HSR Health Services Research

DOI: 10.1111/j.1475-6773.2011.01347.x
ORIGINAL ARTICLE

Applying the PRECIS Criteria to Describe Three Effectiveness Trials of Weight Loss in Obese Patients with Comorbid Conditions

Russell E. Glasgow, Bridget Gaglio, Gary Bennett, Gerald J. Jerome, Hsin-Chieh Yeh, David B. Sarwer, Lawrence Appel, Graham Colditz, Thomas A. Wadden, and Barbara Wells

Objectives. To characterize Practice-Based Opportunities for Weight Reduction (POWER) trials along the pragmatic-explanatory continuum.

Settings. The POWER trials consist of three individual studies that target obesity treatment in primary care settings.

Design. Using the PRagmatic Explanatory Continuum Indicator Summary (PRECIS) criteria, nine reviewers independently scored each trial.

Methods. Average and median ratings, inter-rater reliability, and relationships to additional ratings of the extent to which study designs were explanatory (i.e., efficacy) versus pragmatic (i.e., practical) and related to external validity were determined.

Principal Findings. One trial was consistently rated as being significantly more pragmatic than the others ($R^2 = 0.43$, $p < .001$), although all three were in the moderate range on the PRECIS scales. Ratings varied across PRECIS dimensions, being most pragmatic on comparison condition and primary outcome. Raters, although undergoing training and using identical definitions, scored their own study as more pragmatic than the other studies/interventions.

Conclusions. These results highlight the need for more comprehensive reporting on PRECIS and related criteria for research translation. The PRECIS criteria provide a richer understanding of the POWER studies. It is not clear whether the original criteria are sufficient to provide a comprehensive profile.

Key Words. Methodology, CONSORT, RCT design, research design, pragmatic trials, dissemination, external validity

The gap between research and practice is well documented (McGlynn et al. 2003) and has been characterized variously as "lost in translation" by a National Institutes of Health (NIH) director (Zerhouni 2005) and a "quality

chasm" by an Institute of Medicine review group (Institute of Medicine, Committee on Quality Health Care in America 2003). There are multiple reasons for this gap, but many observers have focused on the discrepancy between the conditions found in real-world settings in which research results need to be applied, and the carefully controlled conditions under which interventions are often tested. Some have argued that there is an important need for more practical or pragmatic trials tested under more representative conditions (Tunis, Stryer, and Clancey 2003; Glasgow et al. 2006). Others have debated the types of research designs that are most relevant for effectiveness research (Simons-Morton et al. 1998; Rothwell 2005; Mercer et al. 2007). Regardless of one's position on these issues, almost all reviewers concur that greater detail and transparency in reporting of intervention characteristics, context, and assessment specifics are needed (Glasgow et al. 2004; Zwarenstein et al. 2008; The Dartmouth Institute For Health Policy & Clinical Practice, 2010).

The PRagmatic Explanatory Continuum Indicator Summary (PRECIS) is an important effort to increase such transparency in reporting (Thorpe et al. 2009). The CONSORT Work Group on Pragmatic Trials (Zwarenstein et al. 2008; Thorpe et al. 2009) developed the PRECIS criteria to help increase clarity about the extent to which a trial is applied and widely applicable (pragmatic) or is more basic and mechanism focused (explanatory). It consists of 10 dimensions: flexibility of the comparison condition; flexibility of the experimental intervention; practitioner expertise—in both experimental and comparison conditions; eligibility criteria; primary analysis; practitioner adherence; participant compliance; follow-up intensity; and outcomes. More standard use of PRECIS would provide guidance in intervention and trial

---

Address correspondence to Russell E. Glasgow, Ph.D., Deputy Director, Dissemination and Implementation Science, Division of Cancer Control and Population Sciences, National Cancer Institute, 6130 Executive Blvd., Room 6144, Rockville, MD 20852; e-mail: glasgowre@mail.nih.gov. Bridget Gaglio, Ph.D., is with the Mid-Atlantic Permanente Medical Group, Mid-Atlantic Permanente Research Institute, Rockville, MD. Gary Bennett, Ph.D., is with the Department of Psychology & Global Health, Duke Obesity Prevention Program, Duke University, Durham, NC, and Harvard School of Public Health, Boston, MA. Gerald J. Jerome, Ph.D., is with the Department of Kinesiology, Towson University, Towson, MD. Hsin-Chieh Yeh, Ph.D., is with the Departments of Medicine and Epidemiology, Johns Hopkins University, Baltimore, MD. Lawrence Appel, M.D., is with the Welch Center for Prevention, Epidemiology, and Clinical Research, Johns Hopkins University, Baltimore, MD. Thomas A. Wadden, Ph.D., is with the Department of Psychiatry, University of Pennsylvania, Philadelphia, PA. David B. Sarwer, Ph.D., is with the University of Pennsylvania, Philadelphia, PA. Graham Colditz, M.D., Dr.P.H., is with the Department of Surgery, Washington University School of Medicine, St. Louis, MI. Barbara Wells, Ph.D., is with the Clinical Applications and Prevention Branch, Division of Cardiovascular Sciences, National Heart, Lung and Blood Institute, Bethesda, MD.

design, as well as a broader, more comprehensive context to interpret the generalizability of results. It would also aid practitioners and policy makers in evaluating the relevance of a given study to their situations (Rothwell 2005; Glasgow 2008) and allow systematic reviews to have a much richer, more contextual database from which to draw general conclusions. We note that no study is completely pragmatic or completely explanatory, and it is not better or worse to be more pragmatic versus explanatory; this depends on the purpose of a given study.

The primary purpose of this study was to apply the PRECIS criteria to the set of National Heart Lