe this is a fair score to use and moreover there is not a significant difference when compared to running a sensitivity analysis using randomly generated values of “1” to “5”. E.g. Eligibility domain ICC 0.88 when using “3” compared to ICC 0.84 random values.

Using sensitivity analyses with randomly generated values “1” to “5” to calculate the ICC scores for 15 trials using 19 raters produced ICC scores for the different PRECIS-2 domains which were noticeably worse than the ICC scores for the five trials scored by all raters. These had three to six imputed values (apart from Flexibility Intervention (adherence) which had 18 imputed values – as stated earlier many said not relevant in the specific trials). For nearly all PRECIS-2 domains, increasing the number of raters increased the ICC so we believe that this trend suggests that our results are robust when all the ICC scores are considered for rating 5, 10 and 15 trials with 18, 12 and 7 raters. So for instance considering the Eligibility domain when 15 trials were scored by seven raters the ICC was 0.877, 10 trials scored by 12 raters ICC was 0.891 and five trials scored by 15 raters gave an ICC of 0.938. All Eligibility domain ICC results were statistically significant. This trend followed through to Primary Analysis so when 15
trials scored by seven raters ICC was 0.666, 10 trials scored by 12 raters ICC was 0.727, with five trials scored by 18 raters giving an ICC of 0.831. Like the Eligibility domain, the results were statistically significant.

Usual care
Difficulties also arose if there was lack of expertise for raters looking at trials outside their field so did not have the “usual care” knowledge. This would not be an issue for a trial team working with PRECIS-2 at the design stage. Some of the range in confidence intervals can be explained, in part at least, by this issue.

Tables or PRECIS-2 wheels
Raters were given the option of using the PRECIS-2 wheel or a table with the nine PRECIS-2 domains to score the trial protocol. Like the pilot, one rater used the wheel in the validity and reliability testing while the other raters used the table, usually giving rationale for scores. Thus the addition of the PRECIS-2 table to help raters consider the different domains and how they decide to score allows a transparency in the process which was not so clear in the original use of the tool. This suggestion from the Delphi appears to have been very well accepted during the validity and reliability testing.

Discriminant validity results (Method 1)
Considering the wide range of raters and trials that were scored, PRECIS-2 has stood up well to discriminant analysis. We believe this initial testing indicates that PRECIS-2 domains can be used to help trialists consider how pragmatic their trial domains are to varying degrees.

The data for the discriminant validity analysis, however, had limitations as all domains had raters who scored opposite to others or the overall pragmatic score was in a different direction. This is demonstrated well in the pragmatism rating by KL and ST for three very different trials (Table 7.18). However, when the median scores for all domains for trial 6 are looked at seven out of the nine PRECIS-2 of domains were rated with 3 equally explanatory/pragmatic but in discriminant analysis only two
choices are possible: more pragmatic or more explanatory. The occupational therapy rehabilitation trial, number 8, had three domains more explanatory, 3 domains equally pragmatic/explanatory and 3 domains more explanatory. Considering the surgical trial, number 13, all of the median scores indicated rather or very pragmatic so this was the only score in this trial (and the only trial) that did not agree with the majority scoring. A score like this in the normal use of PRECIS-2 would lead to discussion and then consensus with other raters. But all the scores were determined independently by raters and reaching consensus was not part of the validity and reliability testing of PRECIS-2. Overall, the opinions of KL and ST versus the opinion of other users is a limitation of this methodology. Ideally we would have liked a global assessment for pragmatism for the 15 trials. Nevertheless while it would have been a better approach, we do not believe the results would have been very different if the global assessments of all of the raters had been used. Our approach allowed us to reduce the work required of raters, important since it proved difficult enough to get raters to complete scoring of all 15 trials.

AUROC measurements were used to measure discriminant validity and was believed to be an acceptable way to analyse the data. The key problem is discriminant validity simply answers the question if the score of a domain can determine if a trial is explanatory or pragmatic so binary YES or NO. The problem arises when a trial is midway between explanatory and pragmatic. However in the PRECIS-2 tool have a Likert 1-5 scale from explanatory to pragmatic and discriminant validity cannot work if a trial is midway between explanatory and pragmatic.

**Discriminant validity results (Method 2)**

Yoong et al [69] determined global rating of explanatory, combined or pragmatic trial (Figure 7.16) in trials for a systematic review by splitting the 5 point score of 1-5 into explanatory 0-1.7, combined > 1.7 -2.2 and pragmatic >2.2-4.0 for all domains with data and dividing by number of domains with no missing data. We were unable to do that as the scores did not appear to reflect the overall pragmatism (Figure 31) indicating the individual domains were more important than the overall total score.
In attempting to deconstruct pragmatism by looking at explanatory scores of “1” or “2” for PRECIS-2 domains, it appears that explanatory scores can occur in any of the PRECIS-2 domains and tend to occur in more than five domains for a global explanatory rating. While there is some overlap with pragmatic global rating of trials, pragmatic trials generally have fewer domains with scores of “1” and “2” than scores “3” to “5”). The overlap does emphasise that there is a continuum and it is very hard to classify trials as either pragmatic or explanatory when nine domains are involved that may push the design of the trial towards being more pragmatic or more explanatory.

Using method 2 to consider discriminant validity for PRECIS-2 a threshold had to be selected to determine whether a trial was pragmatic or explanatory. The threshold below which a trial was classed as pragmatic was set at “1.4” because a the lowest total score for an explanatory trial KL and ST agreed on was “1.5” and the highest score for a pragmatic trial that KL and ST agreed on was “1.4”. But a different choice would lead to different result. It would be useful to apply the “1.4” threshold to a different set of raters and trials to see if this is generalizable.

**Discriminant validity methodological limitations**
The key weakness in this discriminant analysis is that there were median scores from only 15 trials; the model would have benefited from additional points in each domain to test its validity. However, each of the 15 median values was constructed with up to 19 scores from participants. As before, all data was included in the analysis, including scores from raters who said they had used “3” if uncertain or
“1” if uncertain and there was inadequate information and in surgical trials where domain *Flexibility (Adherence)* for patients undergoing surgery was not applicable.

Using method 2 in the discriminant validity study, we were unable to calculate a global rating to assess if PRECIS-2 could accurately discriminate trials of *varying* pragmatism as judged by the subjective rating by KL, ST on an explanatory to pragmatic basis. We were only able to determine if PRECIS-2 could discriminate between a pragmatic or explanatory trial. As there is a continuum from very explanatory to very pragmatic, this almost defeats the purpose of testing if PRECIS can detect pragmatism if only pragmatic or explanatory trial, particularly as three trials were assessed by KL and ST as being equally explanatory/pragmatic. In addition, PRECIS-2 (like its predecessor PRECIS) is composed of domains and each of these can be on a continuum of pragmatism from very explanatory to very pragmatic so deciding if a trial is pragmatic or explanatory is not the goal of PRECIS-2. Assisting trialists to consider each domain and determine how pragmatic or explanatory each domain in the trial design is the aim of the tool.

**Validity and reliability study strengths**

The tool Gartlehner developed to differentiate between efficacy and effectiveness studies was validated through a small sample size of two raters blinded to the type of trial and using 24 trials [37]. The authors state the seven criteria have face value due to the consensus input from the Directors of the 12 U.S. and Canadian Evidence-based Practice Centers and this relatively simple validation study [37, 116]. Their inter-rater agreement was 78.3% (kappa: 0.42) if six out of seven criteria were applied, sensitivity was 72%, specificity was 83% [37]. However, a sample size of two raters evaluating 24 trials is too small a sample size to demonstrate construct validity. The tool by Gartlehner also distinguishes between pragmatic or explanatory trials whereas the PRECIS-2 tool helps trialists consider where their trial is on the continuum of explanatory to pragmatic; few trials are completely pragmatic or explanatory. Also, in subsequent methodological studies using PRECIS, none of them had an adequate sample size of raters and trials to rate to demonstrate both face and construct validity. Our study is
bigger than previous studies, and although we had difficulty getting an adequate sample size of 15 raters use PRECIS-2 to score 15 trials we have managed to improve the sample size of previous studies.

One of the interesting issues that emerged from one of the methodological studies using PRECIS studies was that there appeared to be a bias for trialists rating their own trials to be more pragmatic than independent raters [73, 150]. Bratton recognised this and advised that independent committees with sufficient knowledge of the trial should undertake the PRECIS scoring as there may be a vested interested in suggesting a trial was more pragmatic and this would reduce any bias [58]. In selecting the trials to test PRECIS-2 validity and reliability we thus endeavoured not to select trials that validators had been involved in. However, we did inadvertently include the FLOW study [163] that one of the raters (DA) was currently working in as the chair of the Data monitoring committee so knew a great deal about this trial, interestingly his scores were more pragmatic overall when every domain was considered.

**Validity and reliability study limitations**

The reliability scores could perhaps have been higher and a few factors might have influenced this. Training was minimal; we used a three-page crib sheet with basic definitions on domains and instructions for use. We did not use examples or more extensive information sheets as we were concerned that reading extra materials would have increased the work load for PRECIS-2 raters. This would have made it harder to get agreement to assist with validity and reliability testing and make it more unlikely to get a large enough sample size. It is of note that the three (out of the 12) methodological studies on PRECIS which involved more training than other methodological studies had better agreement scores [67, 73, 150] and included web training PowerPoint on PRECIS development and the original PRECIS publication [65].

The treatment of missing values and the selection of “3” might appear to be a limitation in the PRECIS-2 validity and reliability work. The value of “3” was selected for missing values as we believed it
represented an objective score for equally pragmatic/explanatory (and was similar to the advice we gave to PRECIS-2 users if there was uncertainty). The score of “3” may however artificially increase the ICC as correlation between missing values will become high. We strongly believed this was a fair score to use but there is a difference for ICC and Rater scoring using “3” when compared to randomly generated values of “1” to “5”. For instance in the validity and reliability of the Follow up domain there were eight imputed values of “3” and ICC 0.80 for 10 trials and 12 raters and randomly generated imputations gave ICC of 0.707.

We also used two groups of people: trialists and researchers who were familiar with PRECIS but had not been involved in the development of PRECIS-2; and researchers who had been involved in PRECIS-2 development (Delphi or early User testing or both) so everyone was reasonably experienced in the ideas behind PRECIS so completely new users may need additional training.

Use of protocols instead of full published review articles was raised as an issue with one rater who had worked with the original PRECIS tool (physiotherapist). “Given that protocols are not really trials but only the idealized version of the trial, I wonder whether one should even be examining protocols when developing PRECIS. Clearly a different (potentially) version of the trial than the actual trial as published. Also, I had a very difficult time with trials that are far afield of my content area. Had to leave several cells blank because I simply don’t have any clue what “usual care” might look like...” this issue was discussed with another who said that changing protocol unethical and unlikely to happen. Issue of expertise also concern raised by others (Daniel Bratton – TB expert/methodologist).

Clear concrete guidance on using PRECIS-2, should reduce subjectivity though in many cases a judgement is still needed to score a trial. If a trial team use PRECIS-2 to rate the trial, including the chief investigator, clinicians that are experts in the trial area, statistician then information sharing will ensure everyone has sufficient knowledge of the trial to score the different PRECIS-2 domains. As
mentioned previously, the input of an external observer is also likely to reduce bias in PRECIS-2 scoring as well.

Ideally to validate PRECIS-2 we would have used the tool with at least 15 trial teams to design a trial in the way the tool was intended to be used, using at least 15 raters in each trial team to score the trial. Use of PRECIS-2 retrospectively by raters unfamiliar with the trials is not the intended use of PRECIS-2 and does act as a limitation. Unfortunately prospective use of PRECIS-2 with trial teams was not practical, necessitating retrospective work instead. However, although there is some unreliability in PRECIS-2 due to reporting issues when used retrospectively, this is not in itself a problem. The primary purpose of PRECIS-2 is to encourage discussion and consensus and to assist trialists being clearer on their trial design by considering the end user of their trial results. This viewpoint on validating PRECIS-2 and its utility was also discussed by Glasgow (2012) on using PRECIS with RE-AIM to consider translation of study evidence into real world settings [73]. Our retrospective validity and reliability testing is a hard test for PRECIS-2 and, despite this, it has performed reasonably well. It suggests that PRECIS-2 could be used retrospectively in, for example, systematic reviews to rate included trials. Recently, there has also been work, post PRECIS-2 development, using the tool to consider how pragmatic trial grant applications are, at the National Institutes of Health in the USA ([170] and the PRECIS-2 website (see Chapter 8B) is being increasingly accessed.

**Conclusion**

We have succeeded in testing the validity and inter-rater reliability of PRECIS-2. Due to the lack of reporting of information in particular domains (e.g. recruitment) there were problems for raters in scoring some domains. These results indicate that the PRECIS-2 model performs reasonably well to identify pragmatism even when used retrospectively - which is not its intended use. Although its discriminatory power seems promising, further work with individual raters assessing overall pragmatism as well as using PRECIS-2 for a group of at least 15 trials would be desirable.
SUMMARY
Validity and reliability was tested using 19 raters (international trialists from seven countries) who used PRECIS-2 to score a varied sample of 15 RCT protocols. Trials were purposefully selected so that they varied between pragmatic and explanatory, with both drug and non-drug trials included. Inter-rater reliability was generally good, with seven of nine domains having an ICC over 0.65. Discriminant validity was reasonable with better than chance discrimination for all domains.

CONCLUSION
We had succeeded in validating PRECIS-2 unlike previous versions of PRECIS. Work on guidance and the website should improve the tool further.
Chapter 8A: Developing a search strategy for pragmatic trials

Background

Following work to improve PRECIS through the modified Delphi and Brainstorming meetings, we then validated PRECIS-2 using 15 trial protocols and 19 raters. At this stage in the PhD project we had a tool to judge how explanatory or pragmatic clinical trials are. However, we needed to develop the methodology for searching for pragmatic trials so that we could create a database of pragmatic trials that could be stored in the PRECIS-2 website. Once we had a set of pragmatic trials these could be matched with explanatory trials to consider internal and external validity and estimates of treatment effect of the different trial designs.

Aims and Objectives

Aims

The aim of this chapter was to develop a search strategy for pragmatic trials which could then be stored in the database of pragmatic trials on the PRECIS-2 website.

Objective

Develop a search strategy for pragmatic trials.

Methods

A search strategy was developed to select pragmatic trials (Box 8.1). This went through two versions following a pilot using the first version. The initial one was developed through discussion with AJ, lead information scientist for the Library and Learning Centre at the University of Dundee College of Medicine, Nursing and Dentistry. Articles of trials that were widely accepted as being pragmatic were
used as the basis to develop the search strategy using MEDLINE in Ovid and to test if the articles had been selected using our search strategy.

| 1. | “pragmatic trials” |
| 2. | “management trials” |
| 3. | “efficiency trials” |
| 4. | “practical trials” |
| 5. | “effectiveness trials” |
| 6. | “efficacy trials” |
| 7. | “field trials” |
| 8. | ("real-world" OR “real life”) AND trials OR trial) |
| 9. | MH “Clinical Trials as Topic+/ST/SN/MT”) |
| 10. | S1 or S2 or S3 or S4 or S5 or S6 or S7 or S8 |
| 11. | S9 NOT S10 |
| 12. | S9 NOT S11 |
| 13. | S10 OR S12 |

*ST – standards; SN – statistics & nominal data; MT - methods*

**Box 8.1 Version 1 Search strategy for pragmatic trials**

**Pilot testing of search strategy**

One hundred studies from the earliest trials randomly selected from the pragmatic trial search (Version 1) were chosen for the pilot study to detect any problems with inclusion criteria. ST and KL had 77% overlap with selection of pragmatic trials. Following this pilot exercise KL decided to select only English language articles for full text (not just abstract). KL also realised that in searching for pragmatic trials, discussion articles and pilot studies were being selected which were excluded. In addition, behavioural trials were picked up in the search strategy as “real life” due to the difficulties of testing driving for instance under the influence of drugs or alcohol or lack of sleep. It was decided to exclude these virtual simulation of test situations as usually laboratory conditions. The second version of the search strategy (Box 8.2) was developed with assistance from CF at Health Services Research unit (HSRU) at the University of Aberdeen.
Database: Ovid MEDLINE(R) In-Process & Other Non-Indexed Citations and Ovid MEDLINE(R) <1946 to Present> Search Strategy:

1  randomized controlled trial.pt. (388158)
2  controlled clinical trial.pt. (89760)
3  1 or 2 (472997)
4  effectiveness trial?.tw. (849)
5  (pragmatic or real world or real life).tw. (27492)
6  3 and (4 or 5) (2581)

Included only: intervention studies in English only.
Excluded: pilot studies, study protocols, longitudinal studies, controlled clinical trials, before and after studies, cost effectiveness trials, case reports, opinion pieces on trials (pt – publication type; tw – look in titles and abstracts)

Box 8.2 Version 2 Search strategy for pragmatic trials

Screening and identification of pragmatic trials

KL ran the search strategy detailed in Box 8.2 on 11th November 2013 and obtained 2581 titles and abstracts for screening. The title and abstracts were downloaded into EndNote X4 in batches of 100. The aim of the search strategy was to identify pragmatic trials. To ensure the sample was pure, all titles and abstracts downloaded into EndNote were screened to ensure only intervention pragmatic trials were included. Thus pilot studies, study protocols, longitudinal studies, controlled clinical trials, before and after studies, cost effectiveness trials, very explanatory trials, case reports, and opinion pieces on trials were excluded. Duplicates were then removed. Trials classified as explanatory were clearly investigating if an intervention worked in ideal conditions and included simulations or laboratory tests to simulate real world situations e.g. driving and drinking or use of drugs.

Results

Our search for pragmatic trials identified 2581 studies, which was reduced to 763 trials after removing studies that were not trials and duplicates. This database contained 234 (30%) references from American Journals e.g. American Journal of Nursing, American Journal of Respiratory & Critical Care Medicine, American Journal of Cardiology, American Journal of Clinical Dermatology and American Journal of Critical Care. One of the largest groups of pragmatic trials was cardiovascular trials – 88 RCTs.
Discussion

Searching for pragmatic trials is hard; the search strategy version 1 or 2 that we developed only found articles that had “pragmatic” in their title or abstract as a keyword. This meant, for instance, we were unable to develop a search strategy that picked up a trial that we knew was pragmatic as determined by MZ and trial team and peer review. For instance, the DOT trial published in the Lancet in 1998 [171] - Randomised Controlled trial of self-supervised and directly observed treatment of tuberculosis (DOT). This trial article had NO MeSH terms for pragmatic, the Publication labels were Type “Randomised controlled trial” and “Clinical trials”.

There were no pragmatic trials picked up before 1979 but there has been an exponential increase in articles describing pragmatic trials, though the National Library of Medicine (NLM) is still not using the term “pragmatic trial” as a MeSH term. KL, and no doubt others, has sent requests that they consider using this MeSH term to assist researchers and practitioners using trial results. Determining what is a pragmatic trial, however, if the authors do not use Pragmatic as a keyword can be difficult. This is probably one of the reasons the NLM is not prioritising “pragmatic” as a new MeSH term.

Considering the Web of Science graph of publications (Figure 32) with term “pragmatic trial” in title or abstract, had 8950 publications on the 6th November 2014. As can be seen by the graph below, there has been an exponential rise in the number of pragmatic trials that have been published.
The search strategy, although developed with specialist input was overly sensitive and not sufficiently specific to select only pragmatic trials; with just over 1 in 3 articles being screened as pragmatic. For instance, in screening the results of the search strategy, KL decided to exclude obviously explanatory trials, in particular if they stated that they were “efficacy and safety trials”. These trials often stated “further studies are needed in a larger population”. The search strategy did select some very explanatory trials, for instance one trial “Steady-state bioequivalence study of clozapine tablet in schizophrenic patients” [172]. The abstract stated “...The present study was conducted under real-life conditions in schizophrenic patients...” but on reading the trial details, this trial is quite clearly using surrogate outcomes and looking at efficacy rather than the effectiveness of the treatment for schizophrenic patients. Having identified 763 pragmatic trials, the next stage of the project (Chapter 7B) was the creation of the PRECIS-2 database where they would be stored.

**Methodological limitations**

Due to the incremental nature of the work there are no details on the different types of articles excluded to identify pragmatic trials.
Conclusion

A search strategy was designed to select pragmatic trials. As there was no MeSH term for pragmatic trials this was not particularly sensitive. Exclusion of studies that were not pragmatic trials left 763 RCTs to be included in the PRECIS-2 database. This is the biggest collection of pragmatic trials currently available.

Summary
Designing a search for pragmatic clinical trials is not easy. In the end, 763 pragmatic clinical trials were selected for inclusion in the PRECIS-2 database.
Chapter 8B: Creation and development of the PRECIS-2 website

Background

To support future users of the PRECIS-2 tool to design clinical trials fit for purpose a website needed to be created which would include guidance and a summary of useful PRECIS literature. A key feature of the website was an electronic tool that enabled raters to create PRECIS-2 wheels based on the data for all nine domains, for their individual and aggregate scores from several trialists in a team using PRECIS-2. The PRECIS 2 website also would contain a database of trials based on the pragmatic trial search (Chapter 8A) which would be easy to search. Through creation of the PRECIS-2 database, work on the next stage of matching explanatory and pragmatic trials by intervention, to then assess the internal and external validity of effect estimates this set of trials could take place.

Aim and Objectives

Aim
To create a PRECIS-2 website for potential users of the tool.

Objectives
The aim of creating a website for PRECIS-2 was achieved through designing a website with several functions: a database of trials with varying pragmatic design, with information for trials groups which would include a tool kit and software to score trials being designed using PRECIS-2. This website had to be accessible for all potential users and in particular KL to construct the database of pragmatic trials.

Methods

Work on the website started in July 2013. The original specifications are detailed in Box 8.3. Starting from scratch to develop the website involved several face to face meetings and regular e-mail
communication with the software developer (LT), in particular to test changes and edit accordingly. Ideas from different websites were suggested for design and front page presentation, in particular the Normalisation Process Theory website that ST had been involved in which appeared to be simple, uncluttered and easy for users. [http://www.normalizationprocess.org/](http://www.normalizationprocess.org/) It was also important that the PRECIS-2 website was freely accessible. We did, however, want to monitor who was using the web tool to create PRECIS-2 wheels for the trials created by a trial team, so we asked trialists to register for this facility using name and password (top right corner of the home page). This password was known only to the user but the database in its entirety can be seen by KL and LT. (Some of the trials are not released to the database so can only be seen by the PRECIS-2 user that has registered to use the site.) When a user registers, an email is automatically generated and sent to KL, the user can then be activated in the Admin page by KL or TL.
Database Specifications

- Database will be available via a website, which will be hosted by the University of Dundee (and, in due course, by the University of Aberdeen).
- The website must work without an internet connection, so offline as well as online version.
- We want to display some mainly text-based information about roughly 200 trials. We will extract the data.
- The fields that we will extract and which we want to be able to display in the database include: topic of trial, intervention type, number of centres, sample size, randomization, allocation and blinding methods, type of statistical analyses, participant characteristics (e.g. age, condition, baseline severity etc.), duration of trial, effect size and measures of variance for the primary outcome, bias. Each of these essentially boils down to a single numeric or text field. We might add an image (PRECIS-2 wheel) to each entry as a pdf.
- We would like to have search so we can, for example, look for all trials of aspirin in the database.
- We would like to be able to have access options so that we can add more trials but others can only view it, or we can give others access rights.

PRECIS-2 software

- There is no easy way of producing a PRECIS-2 picture at present so a tool that produced a picture that could then be exported/cut & pasted as Word and pdf would be great.
- PRECIS-2 has 10 domains, each with a score from 1 to 5 (probably worth making it from 0 to 5 at this point).
- It would be useful to have a bit of software with a table that gave the name of each of the 10 domains (see chapter 6, table x), a place to score and a field for rationale.
- Once the scores for the 10 domains are filled in, the PRECIS-2 wheel should be created from this. Ideally there would be two options: 1) just mark the score on each domain with a dot (red, say, although maybe user could change this) and 2) mark with dots and then join the dots.

Box 8.3 Original specifications for the Database and PRECIS-2 website (prior to final PRECIS-2 tool)

The website created was accessible offline as long as the web page was cached on the system. The initial prototype prior to input validation, normalization and cleaning was the basis for further development. Offline “refresh” buttons still worked in the PRECIS-2 website and it was possible to generate as many wheels as a trialist needed. It was fully tested by KL and GF with trial teams using the PRECIS-2 website (Chapter 9). One of the useful aspects of the PRECIS-2 website is the ability to generate thumbnail images that can be saved as jpeg pictures.

To assist in further work for internal validity assessment of a sample of the database trials we also added information on Risk of Bias fields (Box 8.4).
To assist in presenting Effect Measures of trials results in a similar fashion and thus ensure enabling comparison, LT developed an algorithm similar to the Cochrane software RevMan [173] to automatically enable users to calculate the Effect measure of Risk Difference e.g. 83/100 - 23/100 = 60%. However, if some trials have adjusted data this can also be entered manually and put into the comments field too. It is thus possible for trialists entering the data to explain why the adjusted figure is different from just doing the basic calculation. If it is not possible to calculate Risk Difference then the alternative presentation of Effect measure can be inserted into the Comments box as well, or any additional information (box 8.5). If PRECIS-2 is being used to assist in trial design and the trialists are working with developing the protocol the Effect Measure or Estimate of Treatment effect field is optional.

<table>
<thead>
<tr>
<th>Effect Measures</th>
<th>Events Intervention</th>
<th>Total</th>
<th>Events Control</th>
<th>Total</th>
<th>Risk Diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nicardipine: Therapeutic response approx. 15min</td>
<td>83</td>
<td>100</td>
<td>23</td>
<td>100</td>
<td>60.00 %</td>
</tr>
<tr>
<td>Nitropruss: Therapeutic response approx. 60mins (more variation)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Box 8.5 Risk Difference data

To assist searching the database with trials of varying pragmatic trial designs, LT created a drop down list. Trialists can preselect the type of trial they were entering. The trial categories are detailed in Box 8.6.
To assist presentation of trial scores, maximum and minimum scores were also overlaid in red on the PRECIS-2 wheels that were created from the median scores. It was thought this would help in initial trial discussion of design if trialists had scored the trial independently and could clearly see in the PRECIS-2 wheel diagram the differences in scores by raters. This wheel with maximum and minimum red lines (if different) would assist trialists in focusing discussion on areas of disagreement.

**Pilot feedback of Website**

IG, a German GP trialist, who assisted in the pilot for the validity study (Chapter 7A) was also asked to give feedback on the website 5th March 2014 (Box 8.7). Her suggestions were discussed with LT and a new PRECIS-2 logo was put on the home page and changes made wherever possible. The main trial mentioned by IG that had just finished recruitment following publication of the pilot, a pragmatic design treating urinary tract infection, was also put into the PRECIS-2 database[174] [175].

All new users of the website who register to fully access the PRECIS-2 tool and database are advised that the website is under development and feedback is encouraged. We also have on the PRECIS-2 home page “*We would be very grateful if users would give us feedback on using PRECIS-2: please visit the Contact us page. These PRECIS-2 criteria are constantly being reviewed and we welcome your input.*”
Congratulations, it looks really good. I am very impressed!

Here my remarks:
- The Home-Page is a kind of introduction/advertisement but for this purpose it is not well designed. Bigger letters, a small wheel as a symbol in one of the corners a bit more structure might be useful.
- Trials: when you show them it would be useful to give the references because it is not to understand what are the scores were given for. I would be appreciated if ICUTI was there as well. We have finished the recruitment last week.
- The PRECIS 2 Info. Here is the text bounce around the wheel not uniform. Thea are partly in two lines, partly in 5-6. But it is just an aesthetical point
- How to? The distances between the single paragraphs are not uniform. I’m not sure if the difference of sizes between the letters are not too big (I mean the text and the subheadings). Altogether it gives a “restless picture”.
- The “Kits” is very useful. You have here two different blue colors which don’t suit very well. The Letters in your graphs appear a bit fuzzy

Each other side is well done! Thank you for asking me for a feedback.

Box 8.7 Feedback on PRECIS-2 Website after pilot testing

The tool kit was being developed simultaneously through user testing, validity testing and the PRECIS-2 elaboration paper [153] and was edited several times. From December 2014, we have used Google analytics to get PRECIS-2 site metrics to capture the number of people viewing the site and unique visits.

RESULTS

PRECIS-2 website

We purchased a domain and created a publicly available website www.precis-2.org that is currently located on the Health Informatics Centre (HIC), University of Dundee server: https://crs.dundee.ac.uk/precis/Trials/. This will soon be moved to a server at University of Aberdeen.

KL, working with the software developer (LT) designed the website which has the PRECIS-2 training kit, scoring system for trial teams plus additional guidance, a summary of other PRECIS literature and useful references to support the design of trials via the website.
Pragmatic trial database

The plan was to use the PRECIS-2 tool to create a database of 200 trials that take a pragmatic approach to design. The challenging nature of matching meant that it was not possible to reach 200. We have, however, created a database of 58 trials of varying pragmatism which is searchable and extendable (that is new studies can be added). Many of the trials included were selected with matching to explanatory trials of the same intervention in mind, with a focus on cardiovascular drug trials, although other trials were also included to test the database design. Currently, in the database, there are an additional 11 trials protocols with varying PRECIS-2 scores from individual trialists scoring the design of their trials, so in total there are 89 trials in the database as of 10.11.14. These trials are only visible to the trialists who enter the trial data and the administrator (KL) until they are released by the individual trialists into the pragmatic trial database. We were, however, able to make the database itself more functional than originally envisaged. There are 30 registered users (December 2014) of the PRECIS-2 website (not including KL and LT) see Table 46.

<table>
<thead>
<tr>
<th>Known users of PRECIS-2 website</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pragmatic Clinical Trials Unit (PCTU) at Queen Mary’s University of London</td>
<td>UK</td>
</tr>
<tr>
<td>University of Exeter</td>
<td>UK</td>
</tr>
<tr>
<td>Clinical Trials Evaluation Unit at Bristol University</td>
<td>UK</td>
</tr>
<tr>
<td>National Institute of Health (NIH)</td>
<td>USA</td>
</tr>
<tr>
<td>Paraxel</td>
<td>USA</td>
</tr>
<tr>
<td>The Center for Medical Technology Policy (CMTP)</td>
<td>USA</td>
</tr>
<tr>
<td>University of Colorado, Denver</td>
<td>USA</td>
</tr>
<tr>
<td>Family and community Medicine, University of Toronto</td>
<td>Canada</td>
</tr>
<tr>
<td>Centre for Online Health and Queensland Children’s Medical Research Institute (QCMRI) at The University of Queensland</td>
<td>Australia</td>
</tr>
</tbody>
</table>

Current testers of the PRECIS-2 website (Table 46) include trialists in the UK, USA, Canada and Australia. Some of these people have heard about the website through work underway to test PRECIS-2 (APT see unpublished protocol) at the Pragmatic Clinical Trials Unit. Others had requested to use the website
(National Institute of Health) and researchers from the University of Colorado used PRECIS-2 to consider how pragmatic grant applications were for trial protocols. Additional testers heard about the website through personal communication and others have enquired following the publication of the protocol which specified that we would create a PRECIS-2 website [52]. In addition, individuals from the Centre for Online Health and Queensland Children’s Medical Research Institute (QCMRI) at The University of Queensland and the Clinical Trials Evaluation Unit at Bristol University, http://cteu.bris.ac.uk/ enquired about the website but did not register – this means that they are not able to create their own trial wheel but the former used the PRECIS-2 tool as part of a systematic review considering how explanatory and pragmatic included trials were. Feedback has been positive epitomized by this comment “I have used the website and it’s very intuitive and easy to navigate.”

The following figures demonstrate the home page layout (Figure 33), the information layout for “PRECIS-2 info” (Figure 34), HepFree trial (trial on screening and treating immigrants at high risk for Hepatitis) scores from five individual scorers (Figure 35) and the overall presentation of the HepFree trial including thumbnail wheel – clicking on the “+” gives more information (Figure 36). Trials can also be aggregated by the administrator using the “aggregate” function and clicking in boxes marking trials (Figure 37). These can then be joined to tidy up the database and fuse separate entries for the same trial by trialists individually scoring the trial. This is different to one administrator taking scores from individual raters and putting them into one trial entry. Finally the maximum and minimum scores for a trial have been visually created for each domain, in addition to the wheel creation using median scores to assist trial teams in discussing the design of individual domains. If median scores are the same as max and the red line will be superimposed on the median line but if there is divergence in opinions this will be clearly visible (Figure 38).
Figure 33 Home page of PRECIS-2 website

PRECIS-2

Designing clinical trials is challenging. PRECIS – PRagmatic Explanatory Continuum Indicator Summaries – is a clever acronym for a tool to help trialists designing clinical trials consider where they would like their trial to be on the pragmatic/explanatory continuum.

This PRECIS-2 website has two functions:
1. a training resource;
2. a database of trials that have been scored using PRECIS-2

Trialists working on their own trial can apply for a password so that their team can score their trial while developing the trial design and protocol.

This trial design information will only be visible to trialists using a password until they decide to make this information publically available. We advise one password per trial team so you all have access to score the same trial. A PRECIS-2 wheel will be generated, based on all the scores, and can be used for discussion and consensus.

The database of trials contains trials that are a spectrum of pragmatic trials. We hope this will be helpful to researchers who can then search for trials on particular topics and consider the trial design. In addition trialists can look at the internal validity using the Risk of Bias tool.
PRECIS-2

In 2009 a tool called the Pragmatic-Explanatory Continuum Index Summary – PRECIS – was published to help trialists to think more carefully about the impact their design decisions would have on applicability (Thorpe 2009). This tool has been improved and validated to create PRECIS-2 (Loudon 2013). This is a 9-spoked ‘wheel’ with nine domains based on trial design decisions (i.e. Eligibility criteria - who is selected to participate in the trial? Recruitment - How are participants recruited into the trial? Setting - Where is the trial being done? Organisation – what expertise and resources are needed to deliver the intervention? Flexibility delivery – How should the intervention be delivered? Flexibility adherence – what measures are in place to make sure participants adhere to the intervention? Follow-up – How closely are participants followed-up? Primary outcome – how relevant it is to participants? and Primary analysis – to what extent are all data included?).

The PRECIS-2 ‘wheel’ can visually represent how explanatory/pragmatic a trial is on the pragmatic to explanatory continuum. Trials that take an explanatory approach produce wheels nearer the hub; those with a pragmatic approach are closer to the rim.

Figure 34 PRECIS-2 Info

Scores for Trial: HepFree

Figure 35 individual PRECIS-2 scores for one trial
Figure 36 scoring – end result PRECIS-2 wheel

Figure 37 scoring – aggregate function
Finally Google analytics was used with this link for the PRECIS-2 website [153], the number of users is increasing (Figure 40). The countries using the website after six months development and testing (Figure 39), includes: Russia (31 sessions), UK (26 sessions), United States (16), France (6), Canada (2) and Germany (2).

<table>
<thead>
<tr>
<th>Country</th>
<th>Sessions</th>
<th>% Sessions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Russia</td>
<td>31</td>
<td>37.35%</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>26</td>
<td>31.33%</td>
</tr>
<tr>
<td>United States</td>
<td>16</td>
<td>19.28%</td>
</tr>
<tr>
<td>France</td>
<td>6</td>
<td>7.23%</td>
</tr>
<tr>
<td>Canada</td>
<td>2</td>
<td>2.41%</td>
</tr>
<tr>
<td>Germany</td>
<td>2</td>
<td>2.41%</td>
</tr>
</tbody>
</table>

Figure 39 PRECIS-2 website users by country (8th January 2015)
Figure 40 Google analytic user information (18th December 2014)
Discussion

Database pilot testing issues

Setting up fields for data required testing to ensure numeric integer or text field. For example, input of trial countries trials, KL tried to put in USA and got error message: The value ‘USA’ is not valid for Countries involved. Font size and colours as well as layout were all discussed with the software designer to make it easier for website users. There were also problems with automatic saving so work was lost which was extremely frustrating, but this was sorted by the software developer so there is now Autosave and a warning message to users if they are about to be timed out.

We had to prevent problems for users navigating round the PRECIS-2 website to ensure that there was a logical order in moving around the website. This was resolved through testing, in particular with trialists based at the PCTU and through e-mail discussion to prevent users becoming frustrated. There also appeared to be a difference in Google Chrome and Internet Explorer in difficulties entering data and the way data was presented which was resolved in further versions of the website.

Developing a simple website with the software engineer was one of the most fulfilling parts of the PhD. At the time of writing (December 2014) the PRECIS-2 website has not gone live but is being increasingly accessed. The website has been pilot tested and used by trial teams at the PCTU. Following publication of the PRECIS-2 elaboration paper, this paper and examples of trials assessed by PRECIS-2, will be placed on the website and it is anticipated website traffic will increase. This can be easily monitored using Google analytics.

Conclusions

The PRECIS-2 website developed has a toolkit to assist trialists designing trials and software to assist trial teams to create their own PRECIS-2 wheels which can then be added to the database of trials on
the website. An increasing number of enthusiastic international testers are using the PRECIS-2 tool, freely accessible on the internet. The development of the PRECIS-2 website enabled the third and final part of the work to compare the internal and external validity and effect sizes of a matched set of explanatory and pragmatic trials.

<table>
<thead>
<tr>
<th>SUMMARY</th>
</tr>
</thead>
<tbody>
<tr>
<td>A PRECIS-2 freely accessible website <a href="http://www.precis-2.org">www.precis-2.org</a> has been created that has a trial database that is searchable and extendable. The website includes information on how to use PRECIS-2 and a tool to assist the trial team to create their own wheels to assist consensus decision making. This has received positive feedback and is continuing to be tested by international trialists.</td>
</tr>
</tbody>
</table>
Chapter 8C: Internal and External validity of pragmatic and explanatory trials

Background

Following work to develop a search strategy for pragmatic trials (Chapter 8A) and the creation of a database for the trials in the PRECIS-2 website (Chapter 8b), the final part of the work to assess the external and internal validity and effect sizes of a matched set of pragmatic and explanatory trials could be undertaken. PRECIS-2 (Figure 41) could then be used to answer speculation [43, 44] that pragmatic trials sacrifice internal validity in order to achieve applicability and help provide an empirical basis for the promotion (or rejection) of pragmatic design choices in clinical trials. As evidence on the relative effect sizes of pragmatic compared to explanatory trials of the same intervention was lacking, I also proposed to directly compare the effect sizes of these different trial designs using trials investigating the same intervention. The hypothesis was that trials that take an explanatory approach tend towards larger effect sizes, which are then not seen when the treatment is used in routine care. Such knowledge is important because some treatment may cease to be cost-effective if the actual benefit is rather less than that measured in the trial.
Aims and Objectives

Aims

To use the validated tool (PRECIS-2) to compare the external validity and the Cochrane Risk of Bias tool to assess the internal validity; and to compare the effect size in pairs of explanatory and pragmatic clinical trials, matched by intervention.

Objectives

1. Match a set of pragmatic trials with more explanatory trials of the same intervention, condition and participants.

2. Use the Cochrane Risk of Bias tool to assess internal validity of a matched set of explanatory and pragmatic trials
3. Use the PRECIS-2 tool to score domains to consider external validity of a set of explanatory and pragmatic trials.

4. Compare effect estimates of a set of matched pragmatic and explanatory trials in a meta-regression

**Methods**

**Objective 1: Identifying matching explanatory and pragmatic trials**

**Method 1**

We aimed to match cardiovascular drug trials taking a pragmatic approach with trials of the same intervention and primary outcome taking a more explanatory approach. Matching trials was challenging. The intention was to match trials with similar patients on the basis of same intervention (e.g. anti-hypertensives to reduce blood pressure), condition (e.g. hypertension – high blood pressure) and same primary outcome (e.g. mortality). It was, however, hard to get matches.

Pragmatic trials had been found (see Chapter 8A) so KL was looking for explanatory trials. For example, considering the pragmatic Randomised Controlled Perioperative Beta Blockade (POBBLE) trial [176], most studies considered for matching appeared to be controlled (but not randomised) trials, for example, Pasternack with 48 patients selected to receive 50mg metoprolol compared to 152 similar but untreated peripheral vascular surgery patients [177]. Another potential match “Perioperative beta-blocker therapy and mortality after major noncardiac surgery” was a cohort study involving 782,969 patients carried out in 329 hospitals in USA [178]. RCTs published with the intervention of metoprolol, that were perhaps precursors to the POBBLE study were hard to find, though there were RCTs involving different intervention anti-hypertensives: atenolol and bisoprolol. Both beta blockers are similar and appear to be used in similar way but it was hard to get identical comparisons for the purpose of matching this RCT. Another RCT enrolled patients earlier than Brady’s POBBLE trial but was published
at the same time as the Metoprolol after Vascular Surgery (MaVS) trial, this was a larger study which was not specifically described as explanatory, POBBLE stated MaVS was a trial of a “pragmatic policy of perioperative beta-blockade...” but on further consideration was entered into the database [as explanatory] for comparison as the intervention, metoprolol, was used in both trials for patients as a perioperative beta-blockade to reduce hypertension during vascular surgery [179].

Techniques to identify more explanatory trials involved searching using PubMed, Google, looking at the included studies in systematic reviews in The Cochrane library, and other published systematic reviews and meta-analysis. In addition, KL used reference lists of possible articles (in particular taken from citations in the introduction and background to a study) and looked at clinical trial websites e.g. [https://clinicaltrials.gov/](https://clinicaltrials.gov/) and [https://www.clinicaltrialsregister.eu](https://www.clinicaltrialsregister.eu).

**Method 2**

As it proved to be so hard getting matches, additional work was carried out using a Cochrane systematic review of first line drugs for hypertension [180]. A Cochrane systematic review was selected as the intervention for all the included trials would be the same (anti-hypertensives – drugs to treat high blood pressure), treating patients with the same condition (hypertension) and there would be a common outcome (e.g. mortality). So The Cochrane Library and the Cochrane Database of Systematic Reviews were searched for the Cochrane Hypertension Group and KL selected the first pharmacotherapy review done on hypertension "First line drugs for hypertension" as a potential source of a mixture of pragmatic and explanatory trials. This Cochrane review had 24 trials which would ensure assessment was feasible and support the work detailed in Objectives 2, 3 and 4. This Cochrane systematic review was first published 2002, updated 2009 and is currently being updated.
Objective 2, 3 and 4: Measuring risk of bias to assess internal validity, PRECIS-2 to assess external validity in a set of explanatory and pragmatic trials and comparing effect estimates

Two different methods had been used to create two groups of trials to assess the internal and external validity, as well as the effect sizes of explanatory and pragmatic trials. To measure internal validity of the pragmatic and explanatory trials in the matched pairs and in the Cochrane systematic review we used the Cochrane Risk of Bias tool (Table 8.2) [46] using a detailed version with criteria for assessing each domain. This considers the trial structure and design and the mechanisms of the trial conduct and covers six areas of bias: selection bias, performance bias, detection bias, attrition bias, reporting bias, and other bias. Using it requires expert judgement. As a result, there may be some variation in results by users who are not very familiar with the trial. Despite this, it is a widely used and accepted tool for measuring whether or not a trial has internal validity.

We used PRECIS-2 (Figure 8.12) and the PRECIS-2 website www.precis-2.org to produce PRECIS-2 wheels to consider the external validity and applicability of the matched pragmatic and explanatory trials and the Cochrane systematic review “First line treatment for hypertension”.

The Cochrane Risk of Bias scores for each of the matched trials and the trials in the Cochrane systematic review were then to be compared, as were estimates of treatment effects, the latter in a meta-regression (see Analysis below).

Data Collection Process

In addition to extracting data to assess internal validity using the Cochrane Risk of Bias tool [46] and external validity using PRECIS-2, KL extracted the following information from the retrieved trials for the matched cardiovascular explanatory/pragmatic trials and for the Cochrane systematic review “First line drugs for hypertension”: trial intervention, number of centres, sample size, participant
characteristics (e.g. age, baseline severity), duration of trial, and effect size and measures of variance for the primary outcome.

**Quality assurance**

Duplicate data extraction was undertaken by ST in 6 out of 28 (21%) randomly selected pairs of the matched pragmatic and explanatory trials. In addition, duplicate data extraction was also done for the Cochrane systematic review “First line drugs for hypertension” by ST for a random sample of 4 out of 23 (17%) of the RCTs. For both methods (matched trials and the Cochrane systematic review) ST used the Cochrane Risk of Bias to assess internal validity and the PRECIS-2 scores to assess external validity.

**Analysis**

Estimates of treatment effects for matched trials were converted to Risk Difference wherever possible using the software on the PRECIS-2 website which was based on the Cochrane Collaboration RevMan algorithm, otherwise continuous variables were presented e.g. mean change, percentage change or for BP in one trial change in mm Hg. The Cochrane Risk of Bias scores were compared between matched explanatory and pragmatic trials to assess internal validity with regard to: sequence generation, allocation sequence concealment; blinding of participants, personnel and outcome assessors; incomplete outcome data; selective outcome reporting; other potential threats to validity like publication bias. Effect sizes were to be compared in a random effects meta-analysis across all trials with type of trial (explanatory and pragmatic trials; binary coded) as an independent variable and overall effect as the dependent variable.

For the Cochrane review “First line drugs for hypertension” the Cochrane Risk of Bias tool was used to assess Internal Validity (Table 47) [46] of all the included trials – 23 in total. It has been hoped to use the scores in the Cochrane review but the authors had only published scores for “Allocation concealment” with an overall Risk of Bias score. PRECIS-2 scores were used as an indicator of external validity and estimates of treatment effect were noted.
Table 47 Cochrane Collaboration tool for assessing risk of bias

<table>
<thead>
<tr>
<th>Bias domain</th>
<th>Source of bias</th>
<th>Support for judgment</th>
<th>Review authors’ judgment (assess as low, unclear or high risk of bias)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection bias</td>
<td>Random sequence generation</td>
<td>Describe the method used to generate the allocation sequence in sufficient detail to allow an assessment of whether it should produce comparable groups</td>
<td>Selection bias (biased allocation to interventions) due to inadequate generation of a randomised sequence</td>
</tr>
<tr>
<td></td>
<td>Allocation concealment</td>
<td>Describe the method used to conceal the allocation sequence in sufficient detail to determine whether intervention allocations could have been foreseen before or during enrolment</td>
<td>Selection bias (biased allocation to interventions) due to inadequate concealment of allocations before assignment</td>
</tr>
<tr>
<td>Performance bias</td>
<td>Blinding of participants and personnel*</td>
<td>Describe all measures used, if any, to blind trial participants and researchers from knowledge of which intervention a participant received. Provide any information relating to whether the intended blinding was effective</td>
<td>Performance bias due to knowledge of the allocated interventions by participants and personnel during the study</td>
</tr>
<tr>
<td>Detection bias</td>
<td>Blinding of outcome assessment*</td>
<td>Describe all measures used, if any, to blind outcome assessment from knowledge of which intervention a participant received. Provide any information relating to whether the intended blinding was effective</td>
<td>Detection bias due to knowledge of the allocated interventions by outcome assessment</td>
</tr>
<tr>
<td>Attrition bias</td>
<td>Incomplete outcome data*</td>
<td>Describe the completeness of outcome data for each main outcome, including attrition and exclusions from the analysis. State whether attrition and exclusions were reported, the numbers in each intervention group (compared with total randomised participants), reasons for attrition or exclusions where reported, and any reclusions in analyses for the review</td>
<td>Attrition bias due to amount, nature, or handling of incomplete outcome data</td>
</tr>
<tr>
<td>Reporting bias</td>
<td>Selective reporting</td>
<td>State how selective outcome reporting was examined and what was found</td>
<td>Reporting bias due to selective outcome reporting</td>
</tr>
<tr>
<td>Other bias</td>
<td>Anything else, ideally prespecified</td>
<td>State any important concerns about bias not covered in the other domains in the tool</td>
<td>Bias due to problems not covered elsewhere</td>
</tr>
</tbody>
</table>

*Assessments should be made for each main outcome or class of outcomes.
Results

Method 1- Cardiovascular trials

Objective 1: Matched explanatory and pragmatic trials
From the pragmatic trial database, KL selected 28 cardiovascular trials that she judged would be most likely to have explanatory trial matches based on the same intervention, patients and primary outcome. In fact, there were only 14 matches and even here comparators were not always identical and estimates of treatment effect could not always be directly compared (Appendix Table 8.1).

Objective 2, 3 and 4: Measuring risk of bias to assess internal validity, PRECIS-2 to assess external validity in a set of explanatory and pragmatic trials and comparing effect estimates
Extracted data on variables (intervention, primary outcome, number of centres, sample size – number randomised, degree of pragmatism, participant baseline severity, duration of trial, estimates of treatment effect – as Risk difference if possible and measure of variance for primary outcome) for the matched trials is shown in Appendix Chapter 8, Table 8.1. Unfortunately the quality of data for the 14 matched cardiovascular trials was inadequate to undertake a meta-regression analysis. There did not appear to be any clear conclusions with regard to a link between pragmatism, extracted variables and effect size looking at the data in Table 49.

Two of the cardiovascular trials [176, 181] from the pragmatic trial search were scored using PRECIS-2 as 3 equally pragmatic/explanatory, and one trial [182] as “2” i.e. rather explanatory but as their matches were even more explanatory they were included.

With regard to internal validity, there was inadequate information in 62 trial bias judgement domains out of a potential of 168 judgments on random sequence allocation, allocation concealment, blinding, incomplete outcome data, selective outcome reporting and other sources of bias giving internal validity assessments of “Unclear” internal validity in half of the matched trials. Risk of bias assessment did not, however, reveal obvious differences with regard to internal validity between the two design
approaches with 10 out of the 14 matched pairs of explanatory and pragmatic trials having the same internal validity assessment (Appendix Table 8.2).

**Method 2 - Cochrane hypertension review**

**Objective 2, 3 and 4: Measuring risk of bias to assess internal validity, PRECIS-2 to assess external validity and comparing effect estimates**

The authors of this Cochrane systematic review identified 57 trials, 24 trials were subsequently included by the review authors with 28 arms and 58,040 participants (Figure 42). We excluded one trial [183] as we determined this was a case control study not a randomised controlled trial and informed the systematic review authors. Of interest was that the review required the control group to be a placebo or an untreated control so they did not include an active comparison group (as used for example, in the Antihypertensive and Lipid Lowering Treatment to Prevent Heart Attack Trial (ALLHAT) [184]. Although ALLHAT was a randomised, double blind trial of 40,000 high risk hypertensive patients, there was no placebo control group. The fact that only placebo or untreated control was the comparator may partly explain why 15 out of 23 of the trials were explanatory. Often if more pragmatic design the comparator is usual treatment not placebo as the control group.
Assessment of the trials in the Cochrane hypertension review found no clear link between design approach (i.e. pragmatic or explanatory) and risk of bias (Table 8.3, 8.6). Considering external validity, using the software on the PRECIS-2 website, the included trials were visually displayed as
“thumb nail wheels” This enabled assessment and comparisons “at a glance” of the included trials in this Cochrane systematic review. KL determined there were 15 more explanatory trials and eight more pragmatic trials included in this systematic review on hypertension.

Table 48 Summary of Cochrane Risk of bias scores for 23 explanatory and pragmatic trials

<table>
<thead>
<tr>
<th>Cochrane Risk of Bias</th>
<th>Explanatory trials</th>
<th>Pragmatic trials</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>High risk</td>
<td>Low risk</td>
</tr>
<tr>
<td>15 Explanatory trials</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>8 Pragmatic trials</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Details on Cochrane Risk of Bias for the 23 trials included in this review are included in Appendix Table 8.3. Table 48 indicates there were five explanatory trials which had overall Cochrane Risk of Bias overall scores of “low risk” compared to three pragmatic trials with “low risk” scores, seven trials with “unclear” overall scores” for explanatory compared to five with “unclear” overall scores for pragmatic trials. There were three “high risk” scores for explanatory trials and none for the pragmatic trials.

Determining the overall score was not particularly easy as a judgement is required and sometimes there was inadequate or unclear information. However guidance from Higgins et al was followed [46], judgement calls were used to select “low risk scores” where all the domains were judged to be low risk, or only one was unclear. “High risk of bias” was used if two or more domains were judged to be “high risk of bias”. “Unclear risk of bias” usually had three or more domains that were “unclear”. If details were included in the trial articles to assist in judging the internal validity and risk of bias for each domain they were detailed in Appendix Chapter 8 Table 8.3.

With regard to judgements within the different domains of risk of bias and internal validity assessments, KL found that there was inadequate information (occasionally conflicting information) in 46 trial bias judgement domains out of a potential of 138 judgments on random sequence allocation,
allocation concealment, blinding, incomplete outcome data, selective outcome reporting and other sources of bias. Thus one third of the Cochrane Risk of Bias domains in the trials included in the Cochrane systematic review were given by KL internal validity assessments of “Unclear”.

There were no obvious differences in the estimates of treatment effects of pragmatic and explanatory trials included in this Cochrane systematic review to treat hypertension (Figure 43, Appendix Table 8.6). The interventions were similar (in that they were all anti-hypertensive pharmaceutical treatments) with the goal to reduce blood pressure. Drug treatments included Beta blockers, ACE Inhibitors, Ca^{2+} channel blockers, High-dose thiazide and low-dose thiazide. The primary outcome was generally mortality with similar participant baseline severity as trial inclusion criteria; the variables were the degree of pragmatism and estimates of treatment effect – as Risk difference if possible (Appendix Table 8.6). PRECIS-2 was used to score the trials to assess external validity and PRECIS-2 tables used to give the rationale for scoring domains, two examples are included in the Appendix Chapter 8, Table 8.4 HYVET [185], Table 8.5 and Wolff [186]. These tables all give the rationale for decision making in assessing how pragmatic or explanatory the trials were and consider the external validity of the trials.

**Quality assurance of data extracted for Method 1 and 2**

There was good agreement for duplicate data extraction between KL and ST for 6 out of 28 randomly selected pairs of the matched pragmatic and explanatory trials (Table 8.4, Appendix Table 8.2 includes duplicate scores). There were, however, different scores in Cochrane Risk of Bias assessment for Blinding for instance with “Unclear” KL with explanation and “High risk” ST which lead to different overall judgements. ST also rated *Flexibility (Adherence)* in a surgical trial for stents as very pragmatic whereas KL did not score this domain as not relevant to the patient. (The BMJ paper describing PRECIS-2 [153] now suggests that the domain should not be scored in surgical trials and is discussed in Chapter 7b). ST and KL always agreed on the description of overall pragmatism despite some differences in individual domains.
Figure 43 Forest plot of 23 Cochrane reviews indicating explanatory/pragmatic trials.
Duplicate scoring was also done by KL and ST to check the Cochrane Risk of Bias and the PRECIS-2 scores for the Cochrane systematic review “First line drugs for hypertension” using a random sample of 4 out of 23 of the RCTs (Appendix Table 8.2 includes duplicate scores). There was good agreement with identical overall scores for three out of four (with only one disagreement[187]; KL assessed as “High risk” due to “High risk” scores for “incomplete outcome data” and “other sources of bias” whereas ST “Unclear” Risk of bias as “Unclear” for both of these questions. Considering the scoring of external validity, there was overall agreement on whether a trial was pragmatic or explanatory for three out of four but one trial for ST was hard to determine as he had not been able to score several domains [187]. KL and ST did have different scores for some PRECIS-2 domains for the four trials with duplicate scores, which could be seen in the PRECIS-2 “wheels”. This was partly due to ST not scoring at all if uncertain instead of scoring “3”.

The Cochrane review authors’ “Concealment of Allocation” Risk of bias scores were not identical to those of KL. Eight of the assessments by KL were different i.e. eight out of 23 (35%). Three of the decisions previously documented in the published Cochrane systematic review tables as “high” risk of bias were different to the rating by KL for “allocation of bias” as “unclear” and five rated as “low” risk of bias by the Cochrane systematic reviewers were rated as “unclear” by KL. Two other decisions in the Cochrane Risk of Bias assessment for “incomplete outcome data” that had previously been assessed as “low risk” by the systematic reviewers were also rated as “high risk” and “unclear” by KL based on the data available.
### Table 49 Fourteen matches of cardiovascular trials: Estimates of treatment effect/Risk of bias table

*With duplicate results in red (ST)*

<table>
<thead>
<tr>
<th>Study</th>
<th>Effect size - Pragmatic</th>
<th>Risk of Bias - Prag</th>
<th>Effect size - Explanatory</th>
<th>Risk of bias - Explan</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bots(rather Pragmatic)</td>
<td>RD 17.16% (p &lt; 0.001)</td>
<td>RD 11.15% (p &lt; 0.001)</td>
<td>Unclear</td>
</tr>
<tr>
<td></td>
<td>/Davidson(rather Explanatory)</td>
<td>RD 22.9% (p &lt;0.001 )</td>
<td>RD  22% (10mg) (p )</td>
<td>Unclear</td>
</tr>
<tr>
<td></td>
<td>Intervention: Rosuvastatin 10mg</td>
<td>Primary Outcome: Achieving European LDL-C goal (&lt;3.0 mmol/l) at week 12</td>
<td>Outcome: Percent change in LDL cholesterol from baseline to week 12</td>
<td>High</td>
</tr>
<tr>
<td>2</td>
<td>Brady – explanatory=pragmatic)</td>
<td>RD -0.34 % (NS)</td>
<td>RD -1.84 % (P = 0.57 NS)</td>
<td>Low Risk</td>
</tr>
<tr>
<td></td>
<td>/Yang(rather Explanatory)</td>
<td>Outcome: Fatal and non-fatal cardiovascular events (namely MI, unstable angina, ventricular tachycardia, or stroke) within 30 days of operation</td>
<td>more explanatory than Brady but still bit pragmatic (3)</td>
<td>Low risk</td>
</tr>
<tr>
<td></td>
<td>Intervention: Metorolol 100mg</td>
<td>At 2 hours (P = 0.006) increase in antioxidant capacity after consumption of EVOO, difference in serum antioxidant capacity between EVOO and Corn Oil group (P=0.013)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Covas (rather Explanatory)</td>
<td>Mean changes for oxidized low-density lipoprotein level after 3 weeks were P = 0.014 – 3.21 U/L (-5.1 to -0.8 U/L) for the high-polyphenol olive oil.</td>
<td>Low risk</td>
<td></td>
</tr>
<tr>
<td></td>
<td>/Bogani(very Explanatory) (1)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Intervention = extra virgin olive oil</td>
<td>(*measurement time outcome different)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Study Author</td>
<td>Intervention</td>
<td>Response Time</td>
<td>Primary Outcome</td>
</tr>
<tr>
<td>---</td>
<td>--------------</td>
<td>--------------</td>
<td>---------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>4</td>
<td>Dorman (Pragmatic)</td>
<td>Nicardipine (NIC) 0.1mg/ml Or sodium nitroprusside (SNP) 0.2mg/ml</td>
<td>NIC 15min/ Nitroprusside (SNP) 30min (p &lt; 0.01)</td>
<td>Rapidity and variability of blood pressure control</td>
</tr>
<tr>
<td>5</td>
<td>Goy (Pragmatic)</td>
<td>Paclitaxel-eluting stents (*different comparators – plain bare metal)</td>
<td>RD 1.88%. NS RD 2% NS</td>
<td>Reduction in major cardiac events (MACE)</td>
</tr>
<tr>
<td>6</td>
<td>Kaiser (Pragmatic)</td>
<td>Sirolimus-coated Cypher or paclitaxel-coated Taxus drug-eluting stents</td>
<td>RD -3.56 % (NS) OR 0.56 (95% CI 0.35 – 0.91)</td>
<td>Reduction in major cardiac events (MACE)</td>
</tr>
<tr>
<td>7</td>
<td>Kedhi (Pragmatic)</td>
<td>Everolimus-eluting stents</td>
<td>RD 3.29% (NS) 2ndry outcome used: MACE 12 months</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Koren (Rather Pragmatic) / Jones (Rather Explanatory)</td>
<td>Mean change 34.3% (p &lt; 0.0001) RD 0.92%</td>
<td>unclear</td>
<td>Mean change 38% (p = 0.0001) RD 12.5%</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td></td>
<td>Intervention = atorvastatin max dose 80mg/day</td>
<td>Primary outcome: % reduction in LDL-C levels</td>
<td></td>
<td>Primary outcome: Mean change in plasma LDL cholesterol from baseline to the end of treatment (8 wks).</td>
</tr>
<tr>
<td>9</td>
<td>Piller (Rather Pragmatic) / Grimm (Rather Explanatory)</td>
<td>RD 0.39% (doxazosin vs Lisinopril and amlodipine)</td>
<td>Low risk</td>
<td>doxazosin lowered (-19 and -16 mm Hg); HCTZ (-22 and 15 mm Hg)</td>
</tr>
<tr>
<td></td>
<td>Intervention: diuretic hydrochlorothiazide (HCTZ) (Chlorothiadone) 25mg</td>
<td>Primary outcome: Cardiovascular disease (CVD) mortality (death due to coronary heart disease (CHD), Stroke, heart failure (HF), or other CVD).</td>
<td></td>
<td>Primary outcome - Composite: blood pressure, biochemistries, lipids/lipoproteins, quality of life, ambulatory electrocardiograms, echocardiograms, adverse experiences, and drug adherence</td>
</tr>
<tr>
<td>10</td>
<td>Schiarit (Pragmatic) /</td>
<td>Effect size – 1 year</td>
<td>Unclear</td>
<td>Effect size 30 days – 7.6% tirofiban group and 6.0% in abciximab group;</td>
</tr>
<tr>
<td>Study</td>
<td>Intervention</td>
<td>Comparator</td>
<td>Primary Outcome</td>
<td>Hazard Ratio</td>
</tr>
<tr>
<td>-------</td>
<td>--------------</td>
<td>------------</td>
<td>----------------</td>
<td>--------------</td>
</tr>
<tr>
<td>Topol (Explanatory)</td>
<td>Tirofiban bolus 10 µg per kilogram of body weight followed by an infusion of 0.15 µg per kilogram per min for 18-24 hours</td>
<td>Eptifibatide</td>
<td>Incidence of composite ischaemic events within one year</td>
<td>1.26</td>
</tr>
<tr>
<td>Smits (prag) with Grube (Explanatory)</td>
<td>Everolimus-eluting stents</td>
<td>Paclitaxel stent vs Everolimus stent:</td>
<td>Composite of all death, nonfatal MI, and TVR at 12 months.</td>
<td>RD 2.95 (p=0.02)</td>
</tr>
<tr>
<td>Suh (explanatory=pragmatic /Douglas (Explanatory))</td>
<td>Clopidogrel 75mg so TAT (aspirin, clopidogrel and cilostazol 100mg BID)</td>
<td>Composite of major adverse cardiovascular events, cardiac death, non-fatal MI, clinically driven target lesion revascularization (TLR) and ischemic stroke at 6 months</td>
<td></td>
<td>RD -0.64% p=0.73</td>
</tr>
<tr>
<td>Von Birgelen (Rather Pragmatic/Serruys (Very pragmatic))</td>
<td></td>
<td>Target-lesion failure - composite of death from</td>
<td></td>
<td>RD 0.11% p=94</td>
</tr>
</tbody>
</table>

*Note: *Different comparator and outcome.
<table>
<thead>
<tr>
<th></th>
<th>Intervention: Zotarolimus-eluting stents <em>(unable to match with explanatory trial)</em></th>
<th>Primary outcome: Acute coronary syndromes at 12 months</th>
<th>cardiac causes, any myocardial infarction (not clearly attributable to a non-target vessel) or clinically indicated target-lesion revascularisation at 12 months</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>Zhu (Rather pragmatic) /Olsson (very Explanatory)</td>
<td>RD 17.16% (p&lt; 0.001) Primary outcome: Achieving European LDL-C goal (&lt;3.0 mmol/l) at week 12</td>
<td>Unclear</td>
<td>Reduction in LDL-C 46% 5mg and 50% 10mg vs 39%, both P &lt; 0.001</td>
</tr>
</tbody>
</table>

NS: Not Significant; RD: Risk Difference; Prag: Pragmatic; Explan: Explanatory; LDL-C
Discussion

Objective 1: Matched explanatory and pragmatic trials

One of the reasons for creating a database of pragmatic trials was to compare pragmatic and explanatory trials as pragmatic designs have been criticized for sacrificing internal validity for external validity [43, 188] although we were not aware of prior evidence one way or the other for this.

Our matching work attempted to provide some evidence by using the Cochrane Risk of Bias tool to assess internal validity of matched trials. Matching was challenging. It was hard to find trials with the same intervention and outcome but with different approaches to design. Many of our attempts at matching failed because of different primary outcomes (physiological measures vs. usual care interaction) or because the follow-up periods were different (short vs. long). If outcomes were matched, follow up periods were often very different. This is in part due to different designs (more pragmatic vs. more explanatory choices of outcome and measurement point) and the fact that many of the cardiovascular trials were older, so design decisions for these cardiovascular trials appears more explanatory; thus physiological measures are more likely to be used with shorter, more intense follow up.

We did, however, find some matches to then assess the internal validity, external validity and estimates of treatment effect of trials taking pragmatic and explanatory design approaches.

OBJECTIVE 2, 3 and 4: Measuring risk of bias to assess internal validity, PRECIS-2 to assess external validity in a set of explanatory and pragmatic trials and comparing effect estimates

The results from both matching explanatory and pragmatic cardiovascular trials and the Cochrane systematic review did not provide evidence to indicate that pragmatic approaches sacrifice internal
validity. More work is required but this is the first ever attempt to provide evidence to support or refute a criticism widely made of a design approach that sets out to design trials that are highly applicable to routine care. We were not, however, able to do the meta-regression on effect size estimates because of considerable heterogeneity between matched trials, particularly their choice of primary outcomes and comparators. We were therefore not able to answer the question posed in the development of the original PRECIS tool [65] that more explanatory trials with lower PRECIS-2 scores would have larger estimates of treatment effect than more pragmatic trials having higher PRECIS-2 scores. To detect the smaller treatment effects in pragmatic trials, larger sample sizes are often needed [43]. Thus the assumption that estimates of treatment effect may decrease as trials are scaled up from highly controlled explanatory trials to pragmatic trials with greater external validity [189] was not proved.

**Cochrane review**

The work using the Cochrane systematic review “First line drugs for hypertension” to assess external and internal validity of the included trials had an unexpected result. There were striking differences in the PRECIS-2 wheels for the trials included in this Cochrane hypertension review [180] (Appendix Table 8.6). This review identified 57 trials and included 24 trials involving 58,040 participants. Our results suggested that as 65% (15/23) of the trials are more explanatory in nature (testing the intervention under ideal conditions) the review results are likely to be less applicable to routine clinical practice, a fact that is not highlighted in the review itself (as in many other reviews). As the Cochrane reviewers used placebo or an untreated control this also meant that more explanatory trials were included in the review. Pragmatic trials often do not include placebo controls as that is less common in usual care, the control is often the current treatment. However, as systematic reviews are the starting point for guideline development for healthcare practitioners in routine practice, the implication of these findings for review interpretation are worth considering further. By their very nature systematic reviews collate trials of varying degrees of pragmatic and explanatory orientations to cover broad
inclusion criteria that usually represent the real world that clinicians are interested in. It is however, interesting in this review that older trials using populations with narrow inclusion criteria, run in periods, carefully controlled protocol administration, compliance monitoring and often exclusion based on compliance, together with extensive follow-up may have been used to guide the treatment of people with hypertension quite unlike those in the trials. As Dahlen in his editorial stated, discussing the trial by Price [124] “Asthma Treatment Guidelines Meet the Real World”, it is very important that RCTs in real-life populations are undertaken to help people pursue their daily life and these trials should be the basis for guidelines [190]. The well documented side effects of hypertension medication are important considerations for assessing whether treatment to reduce cardiovascular risk (particularly in borderline hypertension) offers benefit when weighed against real-world individuals suffering from the undesirable effects of taking regular medication.

Work by Travers also demonstrated that in COPD [22] 90% of patients being treated would not have been eligible for the RCTs that formed the basis of guidelines guiding their care and even fewer (a median of 6%) of asthma patients would have been eligible for asthma trials [26]. Similar results have been found in diabetes and breast cancer [25, 27]. Further work using PRECIS-2 on Cochrane reviews of different subject areas to consider the external validity and applicability of trial results would be worth investigating as the implications of mixing design approaches are currently unclear.

**Overall thoughts on internal and external validity**

Ware [43], Ernst and Canter [44] have made statements suggesting that pragmatic trials are intrinsically poor with regard to their internal validity which may bias the clinical outcome results to give the desired effect rather than the true effect. To address this criticism, we attempted to compare the internal validity of a cohort of matched pragmatic and explanatory trials.

Ensuring a balance between internal and external validity in pragmatic trials in “real clinical settings” is difficult [191]. Assumptions are made, most recently by Sedgwick, that explanatory trials minimise
confounders so that trialists are sure the intervention is the causal effect but pragmatic trials cannot [42]. Discussing a trial on leg ulcers undertaken in the community [192] Sedgwick stated that a pragmatic trial...“would have been undertaken in a routine clinical and healthcare setting and would be expected to have “high external validity” and “low internal validity”. This statement suggests that Sedgwick, like others, is confused about that “internal validity” really is.

This work on a cohort of matched cardiovascular trials and the trials included in a Cochrane hypertension review do not support (or refute) the belief that trials that take a more pragmatic approach sacrifice internal validity. In the matched pair study 10 had similar scores for explanatory and pragmatic and there was little difference in scores for pragmatic and explanatory trials in the Cochrane review of trials testing an intervention to determine the treatment effect. In fact in the Cochrane review, three of the trials that were scored as explanatory using PRECIS-2 had “high risk of bias” scores by KL (although one of these was scored by ST as “unclear”). Sedgwick does not back up his theory with evidence to demonstrate why pragmatic trials have low internal validity but states: “…if the trial lacked internal validity, making it difficult to infer a causal association between treatment and outcome, then external validity might have been limited.”[193] One could argue “external validity WOULD be limited”.

The CONSORT extension for pragmatic trials (Table 3, p3 [32] includes internal validity issues under Methods and Results: Randomisation-sequence generation, Randomisation- allocation concealment, Randomisation- implementation, Statistical methods, Numbers analysed, Outcomes and estimation – for all of these items there is no difference between the standard CONSORT guidelines for more explanatory trials and for pragmatic trials [152]. The only item that is different is Blinding (masking) with the statement “if blinding was not done, or was not possible, explain why” so the expectation is that to ensure internal validity, blinding should be carried out wherever possible or there should be a good reason why this was not done.”
Good reasons would include a desire to combine the effect of the intervention with the effect of the end user’s belief in the intervention they receive i.e. not separate that out, so that the trial effect size would be the same as in the real world. So for instance, in real world situations, the patient is never blinded, and so trialists could choose to replicate that situation and not eliminate the effect of belief in a pragmatic trial. Independent outcome assessment is also not something that occurs in usual care and is contentious (Chapter 5 Brainstorming) but it has been included in restrictions for the Primary Outcome domain in PRECIS-2 as we believe a good pragmatic trial can still be blinded, especially for outcome measures. In fact, in the Cochrane review there were 17 out of 23 “low risk of bias” assessments for blinding with mostly triple blinding. Ernst criticised lack of blinding in trials suggesting that the enthusiasm for an intervention without blinding to prevent the Hawthorne effect may bias the clinical outcome results to give the desired effect rather than the true effect leading the authors to conclude “pragmatic trials can be comparatively weak research tools”[44].

Internal validity is an issue that is of great concern to those designing pragmatic trials, Godwin in his discussion paper on two pragmatic trials in primary care to reduce hypertension, discusses methods to ensure good internal validity [188]. Trial design included cluster randomisation of physicians to the intervention group or control group to prevent contamination or performance bias and doctors influencing the result if a GP cared for patients in both groups. There was also blinding of outcome measurement by research assistants to prevent “assessment bias” and there is blinding of outcome assessment by the statistician analysing the results thus preventing “Detection bias”.

Another group designing a pragmatic trial took a different stance stating that the internal validity of the trial would be compromised if the Cardiovascular Health Awareness Program CHAPS trial [194] was not designed to answer the question posed in the real world i.e. “if no cross-overs were allowed, patients or clinicians were blinded, or artificial follow-up visits were created to minimize loss-to-follow
up”. They suggest that not all criteria used to assess internal validity in explanatory trials are appropriate for pragmatic trials [195]. All of these aspects may however affect the effect size of pragmatic trials.

So while Sedgwick accepts that external validity is increased by including a broad range of participants in the trial who suffer from venous ulcers and require compression bandaging, he suggests this decreases internal validity. “In particular, the internal validity for a pragmatic design will be lower than that for an explanatory design; the participants in a pragmatic trial are a heterogeneous group and may not adhere to the treatment regimen they are allocated [42].” However, I would suggest, it is important that we do not confuse confounding with the reality of practice as voiced by the CHAPS trialists [195, 196].

To conclude, Rothwell states that internal validity is independent of “all aspects of the design and performance that impact on the external usefulness of the result of a trial” - the external validity [45]. He also suggests internal validity underpins the rules for randomised trial design and is thus a given whether pragmatic or explanatory design, whereas external validity and applicability requires clinical judgements from medical practitioners who have clinical expertise [45], as we discovered in the previous chapter when validating PRECIS-2. The issue of what is included in internal validity assessment for both explanatory and pragmatic trials will no doubt need to be returned to, but the essential ingredients of the principles of randomisation are included by both, and using the Cochrane Risk of Bias tool there was no apparent difference in our assessment between pragmatic and explanatory trials.

**Poor Reporting**

Using PRECIS-2 to determine external validity was subject to issues of uncertainty. For instance, if there was uncertainty about scoring due to no or inadequate information then a score of “3” was used (as this was assessing trials in a systematic review) as advised in our Toolkit information on the website.
There was a great deal of information that was missing or inadequately reported and as many of these trials are older (22 out of 24 trials included in the Cochrane systematic review) were published before the CONSORT guidelines for reporting trials that was originally published in 2001, updated and re-published in 2010 [152]. That reporting is a problem is well-known: a Cochrane methodology review evaluating 53 publications looking at a total of 16,604 RCTs concluded that reporting remains sub-optimal with assessment of internal validity judged to be mainly “unclear” [197].

The issue of poor reporting was highlighted in the high proportion of “unclear” judgements for “bias” in internal validity assessments in our study. Many of these trials are older and pre-CONSORT, demonstrating how important the guidelines are to ensure there is adequate reporting of trial details [198]. Only four of the trials were published after 2010 though only two of these trials had sufficient information to categorise trials as “low risk” (Smits (2011)[199]; Von Bergelen (2012 [200]), and two published in 2011 were unclear (Suh [181]; Schiariti [201]). So indicative that reporting could still be improved.

Of particular interest was the number of “unclear” scores for the “Selective outcome reporting” with 16 in total and one “high risk”. This issue was highlighted by Dwan [202] with some trials not reporting results if not statistically significant or not being measured for various reasons even though these outcomes are part of the trial protocol. The former would have a greater impact on risk of bias in the trial than the latter and would reduce the overall internal validity of the trial.

Quality assurance of internal and external validity assessments of trials

There was generally good agreement in duplicate data extraction but differences in scores are related to different judgements of the available data. There was no re-rating of risk of bias scores and PRECIS-2 scores following discussion but many of the scores could be interpreted as similar. The duplicate scores highlight that consensus scoring using PRECIS-2 is highly recommended to be sure each rater is using the same information to rate a domain and how that is interpreted by individuals. For instance,
ST rated *Flexibility (Adherence)* in a surgical trial for stents as very pragmatic whereas KL did not score this domain as not relevant to the patient. This difference in opinion was also highlighted in the validity study and is now included in the 3 page information sheet for users of the PRECIS-2 tool. ST and KL always agreed on the description of overall pragmatism with some differences in individual domains indicating issues again with judging the score of domain information that is not created by the users of the PRECIS-2 tool (or indeed using a score of “3” if inadequate information to score instead of no score).

**Study Strengths and Weaknesses**

Using PRECIS-2 to *design* trials is best done by a rater who has some clinical expertise in the treatment and condition being investigated. The tool was used by KL, although a practising nurse, does not have nurse prescribing certification and cardiovascular speciality experience. The limitations on assessments of the trials were hampered by lack of published information. However, KL was using PRECIS-2 to assess external validity and applicability using the trial description and not to design a trial so clinical expertise should not have been required. Everyone using PRECIS-2 is dependent on explicit clear information in the trial publication and this is often not present. Over time if the CONSORT guidelines for reporting pragmatic trials are more widely used by trialists as a guide to writing up trials, there should be sufficient detail for decision makers and raters to determine the applicability of an intervention in the RCT.

Using a Cochrane review had the advantage of evaluating a set of trials which would include RCTs for the same interventions, reporting the same outcome – total mortality. Ideally, however, we need to use PRECIS-2 and the Cochrane Risk of Bias tool to assess other Cochrane systematic reviews to have a sufficiently large sample size to draw any correlations between effect size and pragmatic or explanatory trials.

While the Cochrane systematic review may not have included more recently published trials as they did not meet inclusion criteria, hypertension is still very topical as evidenced by the recent publication
by the same Cochrane systematic review authors [84]. This suggested that there may be overtreatment of mild hypertension in people with low risk of death due to cardiovascular disease [203]. The authors put out a call for randomised trials that included use of global outcome scores (a quality measure based on health outcomes, that compares current care to a target level of care) instead of blood pressure thresholds and compared drugs with lifestyle interventions and placebo in patients with mild hypertension, as currently there is insufficient evidence to conclude that drug treatment helps all individuals [203]. Designing a more pragmatic trial to test this complex lifestyle intervention would give results that are relevant to “real world” people for whom doctors are considering whether or not to treat for hypertension, and would assist decision making for people who are wondering whether or not to take drugs in addition to, or instead of, making lifestyle changes. The Cochrane authors were co-authors on this paper and are no doubt more aware than most of the dearth of up-to-date relevant trial research with high external validity to inform clinical practice.

**Conclusions**

The sample of matched explanatory and pragmatic trials was small but this sample, together with the study of the hypertension Cochrane review, is the first attempt to systematically look for differences in internal validity between explanatory and pragmatic design approaches. We did not find evidence to suggest that pragmatic trials sacrifice internal validity for external validity [43, 44, 204]. Further work with a larger sample of Cochrane systematic reviews would be useful in this regard and would also unravel information about the trials included and answer the question which remains unanswered: do pragmatic trials have smaller estimates of treatment effect?
SUMMARY

There has been criticism of pragmatic trials suggesting that they sacrifice internal validity for external validity. To address this question, internal validity of a cohort of matched pragmatic and explanatory cardiovascular was considered using PRECIS-2 to score external validity and the Cochrane Risk of Bias tool to assess internal validity. There appeared to be no difference in internal validity between study designs although a larger sample of matched trials would reduce the uncertainty around this conclusion.

This result was checked by assessing the individual trials (23) in a Cochrane systematic review on hypertension, there did not appear to be reduced internal validity in pragmatic trials when compared to more explanatory trials. It was interesting to note two thirds of the trials were explanatory in design. There has been suggestion that more pragmatic designs have smaller estimates of treatment effect and more explanatory trials larger estimates of treatment effect. We were unable to determine if estimates of treatment effect and additional study variables are influenced by more pragmatic or explanatory trial designs.

CONCLUSION

We found no evidence that pragmatic trials have reduced internal validity compared to explanatory trials, although a larger sample is needed to be more certain of this. Further work also needs to be done to ascertain if estimates of treatment effect are different in pragmatic and explanatory trials.
Chapter 9: APT: Applying PRECIS-2 to Primary Care Trials
Pilot and two Case studies

Background

Through the Delphi, Brainstorming and User testing (Chapters 4, 5, 6) PRECIS-2 had been created, and through reliability and validity testing (Chapter 7A and 7B) PRECIS-2 had been validated and its inter-rater reliability established. We believed PRECIS-2 would be even more useful than the original tool in helping trialists to design trials fit for purpose and, moreover, that PRECIS-2 would enable trialists to consider the needs of the end user of the results and the applicability of the research. As 15 groups have now published using the original PRECIS at various stages of design from creating the trial protocol [53], to determining how pragmatic a trial protocol is [40, 54, 61, 62] to retrospective discussion of trials in systematic reviews [39, 64, 67, 69, 205], we were keen to work with trial groups on the improved tool PRECIS-2 during the “design” stage, at the point of time that PRECIS-2 was developed to be used by trialists. Thus, following validity and reliability testing of PRECIS-2 by 19 raters rating up to fifteen trials we wanted to test PRECIS-2 in “real world” conditions to determine if the tool itself was fit for the purpose intended; making the trial team more aware of their design decisions, improving transparency in design decision making, in particular with regard to applicability of trial results.

The Pragmatic Clinical Trials Unit based at Queen Mary’s hospital in London was keen to collaborate and assist in testing PRECIS-2. The centre has been conducting pragmatic clinical trials since the mid 90’s (http://lizard.qmul.ac.uk/pragmatic-clinical-trials-unit.html). The study was entitled APT: Applying PRECIS-2 to Primary Care Trials. This study was the first phase of the APT project, the second is to discuss with policymakers who are responsible for commissioning new research and primary care practitioners who would potentially be involved in applying new research which aspects of trial design
are important, what they think about pragmatic trials and finally if they believe PRECIS-2 will assist in determining evidence to improve clinical practice.

This study description is based on the protocol prepared by the APT study chief investigator Gordon Forbes with input from the APT steering team (unpublished but submitted to Queen Mary’s for R&D and ethical approval). Rater results and some wheels produced by GF have also been used in the Results for this chapter on use of PRECIS-2 by trial teams. Case study discussion on PRECIS-2 and overall discussion on this chapter is solely the views of KL. As we had agreed that all rater data would be anonymised only the role of the rater within the trial team has been stated.

The PRECIS-2 tool was tested out by trial teams at the PCTU by seven trials. Unfortunately most of the trials were past the design stage, the ideal anticipated time to use the PRECIS-2 tool; the Pilot study COPERS was complete, others were recruiting – e.g. Case Study 1 HepFree – but the STOP trial, at the time of discussion with the trial team, was undertaking a small feasibility study but had full flexibility in using the tool to design the full cluster randomised trial, if the pilot was successful.

Aims and Objectives

Aims
To investigate the value of PRECIS-2 for designing primary care trials that are relevant to practice.

Objectives
1. To assess the views of primary care trial teams towards PRECIS-2.
2. To test viability of an online PRECIS-2 tool and training package use by a trial team.
3. To determine if scoring of PRECIS-2 domains prior to a meeting to discuss trial design and at the end of a meeting assists the trial teams at the Pragmatic Clinical Trials Unit in design decisions.
4. To determine if scoring of PRECIS-2 domains assists trials teams in considering the intended audience and creation of trials relevant to practice.

5. To assess if we can improve the methodology for using PRECIS-2 by trial teams

Methods
We applied PRECIS-2 to several Pragmatic Clinical Trial Unit (PCTU) trials and used mixed methods for focus groups of trialists for each trial team. The rating team used PRECIS-2 to score their trial in two stages; before and after group discussion on the trial design. In the first stage the rating team used the PRECIS-2 web resource www.PRECIS-2.org to familiarises themselves with the PRECIS-2 and submit independent PRECIS-2 scores, either on the web or directly to GF. In the second stage the rating team met face to face to discuss areas of disagreement in the independent scores and following the discussion of each domain, each rater produces a final set of PRECIS-2 scores on paper.

Face to face meetings were organised for two hours with trial teams with a simple agenda (Box 9.1). During this APT study meeting we gathered qualitative information through recording the views of individual trial team members by audio taping the meeting, notes taken by GF and KL and using a short questionnaire at the end of the meeting (Appendix, Chapter 9, Figure 9.1) which provided feedback on participants’ experiences. A Likert scale of 1 to 5 was used with 1 = strongly disagree, 3 = neither agree or disagree and 5 = strongly agree. (This questionnaire was adapted from one originally used by data-driven quality improvement in primary care (DQIP) – Making prescribing safer, to gain qualitative information on the DQIP intervention [206]). We did not get the audio recording transcribed; it was simply to clarify discussion notes. In addition, we undertook quantitative assessment of the PRECIS-2 domains using a Likert 1-5 score for the trial design pre and post discussion. We also gained insight to PRECIS-2 scoring decisions by trialists through their rationale in PRECIS-2 tables online and through discussion in the face to face meeting.
The trial data were not shared and we did not discuss any patient identifiable data. We have ensured that all information from the focus groups was fully anonymised.

**Study committee**
The study committee consists of three members from QMUL - Sandra Eldridge, Gordon Forbes, Steph Taylor and PhD student University of Dundee – Kirsty Loudon and from University of Aberdeen – Shaun Treweek. Meetings were organised by GF and attended by KL who acted as an observer and if necessary gave additional information on PRECIS-2 based on the draft elaboration paper (unpublished) which is not currently part of the web resource.

**Participants**
Potential trial teams for the APT study were trialists or members of trial steering committees involved in the primary care trials being undertaken with the assistance of the PCTU. Participants were identified through contact with the APT study chief investigator GF and contacted via email. For each trial included, we tried to assemble a rating team of up to five people; at least one rater from the APT study team, the trial chief investigator, the trial manager, a member of the trial steering committee and a statistician from the trial. If we were not able to recruit these participants we used other members of the trial team with similar experience or had a focus group with a smaller rating team. For all of the case studies described here, including the pilot, KL attended the APT meetings with trialists.

**Statistical Methodology and Analysis**
We collected descriptive statistics from PRECIS-2 scoring before and after the APT trial meeting and qualitative information from completed questionnaires. No statistical tests were carried out.

**Ethics**
Ethics approval was obtained from the Queen Mary’s University London REC – QMREC1360d and informed consent was sought from all participants. There were no conflicts of interest in the study team.
Data handling and record keeping
Information related to participants was kept confidential and managed in accordance with the Data Protection Act, NHS Caldicott Principles, The Research Governance Framework for Health and Social Care, and the conditions of Research Ethics Committee Approval.

Record Retention and Archiving
All study documents will be archived with the Barts Health archiving service for 20 years.

Applying PRECIS-2 to primary care trials (APT)

Agenda

1. Introductions (5 mins)
2. Aim of the meeting (5 mins)
3. Introduction to PRECIS-2 and the APT study (5 mins)
4. Any questions or clarifications regarding any of the domains (5 mins)
5. Discussion of scores (1 hour 20 mins)
6. Questionnaires on PRECIS-2 (10 mins)
7. Verbal feedback on process (10 mins)

Box 9.1 Agenda for APT trial teams using PRECIS-2 to discuss trial design (GF)
Pilot study – COPERS

Aim
To investigate if PRECIS-2 is a valuable tool to assess the trial design of COPERS, a primary care trial, to ensure relevance to practice. In addition to use COPERS as a pilot study to determine if there is a need to refine methods for the first part of the APT study: the application of PRECIS-2 by trial teams.

Objectives
1. To assess the views of the COPERS trial team towards PRECIS-2.
2. To test viability of the online PRECIS-2 tool and training package use by the COPERS trial team.
3. To determine if scoring of PRECIS-2 domains assists the COPERS trials team in considering the intended audience and creation of a trial relevant to practice.
4. To determine if scoring of PRECIS-2 domains prior to a meeting to discuss trial design and at the end of a meeting assist the COPERS trial team at the Pragmatic Clinical Trials Unit. To assess if we can improve the methodology for using PRECIS-2 by trial teams.

Methods
We applied PRECIS-2 to one trial from the Pragmatic Clinical Trial Unit trial, COPERS – (Coping with persistent pain, effectiveness research into self-management) - a trial of a novel group self-management course for adults with chronic musculoskeletal pain [207, 208]. This trial had already been designed and the protocol published [208] with a separate paper indicating the proposed analysis plan [207]. At the time of the PRECIS-2 meeting in the PCTU, 10th April 2014, the trial intervention had been tested out and the participants were post trial, analysing data and writing up. An audio recording was made of the meeting to assist in reviewing the discussion after the meeting. In addition, notes were taken by GF and KL.

Participants
All the participants invited accepted the invitation and the meeting was held within five weeks of invitations being emailed. The rating team consisted of five people; one rater independent to the trial (GF – the Principal investigator for the APT study), the chief investigator, the trial manager, a member
of the trial steering committee and a statistician from the trial. KL took part in the discussion, giving further information on PRECIS-2 and the meeting was also attended by Sandra Eldridge as the Chief Investigator for the APT study.

**PRECIS-2 scores**
Two rounds of independent scoring were undertaken. All COPER participants used the online version of PRECIS-2 [www.PRECIS-2.org](http://www.PRECIS-2.org). Participants were instructed to read the information on PRECIS-2 contained in this website, no other information on PRECIS-2 was provided, and then use the tool to score the COPER trial.

Following the online scoring, we held a two-hour face to face meeting with the COPER participants. We used the simple Agenda outlined in Box 9.1 to ensure timing and meeting objectives were met. First there was a short introduction on PRECIS-2, then the scores from the first round were shared and a discussion was held with the aim of improving consensus in PRECIS-2 scores. The group discussed the domain with the most disagreement first and then the domain with the next most disagreement and so on until finally discussing the domain with the least disagreement last. Participants independently recorded PRECIS-2 score for the second time for each domain as the discussion progressed. The final scores presented in the Results are the median from these second round scores. The median was selected as we believed this measure clearly indicated the most common score given for a particular domain. A mean may not have been a whole number of “1” to “5” and therefore would not have a meaningful value.

**Questionnaire and Feedback**
At the end of the rating meeting the raters completed a short questionnaire (Appendix Chapter 9, Figure 9.1) providing feedback of their experiences with the PRECIS-2 tool. This questionnaire is the basis for the qualitative feedback for the main part of the first part of the APT study. Participants were also asked to provide any verbal feedback on the PRECIS-2 scoring process or questionnaire. GF contacted all COPER participants by e-mail to thank them for attending the APT meeting and to send them the final PRECIS-2 wheel for their trial, along with PRECIS-2 scores for each participant as
well as the median scores for each domain. If any of the trialists had questions regarding the PRECIS-2 tool or APT they were encouraged to get in touch and they were also told they were “free to use the PRECIS wheel (Figure 9.7) in any way that suits”.

Results
Below are the initial (Table 50) and final (Table 51) PRECIS-2 scores for COPERS and the corresponding PRECIS-2 wheels (Figure 45, Figure 44).

The initial round of scoring (Table 50) gave COPERS more pragmatic scores for six domains, neither pragmatic nor explanatory for two domains and more explanatory scores for one domain. There was a large amount of disagreement within the scores with one domain (Flexibility delivery) receiving scores at both ends of the Likert scale from “1” to “5”, three domains with scores ranging across four points of the scale (Recruitment, Setting, Follow-up), three domains with scores ranging across three points of the scale (Eligibility, Organisation, Primary outcome) and two domains with complete agreement (Flexibility Adherence, Primary analysis).

<table>
<thead>
<tr>
<th>Table 50 Initial Scores for COPERS using PRECIS-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scores from the first round of scoring, produced independently by the 5 raters online</td>
</tr>
<tr>
<td>n = 5</td>
</tr>
<tr>
<td>Median</td>
</tr>
<tr>
<td>--------</td>
</tr>
<tr>
<td>4.5</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>5</td>
</tr>
</tbody>
</table>

In the second round of independent scoring of the nine PRECIS-2 domains for the COPERS trial, following discussion, raters gave pragmatic scores to seven domains, neither pragmatic nor explanatory to one domain (Recruitment) and an explanatory score to one domain (Organisation) although one rater withheld scoring (so this was entered as “-1” as the score for this trial domain).

There was less disagreement between raters than with the initial scores (Figure 9.1) with complete agreement on three domains (Flexibility adherence, Follow-up, Primary analysis), a two point range on five domains (Recruitment, Setting, Organisation, Flexibility delivery, Primary outcome) and scores covering a three point range for one domain (Eligibility).
Table 51 Final Scores for COPERS using PRECIS-2

Scores produced by the COPERS participants following the consensus discussion.

<table>
<thead>
<tr>
<th></th>
<th>Eligibility</th>
<th>Recruitment</th>
<th>Setting</th>
<th>Organisation</th>
<th>Flex. Delivery</th>
<th>Flex. Adherence</th>
<th>Follow-Up</th>
<th>Primary Outcome</th>
<th>Primary Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Min</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Max</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

Using the PRECIS-2 website a PRECIS-2 wheel for COPERS (Figure 46) was produced from the final scores using all the scores from the raters, indicating range of scores: https://crs.dundee.ac.uk/precis/Trials/Details/45

This is currently only visible to account holders: User Name: COPERS, Password: APTPilot123 and has not been released into the database. (In one of the domains, Organisation, one of the raters had not entered a score so this was entered as “-1” as the score for this trial domain. (“-1” was the designated best way to indicate a score had not been accidentally missed but the domain had been purposively not scored by the rater due to uncertainty.)
Questionnaire Results

The results from the questionnaire are shown in the tables below (Table 52 and Table 53). The questionnaire was completed by the four participants (the chief investigator, the trial manager, a member of the trial steering committee and a statistician from the trial team). Three out of four participants neither agreed nor disagreed with the statements concerning the online tool and training material. There was strong agreement that the meeting to discuss the scores lead to better scores being produced and agreement that the tool would have been beneficial to use in the design phase.

All participants agreed or strongly agreed that the tool highlighted areas of design which are important for the trial to achieve its goals. There was a range of responses across the domain specific questions (7.1 - 7.9) with mostly positive answers. The domain with strongest agreement on its relevance was Eligibility.
The free text feedback highlighted concerns over interpretation of the domain definitions as well as practical issues completing the online tool.

Table S2 Completed Questionnaire for COPERS – APT study pilot
SA = Strongly Agree, A = Agree, N = Neither agree nor disagree, D = disagree, SD = strongly disagree.

<table>
<thead>
<tr>
<th>No.</th>
<th>Question</th>
<th>No. of Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The online training material contained sufficient information about PRECIS-2</td>
<td>SA 1 A 3 N 1 D 1</td>
</tr>
<tr>
<td>2</td>
<td>The online information on how to use the tool is clear and concise</td>
<td>SA 1 A 3</td>
</tr>
<tr>
<td>3</td>
<td>The online tool is easy to use</td>
<td>SA 1 A 3</td>
</tr>
<tr>
<td>4</td>
<td>PRECIS-2 would have been useful to use in the design stage of the trial</td>
<td>SA 1 A 3</td>
</tr>
<tr>
<td>5</td>
<td>Meeting to discuss independent scores lead to more accurate scores being produced than were produced independently</td>
<td>SA 3 A 1</td>
</tr>
<tr>
<td>6</td>
<td>PRECIS-2 highlights areas of trial design which are important for your trial to achieve its goals, be that informing clinical decision making or increasing knowledge of how an intervention works.</td>
<td>SA 1 A 3</td>
</tr>
<tr>
<td>7</td>
<td>How important are each of the PRECIS-2 domains in ensuring the results from your trial are relevant to their intended audience</td>
<td>SA 4 A 1</td>
</tr>
</tbody>
</table>

|    | Eligibility - Who is selected to participate in the trial? | 4 |
| 7.1| Recruitment - How are participants recruited into the trial? | 2 A 1 D 1 |
| 7.2| Setting - Where is the trial being done?                   | 3 A 1         |
| 7.3| Organisation – What expertise and resources are needed to deliver the intervention? | 3 A 1         |
| 7.4| Flexibility - How should the intervention be delivered?    | 3 A 1         |
| 7.5| Flexibility - What measures are in place to make sure participants adhere to the intervention? | 2 A 2         |
| 7.6| Follow-up - How closely are participants followed-up?      | 1 A 1 1 D 1   |
| 7.7| Primary outcome - How relevant is it to participants?      | 2 A 1         |
| 7.8| Primary analysis - To what extent are all data included?   | 2 A 2         |

Table S3 Free text feedback for COPERS – APT study pilot

<table>
<thead>
<tr>
<th>Participant</th>
<th>Feedback</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Great exercise. Very educational. Highlights the importance of a detailed and clear guide on how to use PREICS-2. I think there will continue to be much debate on how to apply PRECIS-2.</td>
</tr>
<tr>
<td>2</td>
<td>Using the online system: Abstract - you only need a brief description not an abstract. Need a definition of pragmatic and explanatory. Trial appears at the end of the list, needed to look for it. Best at top of the list? Once scored trial asked to press create (like a new score) submit might be better. Comment against Q6: No more than other things, not sure.</td>
</tr>
<tr>
<td>3</td>
<td>When I did the online questionnaire I thought I understood the PRECIS domains - apart from difficulties I had/have with the wording of the flexibility (delivery) domain. Having done this exercise I realise there is more to the domains and my initial understanding of the domains was superficial.</td>
</tr>
<tr>
<td>4</td>
<td>Some domains were confusing - would make using PRECIS to design a trial difficult as the trial group would spend half the meeting arguing over what each domain meant.</td>
</tr>
</tbody>
</table>
Discussion

Face to face meeting
The APT pilot meeting on COPERS was completed within the time allocated in 1 hour and 20 minutes. All of the participants felt that the discussion was beneficial as there were different interpretations within the group for the domain definitions, what comprises usual care. Participants had different knowledge of the trial depending on their input. Valuable contributions were made by all participants and the discussion would not have been possible without the range of backgrounds present – knowledge of the usual care setting came from the clinician present, detailed knowledge of the trial from the trial manager and chief investigator, technical knowledge of the analysis from the statistician and further knowledge of PRECIS-2 was provided by GF and KL.

Most of the domain uncertainty was over the PRECIS-2 domain *Flexibility of delivery* (range 3 to 4 with median score rather pragmatic). Even at the end of the meeting there remained some uncertainty over the definition of this domain amongst the participants. The toolkit stated “*Flexibility (delivery)* – how different is the flexibility in how the intervention is delivered and the flexibility likely in usual care? For example, score 5 for a very pragmatic choice with identical flexibility to usual care; score 1 for a very explanatory approach if there is a strict protocol, monitoring and measures to improve compliance, with specific advice on allowed co-interventions and complications.” This may in part have been due to the uncertainty over what exactly was usual care and although the trial manager offered most insight in this domain, there remained some ambiguity. It appeared to KL, as observer, there was full flexibility and it was often tailored to the individuals attending the training on self-management of pain and this would probably continue in usual care. There was monitoring but there were no measures to ensure strict adherence to the delivery. This is probably what would occur in usual care, though there may be a tendency to be slightly less flexible and adhere to the protocol in the trial more rigidly, than if the intervention was implemented post-trial.
It was apparent from the discussion that a user guide to PRECIS-2 (the PRECIS-2 elaboration paper) which had been prepared and was about to be submitted for publication would be very helpful to clarify to participants how to use PRECIS-2 if there was uncertainty. In addition, example trials with scoring for each domain would also assist trialists designing a trial or indeed, as in this case, assessing the trial they had designed retrospectively. Both of these resources will be placed on the PRECIS-2 website when the elaboration paper has been published [153]).

The discussion of the score for each domain worked best when it followed the following form:

1) What the domain definition means and what would a trial look like which scored a more pragmatic compared to more explanatory score.

2) What happened in the trial?

3) What score should the trial be given?

In the pilot we structured the meeting to begin discussing the PRECIS-2 domain with the greatest score range indicating disagreement, progressing through to the final domain that raters disagreed least on. We did this to ensure that there would be discussion of the most contentious domains first ensuring there would be adequate time. Unfortunately this meant that this was not an easy start to the meeting and facilitating was a little difficult. Starting with an easier domain to initiate PRECIS-2 discussion would have been easier. In addition, switching between domains and not working logically round the PRECIS-2 wheel in order created some minor confusion in recording the scores by individual raters.

**Questionnaires**

All participants except one were able to complete the questionnaire unaided. There was some confusion from one of the participants over what was meant by Question 7 “How important are each of the PRECIS-2 domains in ensuring the results from your trial are relevant to their intended audience?”. The participant did not know who the intended audience was and as a consequence scored every domain as a 5 as they felt that every aspect of the design was essential. The other participants understood what this question was asking and provided different scores for different domains. We discussed changing the wording of this question to specify policymakers and health care practitioners.
but decided to leave this up to the individual trialists to determine who the “intended audience” was and answer the question as they saw fit. Generally, COPER trialists believed that PRECIS-2 domains did help them consider the end user of their trial results.

**Changes following pilot**

Based on the pilot the following changes were made:

1) Change the instruction email on using the online tool to include instructions on what details are necessary when entering a new trial to focus on scoring trial using PRECIS-2.

2) Change the order of the discussion to run in the order of the domains rather than in order of disagreement.

**Web changes**

KL worked with LT (PRECIS-2 website software designer) on three issues to assist website users. Firstly, the PRECIS-2 wheel now shows a range of scores (Figure 47) with a red line for minimum and maximum scores; it is now possible to click on “Show score range”, otherwise the default median scores are simply displayed. Secondly, an error in the scoring range was noticed by KL and this was remedied by LT. For example, is not possible to leave the score for a domain blank – all score fields must be completed using the PRECIS-2 tool. If the field is left blank, there is an automatic message “A score is required, or use -1 to avoid entering a score for this domain.” Using the tool for the COPERS trial, it became apparent that this “-1” score had been incorporated in the range as “(-1-2)” for Organisation. However, this notional score is now not included in the median score and range and for the pilot trial Organisation domain is now “2(1-2)”. It is also possible for raters to insert “0” if they do not know what score to give a PRECIS-2 domain as described in the PRECIS-2 elaboration paper which gives advice to trialists on scoring a trial domain if there is any doubt: “If there is uncertainty over how explanatory or pragmatic your proposed trial design element is for a particular domain, then we suggest the score for this domain should be left blank; this will then highlight uncertainty and encourage discussion. Domains which have greater scoring variation by raters probably require further
discussion and reiteration of Step 1 to 4 to clarify the design of this domain so there is greater agreement.”

Finally, to assist in PRECIS-2 users simply using the PRECIS-2 tool, and not adding a trial to the database, we minimised the fields that had to be completed when creating a new trial. These are indicated with a * and do not include the Abstract – simply the “Title of trial”, the “Intervention” which is a drop down list and further information needs to be briefly given in “Intervention description”, “Number of sites”, “Countries involved”, “Sample size”, “Primary outcome” – knowledge of most of these fields would be used to score the nine different PRECIS-2 domains. The users at this stage would also not unclick “Keep private” so using the website to enter their scores and create a wheel was purely for their use and for GF and KL to register trial scores.

[Diagram of PRECIS-2 wheel]

Figure 47 PRECIS-2 wheel created by website tool showing ranges and maximum and minimum scores, on COPERS, post APT pilot
Scores:
(shows median if more than one score was entered)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>4 (3-5)</td>
<td>3 (3-4)</td>
<td>4 (3-4)</td>
<td>2 (1-2)</td>
<td>4 (3-4)</td>
<td>5 (5-5)</td>
<td>4 (4-4)</td>
</tr>
</tbody>
</table>

Conclusions
The pilot was a success; we received useful feedback on the PRECIS-2 web resource and on trialists using the PRECIS-2 tool to evaluate trial design decisions to consider applicability. We recruited all the participants from the COPER trial for the APT meeting that we had invited. All participants read the online training material and produced a set of scores in advance of the meeting. The discussion was lively with all participants actively contributing and there was an improvement in the level of consensus in the PRECIS-2 scores from initial to final scores, post discussion.

The discussion reinforced the importance of a multidisciplinary rating team including people with detailed knowledge of PRECIS-2, clinical practice in the area in which the trial is being carried out, knowledge of the trial design and any technical aspects. Allowing two hours for the meeting enabled adequate discussion of PRECIS-2 domains and exchange of information. It was apparent no single individual had adequate knowledge to use PRECIS-2 to score the COPERS trial. Thus the meeting to discuss scores cleared up some areas of misinterpretation of domains, improved consensus, and participants believed they were able to produce more accurate scores. The feedback was very positive both verbally and in the completed questionnaires, APT participants appreciated the time to consider trial design decision using the application of the PRECIS-2 tool and believed it assisted in considering the people who would be using their trial following publication.

Using the online tool the COPERS trial team wanted to use the PRECIS-2 tool but participants were unsure if they had to complete all fields, including the abstract, changes to the web version were made to make the PRECIS-2 tool easier to use. As soon as the elaboration paper and examples are published these documents will also be made available on the PRECIS-2 website. This extra information will assist
trialists who are uncertain about scoring domains. Most of the participants strongly agreed that the PRECIS-2 domains assisted trialists to consider the intended audience and applicability of trial results, in particularly the *Eligibility* domain and all believed using the tool at the design stage would have been helpful. Suggested changes were made for future APT meetings based on this pilot.

**Case study 1**

**Aim**

To investigate if PRECIS-2 is a valuable tool for designing the HepFree primary care trial to ensure relevant to practice.

**Objectives**

1. To assess the views of the HepFree trial team towards PRECIS-2.
2. To test viability of the online PRECIS-2 tool and training package use by the trial team.
3. To determine if scoring of PRECIS-2 domains assists the HepFree trial team in considering the intended audience and creation of a trial relevant to practice.
4. To determine if scoring of PRECIS-2 domains prior to a meeting to discuss trial design and at the end of a meeting assist the HepFree trial team at the Pragmatic Clinical Trials Unit.
5. To assess if we can improve the methodology for using PRECIS-2 by trial teams

**Methods**

We applied PRECIS-2 to one trial from the Pragmatic Clinical Trial Unit trial, HepFree - a trial currently in the recruitment phase investigating ways to improve testing and treatment of hepatitis [http://blizzard.qmul.ac.uk/research-generation/738-hepfree.html](http://blizzard.qmul.ac.uk/research-generation/738-hepfree.html) The trial protocol that was used in the meeting is unpublished (Version 3.0, dated 01JUL13), . An audio recording was made of the meeting to assist in reviewing the discussion after the meeting. In addition, notes were taken by GF and KL.
Participants

All four participants from HepFree invited were able to accept the invitation to attend the APT meeting. This included the Primary investigator, the trial manager, a representative from the Trial Steering Committee and the trial statistician. In addition, GF – the Principal investigator for the APT study and KL took part in the APT meeting. GF to act as convenor for the meeting and KL giving further information on PRECIS-2 as required supporting the group’s PRECIS-2 domain discussion. The assistance by KL involved giving information based on the PRECIS-2 elaboration paper [153](in Press) to facilitate optimal discussion of the PRECIS-2 domains. However, KL did not attempt to influence the team in their scoring decisions by giving a personal opinion.

Hepfree Trial

The trial is a cluster randomised controlled trial to assess the impact of identifying, screening and treating immigrants who have viral hepatitis. GP practices are randomised to receive the intervention. As it was a cluster randomised trial, “participants” were GP practices, not patients.

PRECIS-2 scores

Two rounds of independent scoring were undertaken. All HepFree trialists scored using the online version of PRECIS-2 for the first round www.PRECIS-2.org. Participants were instructed to read the information on PRECIS-2 contained in this website, no other information on PRECIS-2 was provided, and then use the tool to score the HepFree trial.

Following the online scoring, we held a two-hour face to face meeting with the HepFree participants. We used the simple Agenda outlined in Box 9.1 to ensure timing and meeting objectives were met. First there was a short introduction on PRECIS-2 by GF, then the scores from the first round were shared and a discussion was held with the aim of improving consensus in PRECIS-2 scores. GF initiated discussion on the HepFree trial starting with the Eligibility domain and finishing with the Primary analysis domain. Participants independently recorded PRECIS-2 score for the second time for each
As previously, in the COPERS meeting for APT, at the end of the rating meeting the raters completed a short questionnaire (Appendix Fig 9.1) providing feedback of their experiences with the PRECIS-2 tool. Participants were also asked to provide any verbal feedback on the PRECIS-2 scoring process or questionnaire. GF contacted all Hepfree participants by e-mail to thank them for attending the APT meeting and to send them the final PRECIS-2 wheel for their trial, along with PRECIS-2 scores for each participant as well as the median scores for each domain. If any of the trialists had questions regarding the PRECIS-2 tool or APT they were encouraged to get in touch and they were also told they were “free to use the PRECIS wheel in any way that suits”.

Results
Only one rater (the statistician) used the PRECIS-2 online scoring tool for this trial, other scores were given to GF via e-mail. Reasons for not completing online were mainly due to shortage of time.

Following the face to face discussion there was with complete agreement on four domains (Eligibility, Flexibility delivery, Follow-up, Primary analysis). There was a one point range on two domains (Setting, Organisation), a two point range on two domains (Flexibility Adherence and Primary outcome), and scores covering a four point range on one domain (Recruitment) (Figure 48, Table 54).
Figure 48 PRECIS-2 wheel created by website tool showing range of scores (5 raters) for HepFree

Table 54 Scores for all 5 raters including median scores with maximum and minimum

<table>
<thead>
<tr>
<th></th>
<th>Eligibility</th>
<th>Recruitment</th>
<th>Setting</th>
<th>Organization</th>
<th>Flexibility Delivery</th>
<th>Flexibility Adherence</th>
<th>Follow-Up</th>
<th>Primary Outcome</th>
<th>Primary Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Min</td>
<td>5</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Max</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Trial manager</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Statistician</td>
<td>5</td>
<td>1</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>TSC representative</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>GF</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>PI</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

Figure 49 Scores for all 5 raters including median scores with maximum and minimum

Case study 1 – Hepfree Questionnaire Results
The results from the questionnaire are shown in the tables below (Table 55). The questionnaire was completed by all four participants: the PI, TSC representative, statistician and trial manager from the HepFree trial team. Two out of four agreed or strongly agreed that the online tool and training material was adequate, helpful and easy to use. There was strong agreement that the meeting to discuss the
scores led to more meaningful scores being produced and agreement that the tool would have been beneficial to use in the design phase. Two out of four participants agreed that the tool highlighted areas of design which are important for the trial to achieve its goals and the other two participants were unsure. There was a range of responses across the domain specific questions (7.1 - 7.9) with mostly positive answers.

The free text feedback was only completed by the PI who thought the meeting to discuss HepFree using PRECIS-2 had been useful.

**Table 55 Completed Questionnaire for HepFree – APT study pilot**  
*SA = Strongly Agree, A = Agree, N = Neither agree nor disagree, D = disagree, SD = strongly disagree.*

<table>
<thead>
<tr>
<th>No.</th>
<th>Question</th>
<th>No. of Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The online training material contained sufficient information about PRECIS-2</td>
<td>1 1 1 1</td>
</tr>
<tr>
<td>2</td>
<td>The online information on how to use the tool is clear and concise</td>
<td>2 1 1</td>
</tr>
<tr>
<td>3</td>
<td>The online tool is easy to use</td>
<td>2 1 1</td>
</tr>
<tr>
<td>4</td>
<td>PRECIS-2 would have been useful to use in the design stage of the trial</td>
<td>1 3</td>
</tr>
<tr>
<td>5</td>
<td>Meeting to discuss independent scores lead to more accurate scores being produced than were produced independently</td>
<td>2 2</td>
</tr>
<tr>
<td>6</td>
<td>PRECIS-2 highlights areas of trial design which are important for your trial to achieve its goals, be that informing clinical decision making or increasing knowledge of how an intervention works.</td>
<td>2 2</td>
</tr>
<tr>
<td>7</td>
<td>How important are each of the PRECIS-2 domains in ensuring the results from your trial are relevant to their intended audience</td>
<td></td>
</tr>
<tr>
<td>7.1</td>
<td>Eligibility - Who is selected to participate in the trial?</td>
<td>2 2</td>
</tr>
<tr>
<td>7.2</td>
<td>Recruitment - How are participants recruited into the trial?</td>
<td>1 3</td>
</tr>
<tr>
<td>7.3</td>
<td>Setting - Where is the trial being done?</td>
<td>1 3</td>
</tr>
<tr>
<td>7.4</td>
<td>Organisation – What expertise and resources are needed to deliver the intervention?</td>
<td>1 3</td>
</tr>
<tr>
<td>7.5</td>
<td>Flexibility - How should the intervention be delivered?</td>
<td>2 2</td>
</tr>
<tr>
<td>7.6</td>
<td>Flexibility - What measures are in place to make sure participants adhere to the intervention?</td>
<td>3 1</td>
</tr>
<tr>
<td>7.7</td>
<td>Follow-up - How closely are participants followed-up?</td>
<td>3 1</td>
</tr>
<tr>
<td>7.8</td>
<td>Primary outcome - How relevant is it to participants?</td>
<td>1 2 1</td>
</tr>
<tr>
<td>7.9</td>
<td>Primary analysis - To what extent are all data included?</td>
<td>1 3</td>
</tr>
</tbody>
</table>
Case Study 1 – HepFree Face to face meeting Results
PRECIS-2 domain Hepfree trial discussion was based on unpublished protocol (http://blizard.qmul.ac.uk/research-generation/738-hepfree.html)

The main part of the APT meeting with the Hep-Free trial team was based on discussion of the nine PRECIS-2 domains.

Eligibility

Protocol Summary
Study summary/synopsis p5 “Female and male patients who have been identified as first generation immigrants born in a country of high risk or second generation immigrants...high risk (as outlined by WHO classification of Chronic Viral Hepatitis prevalence >2%), >18 years of age.” The trial will exclude people lacking capacity.

All raters agreed a score of “5” for “Eligibility” as very inclusive.

Recruitment

Protocol Summary
Primary Objectives p11

- “To assess the impact of either opportunistic (control) or targeted (intervention) screening for chronic hepatitis in primary care patients within ‘at risk’ ethnic minority communities
- To determine whether the provision of an enhanced patient information invitation letters increases attendance for testing when compared to standard information invitation letter.”

In addition, “First and second generation immigrants from known ‘at risk’ communities will be identified utilising GP practice list definitions of ethnicity.”

The group were uncertain how to score the Recruitment domain— using opportunistic screening is very pragmatic and electronic records to target eligible men and women to be screened is also pragmatic but the trial team was using a recruitment methodology that was not usual care. They used “augmented [enhanced] patient invitation letters” including both English and appropriate translation, informing patients about the study in addition to a patient information sheet and consent form. This
could be relatively easily be incorporated into usual care though extra information would increase postage costs, as would translation, though this trial would create letters in appropriate language for future use so KL and two others marked down the trial from very pragmatic to “4”. It was also unclear if it is usual practice to phone potential patients in usual care three times to try and get them to a screening clinic, so the score for the Recruitment domain could also be reduced because of that.

Rating of the “Recruitment” domain ranged from “1” to “5”. Raters median score was “4”.

Setting

Protocol summary

Study Summary/synopsis p4

“There will be 72 centres to be utilised over old Primary care trusts (including Bradford as well as South and East London), known to have a high density of immigrant populations from ‘at risk’ countries (WHO classification of Hepatitis B Virus prevalence >2%)”

The team were nearly in agreement (3/4) in scoring “Setting” as “4”, as the trial setting is very similar to places would want to focus screening on in usual care if trial is successful but not identical to all places would focus screening on. (The statistician gave a score of 5.)

Organisation

Protocol summary

Introduction, Background p1 “The current model of care involves specialist centres with highly trained staff administering therapy at some distance from the patient’s home.”

The third objective p11 Objectives (which involves change in organisation from usual care) is to:

- “To determine whether community based therapy is superior to conventional delivery of treatment (based on referral to local treatment centres) as measured by engagement with management.”

The team spent some time discussing the Organisation domain.
The intervention which is being tested is also changing the organisation for treatment of Hepatitis - it is making it a lot easier for patients with a positive test for hepatitis to be treated at local centre (instead of traveling long distances). The plan is to use experienced hepatitis nurses as the principal point of contact but in line with usual standard of care also use a named local specialist consultant and GP input in addition to nurse management. The team discussed issues with introducing this into usual care. There were two options. Firstly, by increasing the remit of practice nurses, which would involve training and having clinics at GP practices for hepatitis treatment injections but for many hard-pressed practice nurses this may not be possible. Alternatively if the same highly trained staff that previously treated people in specialist centres are involved but now travel to local centres to give injections and follow-up, this could continue post trial with minimal problems. This would make it easier to match trial organisation in the trial with usual care post trial. Thus in this trial changing the Organisation of the intervention, if the trial is successful, would be highly successful and attractive. Either way, additional training or information sessions in GP Practices would be necessary to implement this community-based treatment.

Most of the team scored “3” as the “Organisation” is different to current practice and usual care though as may be easier for nearly everyone (from patients to health professionals (except perhaps the specialised hepatitis nurse who has to travel) to implement into usual care, the trial team may have considered scoring the trial as more pragmatic. The – PI scored 4.

Flexibility (delivery)

Protocol summary
Standard operating procedures and monitoring and no indication other than usual care.

All trial staff scored “4” for “Flexibility (delivery)” domain as decided health care professionals would be more strict in following protocol than in usual care.
**Flexibility (adherence)**

**Protocol summary**

A secondary objective (p11) is:

- To assess treatment compliance between patient’ groups receiving treatment within the community care setting against standard hospital care.

P18 “Patients who undergo therapy after testing positive for viral hepatitis will be assessed for compliance by measurement of medication returns – compliance will be defined as taking more than 80% of the prescribed medication.”

The patient with hepatitis has to have three visits for treatment in 12 months as well as taking medication received three times in the 12 months. So there is monitoring but there appears to be no more than usual encouragement.

*The team all scored “4” for the “Flexibility (adherence) domain after minimal discussion as similar to usual care (PI scored 5)*

**Follow-up**

**Protocol summary**

p20 Schedule of assessment: three visits for treatment and qualitative demographic information - four questionnaires to be completed by participants through phone call (method of choice), interview or postal survey if patient prefers: an adapted version of the Barts Explanatory Model Interview checklists, patient health questionnaire (PHQ-9), the generalized anxiety disorder 7-item (GAD-7) scale as well as the HLQ/HEIQ health literacy questionnaire. Team unclear how long this would take, perhaps 30 minutes.

*Three visits is usual care but the completion of questionnaires is more than “just usual care” so the trial team reduced pragmatism score to “4” for the “Follow-up” domain.*
Primary outcome

Protocol summary

P12 “7.1 Primary Endpoint efficacy analysis

- The proportion of patients eligible to be screened (determined by a review of the number of immigrants registered at the GP practice at the initiation of the study)
- The proportion of potential patients that attend for testing
- The proportion of potential patients that engage in therapy (defined as attending on at least 3 different occasions) in the different treatment arms.”

The trialists thought that these outcomes were very relevant to the participants – but appeared to be thinking of the outcomes from the perspective of the clinicians and trialists. Patient relevant outcomes was mentioned and as that there was more than one primary outcome, there was also some discussion about composite outcomes but the trialists believed they match the three trial objectives.

Median score by trialists was 5 for “Primary outcome”.

Primary analysis

Protocol summary

P22 Statistical analysis “We will use the intention to treat principle when identifying which clusters and arms to analyse individuals in. Thus if patients switch between practices before their test results are available they will be analysed in the practice they were in when randomization took place in relation to comparisons A and B but in the practice to which they moved to in relation to comparison C (because at this stage the trial to test the effect of community care on engagement will not have started).”

There was unity in the trial team and minimal discussion in deciding the trial scored “5” for “Primary analysis”.

Overall HepFree Discussion

The Organisation domain received most discussion. The domain appeared to assist the team to consider issues that they had not specifically considered before, comparing to usual care. There is an argument for scoring the domain as “5” and “very pragmatic” if it was going to be implemented as
usual care post trial as there would be minimal information or training sessions for GPs and practice nurses. However, ensuring there were sufficient adequately trained hepatitis nurses to cover clinics for local care appeared to be a little uncertain. There were also issues of training up practice nurses to take on hepatitis clinics if they could not get sufficient number of trained hepatitis nurses to travel.

Primary outcome – there was unity in the trial team but opinion differed to KL about an appropriate pragmatic primary outcome. Their focus was from the perspective of the clinicians and trialists not the patients diagnosed with hepatitis who would be receiving treatment. KL thought a more pragmatic outcome would be “free of disease” which would be indicated by engaging in therapy. The other outcomes could be subsumed by “free of disease” as in “engaged in therapy” but from the perspective of someone who is testing positive for hepatitis, “free of disease” could be considered to be a more patient relevant outcome. This goal “free of disease” would also be a good ultimate goal for medical practitioners as well as people that tested positive for hepatitis. Unlike the trial team, KL scored this domain “2” as did not fulfil the full pragmatic score for “Primary outcome – to what extent is the trial's primary outcome directly relevant to participants?” – participants being patients and health practitioners in this cluster randomised trial. However, considering the primary outcome from the perspective of commissioners of care, the people who decide whether to implement the intervention on the basis of its results, the issue of screening rates are particularly relevant to policymakers who are keen to reduce infection and increase treatment rates to avoid the significant long term health costs of someone with untreated hepatitis.

For clarification, advice in the elaboration paper suggests that trialists should always consider the Primary outcome from the perspective of the patient: “...Post-trial, an outcome selected using a pragmatic approach would also be relevant to commissioners of care, the people who decide whether to implement the intervention on the basis of its results.” This, in particular, concurs with Patient-Centered Outcomes Research Institute (PCORI) that focuses on research that is patient centred and
includes outcomes most important to patients. If the trial is intended to be of direct use to, say, policymakers and the patient-important outcome is not sufficient to satisfy their decision-making needs, then the trial team may need to add an additional primary outcome. This might need the team to score one or more additional Primary outcomes individually (i.e. the PRECIS-2 wheel has more than 9-spokes but more than one of them is Primary outcome).

Generally the PI scored the trial protocol as being more pragmatic compared to the rest of the team, this concurs with the study by Russell [56].

**Study limitations**
Acting as an observer but attending the trial group meeting to assist the team in using PRECIS-2 had limitations. KL was keen to give input on the PRECIS-2 domains and elaborate on the information, taken from the unpublished PRECIS-2 paper however, KL did not want to influence the team in scoring their trial but simply present them with the information to assist in scoring and thus enabling individuals to make informed decisions about scoring. KL did believe sometimes that if she did not agree with the trial team then perhaps inadequate information had been given but the divergence in opinions highlighted the issues of inter-rater variability in scoring domains. The trial team were the people who were most in agreement in using PRECIS-2 to score their trial and perhaps that was the most important thing.

While all of the trial team gave feedback on the online PRECIS-2 information, only one member of the trial team had used the tool to rate the trial (the statistician who was also involved in another APT case study in which PRECIS-2 had been used). It is therefore unclear how much training individuals in the group had on the PRECIS-2 domains and using the tool. One of the trial group, however, was on the Steering group for APT (SE) but it appeared that the PI and trial manager were not very familiar with the tool so the meeting also was a chance for the team to learn more about using PRECIS-2.
While there was a trial protocol (July 2013 version 3) and there was still some possibility for modifying the trial design, most of the work had already been done prior to the APT meeting. The meeting however did emphasise areas where there was still uncertainty, for instance in the Organisation domain and the delivery of the intervention in the experimental arm – in particularly changing the hepatitis clinics from being hospital based to GP local clinics.

**Conclusion**
The APT meeting with the Hep-free trialists was considered to be a success; there was full attendance of representatives from the trial team who were very engaged in discussing the trial and using PRECIS-2 to score the domains. There was mixed response to the PRECIS-2 tool as only one participant had used it. All participants thought that PRECIS-2 would have been useful to use in the design stage of the trial.
Case study 2

Aim
To investigate if PRECIS-2 is a valuable tool for designing the STOP primary care trial to ensure relevant to practice.

Objectives
1. To assess the views of the STOP trial team towards PRECIS-2.
2. To test viability of the online PRECIS-2 tool and training package use by the trial team.
3. To determine if scoring of PRECIS-2 domains assists the STOP trial team in considering the intended audience and creation of a trial relevant to practice.
4. To determine if scoring of PRECIS-2 domains prior to a meeting to discuss trial design and at the end of a meeting assists the STOP trial team at the Pragmatic Clinical Trials Unit.
5. To assess if we can improve the methodology for using PRECIS-2 by trial teams.

Methods
We applied PRECIS-2 to a trial from the Pragmatic Clinical Trial Unit, STOP - a trial currently in the design phase optimising pharmacist-based treatment for smoking cessation involving randomisation at the pharmacist level and the patient level. The plan was that knowledge gained in the trial could be used to support general diversification of pharmacists’ role, an essential part of the new primary health care environment. An audio recording was made of the meeting to assist in reviewing the discussion after the meeting. In addition, notes were taken by GF and KL.

Participants
Three STOP trialists were able to accept the invitation to attend the APT meeting. Attendees were Primary Investigator, SE a grant applicant and chief investigator for the APT study and the trial statistician. As this trial is at such an early phase in development there is no trial manager or trial steering committee representative.

In addition, GF – the Principal investigator for the APT study and KL took part in the APT meeting. GF to act as convenor for the meeting and KL giving further information on PRECIS-2 as required.
STOP trial
The trial is a 2 by 2 factorial cluster randomised controlled trial to test an intervention to train pharmacists to deliver smoking cessation advice to walk-in customers at pharmacists. This is in line with a move within the NHS for pharmacists to help patients as pharmacies are much more accessible to walk-in treatments. Sixty pharmacies will be randomised to deliver smoking cessation treatments to 1200 smokers wanting to give up smoking. There will be three interventions: in the first group each smoker will be given personalised dosing of nicotine replacement therapy based on their individual requirements detected from their personal metabolic profile. (This is due to individual genetic variations.) So, 25mg patch + active/placebo lozenge according to nicotine profile. In the second group, smokers will be given standard doses of nicotine replacement therapy which comprises 25mg nicotine patch + placebo lozenge. In the third group smokers will all be given high doses of nicotine replacement therapy, 25mg nicotine patch + 15*2mg nicotine lozenges daily. This may mean that some smokers experience side effects but may prevent under-dosing of nicotine replacement therapy for some smokers who otherwise would not get as much nicotine as when they are smoking.

In our consideration of the trial using PRECIS-2 with the trialists, as it was a cluster randomised trials “participants” were pharmacies, however, as patients in each pharmacy were also randomised to one of three treatments patients as participants were also considered separately using PRECIS-2.

In the APT study, we discussed the pilot study of the intervention in 12 pharmacies in Newham, each recruiting about 8 patients in total over up to two months to test the feasibility and acceptability of the intervention.

PRECIS-2 scores
Two rounds of independent scoring were undertaken. All three STOP trialists scored using the online version of PRECIS-2 for the first round - www.PRECIS-2.org. Participants were instructed to read the information on PRECIS-2 contained in this website and then use the tool to score the STOP trial.
Following the online scoring, we held a two-hour face to face meeting with the STOP participants. We used the simple Agenda outlined in Box 9.1 to ensure timing and meeting objectives were met. First there was a short introduction on PRECIS-2 by GF, then the scores from the first round were shared and a discussion was held with the aim of improving consensus in PRECIS-2 scores. GF initiated discussion on the STOP trial starting with the Eligibility domain and finishing with the Primary analysis domain. Participants independently recorded PRECIS-2 score for the second time for each domain as the discussion progressed. The final scores presented in the Results are the median from these second round scores.

**Questionnaire and feedback**
As in the COPERS meeting for APT, at the end of the meeting the raters completed a short questionnaire (Appendix Chapter 9, Figure 9.1) providing feedback of their experiences with the PRECIS-2 tool. Participants were also asked to provide any verbal feedback on the PRECIS-2 scoring process or questionnaire. GF contacted all STOP participants by e-mail to thank them for attending the APT meeting and to send them the final PRECIS-2 wheel for their trial, along with PRECIS-2 scores for each participant as well as the median scores for each domain. If any of the trialists had questions regarding the PRECIS-2 tool or APT they were encouraged to get in touch and they were also told they were “free to use the PRECIS-2 wheel (Figure 9.7, Figure 9.8) in any way that suits”.

**Results**
In this trial two PRECIS-2 wheels were created, one considering the pharmacists as participants and the other considering the individuals who wanted to stop smoking as participants. The current approach to the trial is very pragmatic across most aspects at the pharmacy cluster level.

Following the face to face discussion there was complete agreement on four domains for the pharmacist PRECIS-2 wheel (Eligibility, Flexibility adherence, Follow-up, Primary analysis). There was a one point range on three domains (Recruitment, Setting, Flexibility delivery), a two point range on one
domain (Primary outcome), and scores covering a three point range on one domain (Organisation) (Table 56, Figure 50).

Considering the patient PRECIS-2 wheel (Table 57, Figure 51), patient level scores were not applicable for all domains, Flexibility delivery was not completed as that refers to the pharmacists delivering smoking cessation treatment, the Primary Outcome and Primary analysis domains were the same as for the pharmacist PRECIS-2 wheel. For the patient PRECIS-2 wheel, there was complete agreement on Eligibility, Recruitment, and Flexibility adherence. There was a one point difference in Setting and Organisation. The Follow-up domain had a two point difference.

Figure 50 Pharmacy/Pharmacist Level PRECIS-2 Wheel for STOP trial
Table 56 Pharmacy/Pharmacist Level PRECIS-2 Scores for STOP trial

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Median</strong></td>
<td>5</td>
<td>4.25</td>
<td>4.5</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>3.5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td><strong>Min</strong></td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td><strong>Max</strong></td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td><strong>Statistician</strong></td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td><strong>GF</strong></td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td><strong>Co-applicant</strong></td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td><strong>PI</strong></td>
<td>5</td>
<td>4.5</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

Figure 9.8 Patient Level PRECIS-2 Wheel for STOP trial

Figure 51 Patient Level PRECIS-2 Wheel for STOP trial

Table 57 Patient Level Scores for PRECIS-2 STOP trial

Note: blanks mean PRECIS-2 domains not applicable

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Median</strong></td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4.5</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td><strong>Min</strong></td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Max</strong></td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td><strong>Statistician</strong></td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td><strong>GF</strong></td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td><strong>Co-applicant</strong></td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td><strong>PI</strong></td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td></td>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>
Questionnaire Results
The results from the questionnaire are shown in the table below (Table 58). The questionnaire was completed by all three of the APT meeting participants – the primary investigator, a co-applicant on the grant (SE), and also a statistician who may work on the trial. None of the trialists gave any general comments on taking part in the APT meeting to use PRECIS-2. The statistician neither disagreed or agreed that the online PRECIS-2 tool and information was adequate, helpful and easy to use, the two co-applicants on the grant agreed (and strongly agreed) that it was. As the trial was being used in the design stage it was interesting to note one agreed helpful, another unsure and the statistician did not agree was useful to use at this stage. There was uncertainty that the meeting to discuss the scores leads to greater consensus about scoring and agreement that the tool would have been beneficial to use in the design phase but only one of the participants had scored prior to the meeting. All of the three participants agreed or strongly agreed that the tool highlighted areas of design which are important for the trial to achieve its goals. There was a range of responses across the domain specific questions (7.1 - 7.9) with mostly positive answers. The domain with strongest agreement on its relevance was Eligibility with all three participants stating strongly agreed important in ensuring the results from the STOP trial are relevant to their intended audience.
### Table 58 Completed Questionnaire for STOP – APT study pilot

*SA = Strongly Agree, A = Agree, N = Neither agree nor disagree, D = disagree, SD = strongly disagree.*

<table>
<thead>
<tr>
<th>No.</th>
<th>Question</th>
<th>No. of Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The online training material contained sufficient information about PRECIS-2</td>
<td>SA: 1, A: 1, N: 1</td>
</tr>
<tr>
<td>2</td>
<td>The online information on how to use the tool is clear and concise</td>
<td>SA: 1, A: 1, N: 1</td>
</tr>
<tr>
<td>3</td>
<td>The online tool is easy to use</td>
<td>SA: 2, A: 1</td>
</tr>
<tr>
<td>4</td>
<td>PRECIS-2 would have been useful to use in the design stage of the trial</td>
<td>SA: 1, A: 1, N: 1</td>
</tr>
<tr>
<td>5</td>
<td>Meeting to discuss independent scores lead to more accurate scores being produced than were produced independently</td>
<td>SA: 1, A: 2</td>
</tr>
<tr>
<td>6</td>
<td>PRECIS-2 highlights areas of trial design which are important for your trial to achieve its goals, be that informing clinical decision making or increasing knowledge of how an intervention works.</td>
<td>SA: 1, A: 2</td>
</tr>
<tr>
<td>7</td>
<td>How important are each of the PRECIS-2 domains in ensuring the results from your trial are relevant to their intended audience</td>
<td></td>
</tr>
<tr>
<td>7.1</td>
<td>Eligibility - Who is selected to participate in the trial?</td>
<td>SA: 3</td>
</tr>
<tr>
<td>7.2</td>
<td>Recruitment - How are participants recruited into the trial?</td>
<td>SA: 1, A: 2</td>
</tr>
<tr>
<td>7.3</td>
<td>Setting - Where is the trial being done?</td>
<td>SA: 2, A: 1</td>
</tr>
<tr>
<td>7.4</td>
<td>Organisation – What expertise and resources are needed to deliver the intervention?</td>
<td>SA: 2, A: 1</td>
</tr>
<tr>
<td>7.5</td>
<td>Flexibility - How should the intervention be delivered?</td>
<td>SA: 2, A: 1</td>
</tr>
<tr>
<td>7.6</td>
<td>Flexibility - What measures are in place to make sure participants adhere to the intervention?</td>
<td>SA: 2, A: 1</td>
</tr>
<tr>
<td>7.7</td>
<td>Follow-up - How closely are participants followed-up?</td>
<td>SA: 2, A: 1</td>
</tr>
<tr>
<td>7.8</td>
<td>Primary outcome - How relevant is it to participants?</td>
<td>SA: 2, A: 1</td>
</tr>
<tr>
<td>7.9</td>
<td>Primary analysis - To what extent are all data included?</td>
<td>SA: 3</td>
</tr>
</tbody>
</table>
Case study 2 – STOP Results Face to face meeting

The main part of the APT meeting with the STOP trial team was based on discussion of the nine PRECIS-2 domains which will be described here. There was discussion from the perspective of the pharmacist as this group was randomised to three different interventions and from the patient as this group was also randomised so two separate PRECIS-2 wheels were created. There was overlap in two domains for the PRECIS-2 scores: Primary outcome and Primary analysis. The information on domains is taken from the outline in the NIHR grant application 2010 for the main trial and from the team discussion. Although there was an outline of the trial in the NIHR grant, the pilot had not got underway, and as there was no formal protocol the team did have flexibility in how they designed the pilot and full trial. KL gave information when requested but did not attempt to influence the team in their scoring decisions.

Eligibility

Protocol Summary - pharmacists

All pharmacists responsible for patient care (including ancillary staff if giving care) were considered for inclusion in the trials.

There was agreement by all of the trial team that to give a PRECIS-2 score of 5.

Protocol Summary – patients

P30 of 45 “All smokers seeking smoking cessation advice in the pharmacy where the pharmacist would have offered drug therapy in the NHS Stop Smoking Service, including smokers under the age of 18 (with parental consent), pregnant women and those with psychiatric illness. Smokers who experienced a cardiovascular event within the last three months will be excluded.”

The trial team had tried to be as inclusive as possible in particular considering groups that were often excluded but may need help with smoking cessation, women who were pregnant, people with mental
health problems and teenage smokers. So the trial team scored as very pragmatic as included all smokers wanting to quit so PRECIS-2 Score 5.

Recruitment

Protocol Summary - pharmacists – Population p29 “Pharmacists will be invited by letter (including stamped addressed envelope for response) with one follow-up phone call. The first 12 pharmacists responding will be selected. Pharmacist paid to be part of intervention group.- there will be incentive to take part in trial with £2-300 offered to pharmacies randomised to intervention pharmacies.”

Discussion of this domain was about incentives and the importance of getting sufficient pharmacists to take part in the trial and compensating them for their contribution. The actual amount, perhaps £100 would depend on the public health commissioner. There was discussion about what compensation there was for pharmacists already in the real world, involved in programmes like this and what was usual care incentivisation. The trial team decided this was like usual care and rather or very pragmatic and to score the Recruitment domain as “4” or “5”.

Protocol Summary - patients

Recruitment for patients was drop-in and ad hoc, was opportunistic and any smoker going to pharmacist interested in quitting was eligible. There would be no additional advertising, so recruitment was as close as possible to usual care. If patients chose not to enter the trial they will be still be offered smoking cessation clinic information as per usual care.

The trial team were unanimous in choosing to score the trial as 5.

Setting

Protocol Summary – pharmacists (p29 out of 60 Setting and inclusion criteria). The setting for the trial was community pharmacies in East London. These pharmacies would be recruited from all pharmacies offering NHS Smoking cessation services in Tower Hamlets and Hackney. The results will be used in deprived areas with high smoking rates and low quit rates. Eg The Chief Pharmacist at NHS Glasgow and Clyde has offered to host implementation studies.
The trial team discussed the implementation studies but as the trial was in a similar setting but included only one area of the UK that had a high incidence of smoking, this was considered less pragmatic by two of the team.  

*Two of the team gave a very pragmatic score of “5” for Setting, the other two scored this domain “4”.*

**Protocol Summary – patients**

This was the twelve pharmacies that had self selected themselves following recruitment so it was not ALL pharmacies that a smoker may visit to get help stopping smoking. While these pharmacies were believed to reflect all the pharmacies in the area, there was some discussion that they be better at health promotion and have more enthusiastic team than other pharmacies in the same deprived area the trial where the trial was being undertaken. *This domain was scored as “5” by three out of four of the trial team and “4” by one participant.*

**Organisation**

**Protocol Summary – pharmacists**

This domain included discussion on the extra resources and personnel needed for the trial compared to usual care. There was discussion on training for pharmacists recruited to the trial – the key part of the intervention. This would comprise teaching pharmacists *“effective pharmacist consultation techniques”*. Currently pharmacists receive two hours of training, called a *“light touch”* programme. The trial intervention that the PI was discussing developing, would give pharmacists better training. To do this would probably need different trainers, the trial would aim to employ highly effective ones. The trialists want to ensure fidelity of training courses so there will be monitoring of the training sessions with feedback if necessary. This would thus not be “usual”. However, if the training programme is effective this will replace the existing two hour training programme for pharmacists. There will be no top-up training sessions but there was some discussion about the possibility of a trial “tweet” using Twitter to pharmacists involved in the trial intervention. It was unclear if the trainers for pharmacists would be resources that would be available post trial if this intervention was successful.
Three out of four of the trial team scored the “Organisation” domain as “4” with the statistician scoring this domain as “1”.

Protocol Summary – patients

The trial team appeared to agree that the organisation from the patient’s perspective was no different, indeed two of the team did not score this domain as they did not think it was relevant (Organisation is described in the PRECIS-2 toolkit as resources, provider expertise and organisation of care delivery).

There were two scores of the Organisation domain for the patient as “4” and “5”.

Flexibility delivery

Protocol Summary – pharmacists

There was discussion about the education of pharmacists (see Organisation) The pharmacists delivering the intervention will follow the PACE program http://cmcd.sph.umich.edu/physician-asthma-care-education-pace.html. PACE promotes patient-centred consultation skills, using key consultation strategies and key messages and provides a promising starting point for a smoking cessation intervention. There will be no check up of fidelity of delivery of training so like usual care.

The team all scored the Flexibility delivery domain as “5” as appears to be fully flexible.
**Flexibility adherence**

**Protocol Summary – pharmacists**

This will entail the pharmacist giving advice on behavioural change – in the trial there may be monitoring of adherence by pharmacists to training on how to give advice but there will be no measures to improve adherence. There might be a “mystery shopper” to check pharmacist service but fully flexible as it is completely up to the pharmacist in how he or she delivers smoking cessation information.

*The trial team all scored the Flexibility adherence domain as “5”.*

**Protocol Summary – patients**

There were no measures in place to ensure patients adhered to the smoking cessation programme, other than usual encouragement.

*The trial team all gave this domain for Flexibility adherence “5”.*

**Follow-up**

**Protocol Summary – pharmacists**

The trial follow up for pharmacies will be to collect primary outcome data – the recruitment rate to smoking cessation treatment, pharmacists and pharmacies - through collecting NHS returns, which is a very pragmatic collection of outcome data. Reading the grant application, additional follow up is proposed to gain qualititative feedback from pharmacists about the educational intervention to train them in PACE methods to reduce smoking and to gain information on administration time and costs for these smoking cessation clinics. A lot of this would probably be routine data collection so difficult to determine if extra data collection compared to usual care. (It was not discussed in the meeting.)

*The triaists all scored the Follow-up domain as “5”.*

**Protocol Summary – patients**

Patient follow up is the one-month quit date which is the same as the NHS smoking service but in addition, all participants will be followed up for six months. Smoking status will be determined by self reporting as well as exhaled carbon monoxide at one month and plasma cotinine at six months. It is
unlikely the trial participants will come to the pharmacy for follow-up so they will need to be telephoned or texted and visited at home by the research nurse. This was discussed at the APT meeting but the grant application also states that patients will complete a questionnaire to measure level of satisfaction with pharmacist consultation using a Likert scale. The team were a little uncertain in scoring this domain compared to usual care.

*Trial team scores were: “2”, “3” and “4” for Follow-up.*

**Primary outcome**

**Protocol Summary – pharmacists and patients**

The recruitment rate (from NHS returns) into smoking cessation programmes is the proposed primary outcome measuring “Process of care” for pharmacists in this pilot. There was discussion on how relevant this outcome is to pharmacists but there appeared to have been a great deal of thought by the trialists in selecting this as the primary outcome. Pharmacists are paid by the number of smokers who quit after enrolling in programmes – this is a revenue stream from the NHS. This trial primary outcome for recruitment to quit smoking is a surrogate outcome for the “quit rate” – “outcome of care”. The trialists discussed issues with recruitment to quit, for instance that this could be acceptable if there is the same denominator for quit rate - as in all smokers who get in touch with pharmacists and want to quit. Problems may occur, however, if pharmacists assess smokers wanting to quit and decide that it is unlikely the smoker will follow through and be able to quit so they do not recruit to the stop smoking intervention.

The only primary outcome that is relevant to the smokers enrolling in a clinic to stop smoking is if they succeed and quit smoking so recruitment is only the first step in quitting. From the patients’s perspective recruitment rate is also a surrogate outcome and only quit rates is the outcome of interest (though this is harder to determine accurately – hence carbon monoxide test at 1 month and blood test at 6 months).
After discussion, the trialists considered having a joint primary outcome: recruitment into smoking cessation services and quit rates. There are in fact the outcomes being evaluated for the final phase of the project – the full RCT.

The trial team scored this Primary outcome domain for patients as “2”, “3”, “4” and “4”.

**Primary analysis**

**Protocol Summary – pharmacists and patients**

The analysis for this trial is “Intention To Treat” but at the time of the APT meeting there is no planned imputation, if people are lost to follow-up then they will be lost to the trial. Anybody lost to follow up will be treated as if they had not been given the intervention or control treatment. There was a lot of discussion at the APT meeting that this could cause bias. The standard in smoking cessation trials is not to impute data but there was discussion whether or not this is correct. There was discussion that multiple imputation would be considered.

The trial team was unanimous in scoring “5” for ITT for the NHS returns for the primary outcome.

**Case study 2 Overall discussion**

The key difference in the STOP trial compared to the previous trial Hepfree was using PRECIS-2 to discuss a trial that was still under development; the STOP protocol had not been finalised and trial design was still being considered. This is exactly the stage that PRECIS-2 was intended to be used to assist trialists consider aspects of design that will have an impact on the end users of the trial results. The fourth question in the questionnaire about using PRECIS-2 was not really relevant “PRECIS-2 would have been useful to use in the design stage of the trial” but this received mixed responses, one participant disagreeing but without giving a reason. It would be interesting to see if the design of PRECIS-2 wheels considering randomisation at the pharmacy level, change over time from the initial use at the APT meeting, at the time of protocol publication and when the results of the trial are published.
This trial was designed to be as pragmatic as possible with considerations to the applicability of trial results. The trialists were expert in smoking cessation and familiar with usual care for smoking cessation and had carefully considered an intervention that could make a difference and be implemented post trial into usual care. PRECIS-2 helped make these decisions clearer to others at a glance and encouraged discussion on the issues of recruitment to “stop smoking” clinics. Ideally, the goal of a smoking cessation trial for both pharmacists and smokers is hard data for “quit rates” but the primary outcome for the STOP trial, in the pilot phase, was recruitment rates to smoking cessation pharmacy programmes, using NHS Revenue data that is automatically collected. In this pilot phase for the RCT, the investigators wanted to test out the intervention to promote smoking cessation that would be used by pharmacists and this was different to their overall goals for the main trial to reduce smoking rates. Thus the pilot phase we were focussing on had slightly different outcomes from the main trial, composite outcomes and more detailed breakdown of recruitment and acceptability of the intervention for pharmacies and patients which included quit rates for patients as surrogate outcomes. Unlike the other APT trials, COPER and HepFree, this was a cluster randomised trial that had different domain design for Patients/customers and Pharmacists for Eligibility, Recruitment, Setting, Organisation, Flexibility adherence and Follow-up. PRECIS-2 was used for both patients/customers and pharmacists as both were randomised and so design decisions for different domains affected these participants differently. The other domains had the same PRECIS-2 scores for the trial design at the pharmacy level.

Case study 2 Strengths and limitations
Overall, discussing the trial at the time of protocol development, and in particular discussing the pilot study of the intervention using PRECIS-2 was positive. A meeting at this stage did appear to encourage trialists to consider different aspects of the trial design compared to usual care. This could avoid trialists falling into the trap of not fully considering the likely applicability of their trial results given their design choices. When the APT meeting was held for this STOP trial it was still possible to change the proposed design as the protocol was not yet finalised. Discussion at this stage avoids the problems
Schwartz and Lellouch mention of trialists tending to design trials that are more explanatory than they should be given the designers’ intention [13]

One limitation of this case study was that not all of the trial team had been able to prepare for the face to face meeting to discuss the trial using PRECIS-2. There was only one PRECIS-2 score prior to the face to face meeting so a shift in consensus could not be measured at the end of the meeting, after all domains had been discussed. The PI had not had time to look at the online PRECIS-2 toolkit prior to the meeting and had only briefly scanned the information to familiarise himself with PRECIS-2. This meant that there was a steep learning curve to discuss the PRECIS-2 domains and understand the concept of PRECIS-2 and how it could be useful in trial design. As a practising GP and trialist he had other priorities thus emphasising that information has to be attractively presented and concise to assist optimal use of the PRECIS-2 tool when it is accessed. The co-applicant SE (APT steering group) was very familiar with PRECIS-2 and the statistician had used PRECIS-2 on HepFree (Case 1). It may be that one meeting to use the PRECIS-2 tool may not be sufficient to assist in trial design and protocol development. Nevertheless the APT meeting did appear to be useful to all participants.

**Case study 2 Conclusion**

PRECIS-2 was tested out in the design phase, which is when it is intended to be used. The tool enabled the STOP trial team to discuss different aspects of their trial design compared to usual care and make their design decisions transparent. Using PRECIS-2 supports a closer link between intention (a trial with highly applicable results) and trial design choices. If that is achieved, and the intervention shows benefit, then the results ought to be widely applicable because the trial was explicitly designed for that to be the case, a situation aided by PRECIS-2.

**APT Discussion**

In these case studies, there was a positive response to PRECIS-2 particularly by the primary investigators and the trial steering committee representatives. There was genuine interest in PRECIS-2 with everyone participating in discussion and completing the questionnaires. Without
representatives from the trial team being present in meeting, there would not have been sufficient information and knowledge in the room to use PRECIS-2 to score the different domains. This information is important but not new; raters in Validity testing (Chapter 6) found that a lack of reported information and inadequate clinical expertise prevented confident scoring of PRECIS-2 domains.

The online PRECIS-2 was less well received by the statisticians involved in APT meetings, neither agreeing or disagreeing that “The online training material contained sufficient information about PRECIS-2”, “The online information on how to use the tool is clear and concise” and “The online tool is easy to use”. Without further clarification it is hard to determine how the PRECIS-2 could be presented better for this group. One could speculate that it is harder for statisticians in making a judgement call to use PRECIS-2 to score trials; there is guidance but no specific advice in adding or subtracting to create a score. As domains have varying number of restrictions to score a trial as more explanatory, this was not possible. PRECIS-2 use is a concept which takes time and experience for all users but this group may have particularly benefited from the concrete guidance in the unpublished PRECIS-2 elaboration paper. This was used by KL at APT case study meetings to assist the trial team but was not available for individual trialists to consult. It is worth mentioned though that two of the steering group are statisticians (SE and GF) and the first author of the PRECIS paper published in 2009 was a statistician.

It was obvious that for many at the meeting using PRECIS-2 is very new; it may be that it takes time to get used to the concept of PRECIS-2, comparing an intervention to usual care, and considering the end users of the trial when designing a trial. Judging by our experience of the pilot and the two case studies, the online resources and a presentation on PRECIS-2 prior to trial meetings to discuss the trial design were both important. Having the elaboration paper and examples as additional tools should assist trial teams in using the tool in future meetings. These resources may decrease the range in scores for some of the domain assessments by raters and increase trialists’ confidence in their PRECIS-2 scores.
APT study strengths and limitations
The PCTU, where the APT study took place, is very progressive in designing pragmatic trials that are undertaken in real world settings. As PCTU have been doing this for some time, there is a supportive environment for creating pragmatic trials and improving the methodology for doing this which provided a generally receptive though critical audience. The case studies, however, involved only a small sample of purposively selected trialists discussing PCTU trials so consulting a broader range of people may give different results. A further limitation may have been that trialists did not have the full information on how to use the tool at their fingertips. They were directed to the tool kit on the website www.precis-2.org but at the time of the APT studies the full elaboration paper [153] was not available for trialists to consult. A presentation on PRECIS-2 was however given at the beginning of each meeting which was important for team members who had not had time to prepare before the meeting and also offered a reminder to those who had started learning about the tool for designing trials that are fit for purpose.

Conclusions
Overall, utilising PRECIS-2 with trial teams was a success. Attempting to encourage all APT trial participants to use PRECIS-2 was generally successful but the face to face meeting was enlightening for all participants as each trialist was familiar with different aspects of the trial. This knowledge sharing enabled APT participants to come closer to agreement in scoring domains using PRECIS-2. In addition, most participants acknowledged PRECIS-2 was useful in the design stage and assists trialists in considering the applicability of trial results. The APT studies, however only involved a small number of participants from each trial team using the PRECIS-2 tool in designing trials. Interestingly statisticians as a group were less certain about the usefulness of PRECIS-2 although only two statisticians were involved in the pilot and case studies.
SUMMARY

PRECIS-2 was used with trial teams to discuss the nine domains of Eligibility, Recruitment, Setting, Organisation, Flexibility (delivery), Flexibility (adherence), Follow-up, Primary outcome and Primary analysis. Trialists were active participants in the Applying PRECIS-2 to Primary Care Trials (APT) meetings and were mostly positive about using the tool at the design stage to consider applicability and the end users of the results: patients, smokers, pharmacists, GPs, policy makers.

CONCLUSION

PRECIS-2 could be an important part of the preparation of trial designs in improving transparency in decision making.
Chapter 10: Discussion and Conclusion

Overall discussion

The methodology for each phase of improving the original PRECIS tool and developing PRECIS-2 has been discussed throughout this thesis so this chapter will focus on discussing key issues and future work. The research methods of the 2 round Delphi process, 3 rounds of Brainstorming, User testing and validity and reliability testing were well tried and tested techniques to create the PRECIS-2 tool to assist trialists in designing trial that are fit for purpose.

Participants in all of these research processes included a wide range of researchers, methodologists, trialists, healthcare professionals - some early career but most very experienced. This was important as it assisted in producing a trial design tool that is user friendly and itself “fit for purpose” enabling individual team members designing trials to increase the transparency in decision making with a purpose designed website www.PRECIS.org. We believe we recruited sufficient participants for adequate sample sizes for this work and through collaborating with over 80 international researchers, many leaders in their fields, we have a very enthusiastic group of potential users of PRECIS-2. This provides a sound basis for disseminating the PRECIS-2 tool and increasing use of the tool in designing randomised controlled trials.

The original premise for developing PRECIS was based on the assertion by Schwartz and Lellouch that “most therapeutic trials are inadequately formulated” [18, 28]. Thus a tool was developed to assist trialists discuss whether or not a trial was finding a solution to testing an intervention in more ideal circumstances (explanatory) as opposed to more real world conditions (pragmatic). Through adding a numeric scale of 1-5 and changing the domains we believe we have created a tool that is an even stronger candidate in helping trialists “formulate” or communicate their intentions for designing a trial and enhancing the quality of research proposals.
Although PRECIS-2 emphasises comparisons with usual care, it is important to acknowledge that it is not possible to skip over the important exploratory steps in developing an intervention which necessitate a more explanatory approach. For instance in drug development it is essential that the mechanism of the intervention is researched in pre-clinical trials to demonstrate that the drug can have an effect, and is safe. These steps are necessary before Phase 3 drug trials with a larger sample size when a more pragmatic design could be considered. This is also highlighted in the IDEAL collaboration framework for developing surgical interventions [http://www.ideal-collaboration.net/]. PRECIS-2 however may be particularly useful for Phase 3 and Phase 4 trials when the drug has been approved for use to highlight areas that trialists need to consider to help successful implementation of the intervention being tested.

PRECIS-2 could be used at several stages in testing an intervention to indicate the design decisions; it could be helpful in a grant proposal for a trial pre-intervention testing, as well as during the trial intervention phase and post-intervention phase clearly indicating any design changes. While we have developed the tool to design trials, there is interest in using it to design non-randomised trials and to assess trials in systematic reviews. Applying PRECIS-2 will still require trialists to judge where on the scale of very pragmatic to very explanatory a trial domain will be but we believe the elaboration paper and website will greatly assist this process [153].

The original PRECIS tool has been cited in different types of publication, apart from methodological publication mentioned earlier it has been cited in large number of discussion papers. I would anticipate that PRECIS-2 will also be cited in discussion papers considering trial design challenges. For instance in a recent paper by Holbrook [209] on Comparative effectiveness research PRECIS was cited as the group created a pragmatic trial mulita-centre Uveitis Steroid Treatment (MUST) emphasising that the intervention was similar to usual care with regard to patients and patient centred outcomes and follow up. During the trial, the team used considerable resources to implement the protocol and monitor that
the trial patients were getting the intervention being tested as “usual care”. Monitoring then resulted in changing the protocol during the duration of the trial to make it closer to usual care as this was changing and there was awareness this varied greatly between sites. This is not uncommon as discussed by Wells who highlights organisational issues and context are important for implementation of the results [210]. For instance, testing this intervention in the US and ensuring patients could receive the treatment who were unable to pay meant that treatments were received from the pharmaceutical companies for free and there was reimbursement to patients and clinics to cover treatment. If the treatment was successful then it would be helpful to the reader to know that there had been discussion about these organisational issues so that this intervention could be implemented. It may be that the organisation of the trial by necessity has to be explanatory but being transparent about the trial design will assist everyone from policymaker, doctor to patient to consider this intervention. Use of PRECIS-2 in this trial description may have made the original design and changes visually clearer to all concerned with implementing the intervention post trial through the visual depiction of the trial using a PRECIS-2 wheel.

Dwan (2011) published how trial design decision often change over time from the published report when compared to the protocol and trial registry entry [211] while Getz (2015) suggested simplification of protocols may be beneficial to everyone involved in trials [212]. Through assessing the impact of protocol changes on clinical trials performances between 1999-2005, Getz et al observing a 10.5% increase in administrative work involved in protocol implementation while recruitment and retention of participants actually decreased [212]. As PRECIS-2 assists transparency in trial reporting by enabling trialists to explain changes in trial design, design changes could be monitored through using the image of the PRECIS-2 wheel and table rationale in future.
PRECIS-2 domains in context of the wider research world

**Trial efficiency**
While PRECIS-2 could be used by a trial team as part of their routine process in designing and undertaking trials, it could be part of a package of tools to create “a systematic approach to making trials more efficient” as suggested in Trial Forge (www.trialforge.org) [213]. Trial Forge is a group of trialists and methodologist whose aim is to reduce waste and inefficiency in trials through improving the flow of information from the research question to dissemination of results through developing a trial process similar to https://processmap.tghn.org/ and instigating appropriate research.

**Research and Implementation**
Dissemination or translation of research evidence into practice is the aim of RE-AIM (Table 59) (http://www.re-aim.hnfe.vt.edu/index.html) which was combined with PRECIS-2 in considering the applicability of results from a weight loss trial in obese patients [68].

**Table 59 RE- AIM framework [66]**

| Reach your intended target population |
| Efficacy or effectiveness |
| Adoption by target staff, settings, or institutions |
| Implementation consistency, costs and adaptations made during delivery |
| Maintenance of intervention effects in individuals and settings over time |

The application of the RE-AIM framework assists health professionals and researchers implement trial research. But of great importance is the production of relevant research in the first place, which is why PRECIS-2 is so important to designing trials that consider the needs of the user rather than the trialists. Through an E-book publication Pragmatic trials and PRECIS-2 are being linked to Dissemination and Implementation in Health (http://www.crispebooks.org/PragmaticTrialsEbook_Preview) highlighting their close connection and the potential importance of the tool in future research.
So PRECIS-2 is the tool to use to design clinical trials right at the beginning of the design period. If PRECIS-2 is used at the beginning of the process, RE-AIM is the tool to use once the trial results are known assuming the trial has shown that the intervention is effective. The RE-AIM tool guides the implementation of the intervention as it is disseminated and a larger number of centres, health care providers and patients are involved. The tools are complementary as has been shown by a publication in the USA by a research group that combined the two [68]. PRECIS-2 and RE-AIM could be used to form a continuous chain from design to implementation to ensure continuity and increase the success of both producing relevant trial results and successful implementation of evidence based research. Why else are trials undertaken if not to produce evidence that can then be implemented?

Another group of American and Canadian scientists working at the NIH, however, also considered the use of both PRECIS and RE-AIM purely from the perspective of dissemination and implementation (D&I) [214]. A framework was created to help researchers in D&I with a long term view of developing guidelines for D&I research. The PRECIS-2 tool was described as the tool most ready to be used in “Evaluation/Results Reporting” of trials to indicate the pragmatic/explanatory continuum of a trial and the readiness for D&I. The RE-AIM tool was also included under “Delivery”: Reach, Adoption and Implementation (see Figure 52 Framework for enhancing the value of research for D&I [214][214]. Framework for enhancing the value of research for dissemination and implementation (D&I) [214]. In this case the tools are complementary but both being used for D&I.
Considering the nine domains in PRECIS-2 I will now discuss conclusions from the work developing PRECIS-2 and additional ideas that may help trialists in future.

Using PRECIS-2, the first step in designing the trial is considering who is going to be included in the trial – the Eligibility criteria domain. This is the domain trialists are most familiar with and the key issue with deciding inclusion criteria is expert knowledge in the field, to be clear on inclusion and exclusion criteria. This can be simplified for a pragmatic design through considering everyone who would be offered the intervention in the real world. Then considering the Recruitment domain recruiting patients has to be planned. STEPS [146] uses a business model to manage the recruitment of the participants the trialists would like to include through various strategies including ensuring the trialists engage the support of the key stakeholders. In a pragmatic trial, trialists would want to recruit as many people as possible who would receive the treatment in usual care and STEPS can streamline this process in trial management. Further methods are being tested through the MRC START project by nesting recruitment and retention trials into treatment intervention trials (www.population-health.manchester.ac.uk/mrcstart/about/). Recruitment is an issue that is more likely to be poorly
reported in trials [215] and as we found in the validity and reliability study can make it harder to assess this domain. Trialists designing their own trial will not face this difficulty but by being explicit in their recruitment process description in publications through using CONSORT guidelines (www.consort-statement.org/) future development of trials will be supported.

For optimal consideration of the trial intervention setting, familiarity with usual care for the intervention is needed to enable true matching of usual care in the most pragmatic trial for the Setting domain. So for instance testing out the intervention in primary care if that is usual setting, not secondary care clinics and using primary care setting as diverse as they would be in usual care. Organisation is a completely new domain and this aspect of trial design should focus the trialists on considering what extra resources are needed to deliver an intervention, expertise is again needed to ensure full understanding of the implications of the way an intervention is tested and subsequently fitting this process into usual care. As mentioned in the introduction to this chapter I would anticipate that this domain will have far reaching consequences on testing and implementing the trial intervention if discussed early on in the design process. The sad consequences of not preparing for post-trial implementation of a successful implementation were emphasised in the Theta brainstorming meeting (Chapter 4, 2nd brainstorming meeting) when the ARDS trial was discussed [131].

Successful trials may well change the process of care and the Organisation domain can be used as an indicator for issues surrounding implementation. An example of a trial area that led to huge shifts in usual care which did not reflect current care structures at the time (and so would have scored low for pragmatism) was stroke thrombolysis Boysen 1995 [216], a trial in the Cochrane systematic review on thrombolysis for acute ischaemic stroke [217]. This was not tested in an environment anywhere near usual care at the beginning – but which has led to usual care changing (hyperacute stroke units, emergency neuroradiology, thrombolysis teams etc.). In this particular case the evidence for thrombolysis use was so great that its use increased. There are at least two way this can be dealt with
by PRECIS-2. Firstly, Organisation is scored as highly explanatory (it is far from usual care) but the trial team knows that this is because a change in organisation is part of what is being evaluated. The second approach would be to consider the organisational change to be part of the intervention (which it is) and score Organisation only of organisational changes that are not part of the intervention. The key issue with either approach is that the trial team recognises that a substantial change to organisation is part of what is being tested. The PRECIS-2 elaboration paper differentiates between a change in organisation that is part of the intervention and detailed in the protocol and that which is not: “In that case these resources should not be counted when this element of the trial is judged”. In other words, it is the very difference to usual care that is being tested. It is however, part of the intervention, it is over and above what is required in the intervention itself.

For the PRECIS-2 domain Flexibility delivery again expertise is invaluable. This is the domain that was hardest for the participants in the validity and reliability testing of PRECIS2- as they were not specialists in the field of interest for each of the trials included in the sample. In APT (Chapter 8) this domain also produced most discussion as the team clarified exactly what this entailed. If this occurs right at the beginning of the trial design process, then the team clarify from the outset the aim for the health professionals involved in delivering the intervention; if the fidelity of protocol delivery is rigidly enforced through feedback or if delivery will be similar to usual care. The next domain that was in the original PRECIS tool, Flexibility adherence, we believed it was important to include. Historically compliance has been acknowledged but frowned upon and “per protocol” analysis analysing only those who were compliant has been undertaken in many trials. However, as Dave Sackett stated compliant patients do better whether or not they are taking the intervention treatment or comparison (as long as there are not serious adverse effects) so Intention To Treat (ITT) analysis is vital [218]. If correct procedures for randomisation and allocation of treatment the number of compliers in the intervention and comparison groups should be equal but we believed it was important for trialists to
consider exactly if they were going to monitor and encourage compliance or do no more than occurs in usual care.

Follow up, as in the original PRECIS-2 was also included to continue encouraging trialists to consider the full spectrum of follow up. This may range from routinely collected information that may be minimal and obtained through datasets using objectively measured endpoints as in the originally proposed large simple trials [219] to intensive information gathering that requires patients to attend far more than usual clinic appointments and includes excessive gathering of information that is not directly relevant to the primary outcome of the research question [220]. The latter is currently being assessed as part of Trial Forge to give empirical evidence to determine the amount of time trialists spend gathering primary and secondary data (much of which is not used or deemed unnecessary). Follow up is dependent on the Primary Outcome. This next domain in PRECIS-2 is vital to producing relevant answers to research questions that are important to trial participants and policymakers funding research and implementation of research findings. Initiatives like the Core Outcome Measures in Effectiveness Trials (COMET) in the UK (http://www.comet-initiative.org) and recent recommendation to trialists to use these core outcome sets and publish these in trial registries [221] could make it easier to disseminate trial results and undertake systematic reviews. While in the USA, the Patient-Centered Outcomes Research Institute (PCORI) is encompassed within the National Patient-Centered Clinical Research Network (PCORnet) which provides a platform for patients and trialists to work together to answers research questions that are important to stakeholders from patient to healthcare professionals and thus policymakers. Thus, both of these UK and US are geared to producing more useful results through focusing on relevant outcomes.

The final domain in PRECIS-2 has already been touched on. And remained from the original tool for completeness and as it could indicate the continuum from pragmatic to explanatory analysis of trial results. I believe we have created a tool that is innovative and has great potential for high impact on
trial design, there is unlikely to be complete agreement on the terminology for pragmatic trial, management trial ([222], practical trial [19], Large simple trial [219] but as long as the trial team agree what is usual care in the real world in the context of the trial they are designing then the nine domains of PRECIS-2 should enable discussion and consensus. Like PRECIS, using PRECIS-2 is a judgement call, easiest to use with expert knowledge of the trial intervention; the tool can be used to highlight uncertainty and disagreement within the team and bring the team closer to consensus to produce more relevant research.

Finally the use of PRECIS-2 to consider the issues of internal and external validity of randomised controlled trials of varying pragmatism has commenced suggesting pragmatic trials do not sacrifice internal validity to create external validity, despite asking different questions when testing the same intervention. This work will be considered in “Future work”.

**Future work**

**Using PRECIS-2 at trial design stage**

**Pre-funding**
The ideal use of PRECIS-2 in practice would be pre-funding in research proposals for randomised controlled trials. If the tool was applied at this earliest stage in considering the design of trials at their inception then the ultimate aim of the PRECIS-2 tool would have been fulfilled. This was one of the suggestions proposed during the APT study at the PCTU by several of the participants in trial teams. Through using PRECIS-2 this would ensure that the design of trials was carefully considered right at the very beginning, with the reason for the trial (more explanatory or more pragmatic) being a central tenet for deciding the design choices of the nine different domains in PRECIS-2. Only by using PRECIS-2 at this early stage will the applicability of the trial results for the end user ensure that trials are not designed by accident as suggested by Schwartz and Lellouch [28].

373
To encourage trialists to use the tool for the common good, improving and fulfilling research aims for applicability, the NIHR in the UK could stipulate that PRECIS-2 should be included in grant proposals. As the NIHR already have on their website the elaboration paper for PRECIS-2 as one of their most useful papers this may be achievable. There has been some preliminary discussion with the NIHR regarding this. As the largest of the NIHR programmes, the HTA, funds “independent research about the effectiveness, costs and broader impact of healthcare treatments and tests for those who plan, provide or receive care in the NHS” this should be a goal to improve uptake of the PRECIS-2 tool.

In addition to work described earlier in Chapter 8 with the Pragmatic Clinical Trials Unit with trial teams, additional in London is continuing as part of APT (Applying PRECIS-2 to Primary Care Trials) with the second part of the study. This involves consultation with policymakers and others who are responsible for commissioning new research and primary care practitioners who would potentially be involved in applying new research. These stakeholders have been consulted on which aspects of trial design are important when assessing evidence from clinical trials to improve clinical practice, and secondly, if they believe the different PRECIS-2 domains will assist in producing evidence to improve clinical practice. Thematic analysis of themes arising from the interviews is encouraging and it is hoped that this will provide an informed basis for further discussion with the NIHR to encourage official use of the tool in research proposals.

Post-funding
In the USA, work started in December 2014 with a webinar at the Center for Medical Technology Policy (CTMP) in Washington with information about the PRECIS-2 tool being presented by KL. This project, led by the director Dr Tunis, involves evaluating a set of National Institute of Health (NIH) pragmatic trial protocols over the planning year preparing for the trial start date. Key questions include:

- Does the design change over the planning period? For instance does the trial become more pragmatic or more explanatory?
- Does PRECIS-2 influence this process?
Earlier work in the summer of 2014 was also undertaken at the NIH with colleagues of Professor Glasgow using the PRECIS-2 tool working with the NIH on a 'common fund' project called the Health Care Systems Collaboratory (HCSC) [223] to apply the PRECIS-2 criteria to help understand where and how this set of projects is more and less pragmatic [170].

Further work in the UK could mirror the US project with NIHR grant proposals for pragmatic trial applications that were accepted. Although there is currently no routine, funded planning year for trials in the UK model, using PRECIS-2 to initially assess the trial protocol and then working with the trial team to discuss their trial as the intervention rolls out could assist the trial team and help to ensure that the trial does indeed provide results relevant to decision makers. PRECIS-2 could then be used to assess the trial at various stages to visually give an indication of how a trial changes and rationale could help others designing trials. Finally, this work could help policymakers and healthcare professionals in facilitating uptake of the trial by highlighting applicability.

**PRECIS-2 website and pragmatic trial database**

As the website [www.PRECIS-2.org](http://www.PRECIS-2.org) is live we are aware of several teams that are using the PRECIS-2 tool to design trials around the world. Every registered user is contacted personally by KL offering advice and help if needed to use the design tool. We are assuming that the tool is being used to design randomised controlled trials but one non-randomised study was published using PRECIS so it is possible PRECIS-2 will also be used to design observational studies. We are also aware that PRECIS-2 is being used for a review of diagnostic studies being done by Eleanor Barry with Trish Greenhalgh. As of 20th April 2015 there are 43 registered users and the tool is being increasingly accessed which has exponentially increased since the publication of the PRECIS-2 elaboration paper [153].

The database of pragmatic trials still needs to be fully entered into the PRECIS-2 website. Currently this is in an EndNote database and cannot be publically accessed. It would be of value to users of the website if many of the trials indicated their pragmatic design through PRECIS-2 wheels so this could
comprise future work. While users of the online PRECIS-2 tool can “release” their trials once published and make them available to other users, this has only been done for a few trials so far during collaboration with trial teams at the PCTU on the APT project.

Another way of enhancing the uptake of PRECIS-2 could be through newsletter to the people who have signed up to use the PRECIS-2 software to design trials. Dissemination of information may be best served through personal communication using Twitter or through personal correspondence answering specific questions of users.

Use of PRECIS-2 in trial publications
There are research questions that remain following the work described in this thesis. One that is no longer on the “to do list” is getting the National Library of Medicine to classify pragmatic trials, in October 2014 they started using the MeSH term “pragmatic clinical trial”. While pragmatic randomised clinical trial would have been preferable, this is at least a start in assisting pragmatic trial designers but it is still dependent on researchers assisting NLM with appropriate use of the work pragmatic in the title, abstract, main text and keywords.

The CONSORT extension for pragmatic trials was developed to encourage use by all trialists publishing trials [32]. This guidance may need to be updated in light of the development of the PRECIS-2 tool. In addition, it may be helpful to encourage the use of the two tools (PRECIS-2 and CONSORT) in conjunction with each other. It may also be helpful to readers of trial publications if we could get journals to use PRECIS-2 checklists and production of a PRECIS-2 wheel in their trial description. This could be very helpful to healthcare providers and policymakers to determine applicability of trial results if all trialists used this pre-publication; it would also help systematic reviewers (see below). While PRECIS-2 was created to be used by the trial team, if this was not done to assist in trial design it could still enable transparency in the design that would assist users of the trial results. An extra paragraph in a journal article, with information on the PRECIS-2 score for each domain and discussion
on why trialists decided to create different aspects (domains) in the trial could be very helpful to health care decision makers trying to improve patient care.

**Using PRECIS-2 with systematic reviews**
The work on the Cochrane systematic review indicated that further work using the tool retrospectively may be of use to determine applicability of reviews. As in the idea for trial publications, it could be useful for Cochrane systematic reviews of interventions to have PRECIS-2 wheels for the included trials. Some of the information for this visual depiction of pragmatism may well need to be gleaned directly from the primary investigators as it is doubtful that all of the information required to score the PRECIS-2 domains will be in the published trial articles. This would, of course, be easier if the publications already included a PRECIS-2 wheel as suggested above. An additional idea is create wheels and then present them to the primary investigator of the trial team and/or first author and discuss if there is agreement with the PRECIS-2 description of their trial testing face validity of the PRECIS-2 tool. My personal preference, however, would be to use the tool as intended to assist trialists in designing trials.

**Internal and external validity**
In addition, the work to consider internal and external validity of explanatory and pragmatic trials with a matched intervention should be completed to further investigate results so far that pragmatic trials do not appear to sacrifice internal validity for increased external validity. As different questions are being asked in explanatory and pragmatic trial design some have suggested this may not be relevant but pursuing this work may encourage further interest in real world pragmatic trial design by sceptics [43, 44, 191]. Two attempts were made on the work during the thesis but a further study including a larger set of Cochrane reviews of different interventions (e.g. behavioural change intervention or surgical in contrast to the previous cardiovascular drug trial), with more raters, should give evidence one way or the other. As before the Cochrane Risk of Bias tool would be used to assess the internal validity of trials and PRECIS-2 to assess the external validity. If the result of that work confirmed what was found in this thesis, then this may encourage pragmatic trial sceptics to reconsider their stance.
Finally, it is possible that future publication will reveal how PRECIS-2 is being used, so monitoring the future use of the tool could assist further development and dissemination of the tool, as it did for the development of PRECIS-2. For instance it would be helpful if all future trial PRECIS-2 publications are included in the database and as useful references for users. Of interest, reviewing if PRECIS-2 is published in more general medical journals or more specialised medical journals may assist in ascertaining who is using the tool and targeting those who may benefit from the tool.

**Teaching PRECIS-2 in research courses**
Informally (through user testing and personal communication), we know PRECIS and now the new PRECIS-2 tool and website are being used to teach trial design. The PRECIS-2 tool has also been included in a pragmatic trial training course, the Pragmatic Trials e-book by the Colorado Research and Implementation Science Program (CRISP): [www.crispebooks.org/PragmaticTrial](http://www.crispebooks.org/PragmaticTrial). Following publication of the elaboration paper [153] this has the potential to be increasingly used in Massive Open Online Courses (MOOC). There is definitely scope to include further learning material in the form of a webinar on the website for PRECIS-2, in addition to the two minute u-tube podcast on the PRECIS-2 tool which was prepared for the BMJ publication [https://www.youtube.com/watch?v=Sj7cNCyvHVE](https://www.youtube.com/watch?v=Sj7cNCyvHVE).
PRECIS-2 may be useful not only in specialised trial design courses but also as a tool to teach healthcare professionals in introductory research courses. Learning about designing randomised clinical trials could have two positive outcomes:

1. greater understanding of the importance of applicability of research and knowledge translation into practice when reading research articles;
2. more useful generation of evidence if healthcare professionals become involved in trials or move on to become policymakers

**Strengths of the thesis methodology**
A large number and variety of international researchers, methodologists, trialists and healthcare professionals were involved in the creation of PRECIS-2 with no or little incentives. I tried to include representation of the different users of the tool at some point in the development process. Participants involved were from 19 different countries across the globe. Some of these participants even travelled at their own expense to contribute to a brainstorming meeting in Toronto from the USA for an entire day. Very few of the individuals who assisted in the project were personally known to me (those that I did know prior to the PhD were through working with The Cochrane Collaboration). A triage of methods (brainstorming, Delphi process, user testing) were used to encourage input, re-inforce conclusions and also help in developing my skills as an independent researcher. The fact that we managed to test the validity and reliability of the PRECIS-2 tool in limited time indicates the enthusiasm of the researchers involved, the success of the methods used and project management.

**Weaknesses of the thesis methodology**
There is a possibility that the PRECIS-2 tool could have turned out differently if different participants has been involved. Clearly it would have been good to have the input from those who were unable to participate although many that did respond to say lack of time rather than lack of interest. Also some of those who did participate did not like the tool and did not really understand its purpose focusing on other aspects of trial development and management. In addition, considering in particular the Delphi
process if further time had been possible then an additional round may have led to complete consensus instead of a range of responses that were then used as the basis for discussion in the Brainstorming meeting in Toronto.

**Limitations of the final PRECIS-2 tool**
Any tool is only as good as the user and the instructions for use. Every individual using the tool will have different experience in trial design, some may be complete novices while others almost do not need the PRECIS-2 tool as the information is already part of their skill set. Thus only through further testing can we be certain that the tool does indeed help trial teams deliver trials that are *fit for purpose* be that more explanatory or more pragmatic. This further testing may come through individual use described in publications citing the tool or through grants to test the tool further. Work being done now by the NIH Collaboratory and PCORI may help. These organisations are increasingly keen on pragmatic design, whereas in the past the FDA has pushed designs towards the explanatory end. Increasing the pragmatism of grant applications for trials is also being considered by the NIHR funded trials and PRECIS-2 may assist with this process through increasingly the transparency of trial design. Early discussions with NIHR have taken place. Finally, PRECIS-2 was created with the intention of developing a tool for designing trials, however users are free to use it as they choose. Further advice will need to be given if it is to be used, for instance as a tool for assessing the pragmatism of trials in systematic reviews, something the original PRECIS tool was used for.
Conclusion
I believe that PRECIS-2 can increase the transparency of decision making in trial design and increase the applicability of trial research through the creation of real world pragmatic trials. This tool has been designed to make researchers even more aware of the match between usual care and the trial intervention. Further work however needs to be initiated to actively encourage the inclusion of the PRECIS-2 tool in research proposals to assist funders in selecting trials for funding and helping researchers meet the aims of funders. This will only actively be undertaken by trialists if stipulated by research funders and commissioners of research.

The PRECIS-2 tool has potentially many uses: it is not only relevant in randomised trial design but potentially other trial designs; research dissemination through Cochrane systematic reviews and through using the tool to teach trial design. PRECIS-2 was developed as a tool to design trials, however it has incorporated some of the suggestions included in the CONSORT statement for reporting trials [127] and has now been included in a framework for D&I [214]. Personally I would be most keen to produce evidence that if trialists use PRECIS-2 that does indeed increase the applicability of trial results. Thus trials with higher PRECIS-2 domain scores do go on to be implemented and make a difference in improving healthcare and health services.

PRECIS-2 therefore has high impact potential to not only improve the production of relevant evidence and knowledge translation but, ultimately, patients through better, more applicable healthcare. Finally, through the development of PRECIS-2, I believe trialists now have a tool to accomplish the following: “Making trials matter: providing an empirical basis for the selection of pragmatic design choices in clinical trials”.

381
Publications and Presentations

Publications


Loudon K. Rapid Response - Endgames Statistical Question Explanatory trials versus Pragmatic trials. *BMJ* 2014;349:g6694


Presentations

Loudon K, Treweek S. PRECIS-2: podcast for BMJ publication [http://www.bmj.com/thebmj](http://www.bmj.com/thebmj) and utube: [https://www.youtube.com/watch?v=Sj7cNCyvHVE](https://www.youtube.com/watch?v=Sj7cNCyvHVE)


Loudon K. “PRECIS-2: a tool to improve the applicability of randomised controlled trials”: Oral Session PS2.3-O2: MRC Methodology Conference, Edinburgh, 18th to 19th November 2013

Loudon K. “PRECIS-2: a tool to improve the applicability of Randomised Controlled Trials”: Final year presentation: University of Dundee Student symposium, 6th June, 2014

Posters

Loudon K. “PRECIS-2: A tool to improve the applicability of randomised controlled trials” University of Dundee Student symposium 2013 (2nd year prize)
Workshops

Loudon K, Sullivan F. “Pragmatic Primary Care Trials – plausible poppycock or pertinent? Scottish School of Primary Care - NADEGS, Dundee, 24th and 25th January 2012.

Grants


Loudon K. Dowe Travel Grant to attend 2nd Sackett Symposium 27th/28th September 2013 in Niagara Falls.
References


2. Ambroise Paré’s account in the Oeuvres of 1575 of new methods of treating gunshot wounds and burns [www.jameslindlibrary.org]

3. A treatise of the scurvy. In three parts. Containing an inquiry into the nature, causes and cure, of that disease. Together with a critical and chronological view of what has been published on the subject [http://www.jameslindlibrary.org/illustrating/records/a-treatise-of-the-scurvy-in-three-parts-containing-an-inquiry/key_passages]

4. A brief history of clinical trials [http://genome.wellcome.ac.uk/doc_WTD020948.html]


7. Why the 1948 MRC trial of streptomycin used treatment allocation based on random numbers [www.jameslindlibrary.org]


30. Research: increasing value, reducing waste [http://www.thelancet.com/series/research]


64. Harden SM, Burke SM, Haile AM, Estabrooks PA: Generalizing the Findings From Group Dynamics-Based Physical Activity Research to Practice Settings: What Do We know? *Eval Health Prof* 2013.


74. Damschroder LJ: ASPIRE and PRECIS. Personal communication; 2014.


100. Fourie I: Content analysis as a means of exploring research opportunities from a conference programme. Health Info Libr J 2012, 29(3):197-213.


111. Treweek S, Barnett K, MacIennan G, Bonetti D, Eccles MP, Francis JJ, Jones C, Pitts NB, Ricketts IW, Weal M et al: E-mail invitations to general practitioners were as effective as postal invitations and were more efficient. *J Clin Epidemiol* 2012, 65(7):793-797.


125. Nichols J: **Pragmatic Trials - how pragmatic can we be?** In.; 2011: 30.


142. *Brainstorming Principles and Rules* [www.brainstorming.co.uk](http://www.brainstorming.co.uk)


trial evaluating the effectiveness of aripiprazole or haloperidol in combination with clozapine for treatment-resistant schizophrenia. Trials 2009, 10:31.


221. Clarke M, Williamson P: **Core outcome sets and trial registries.** *Trials* 2015, 16:216.
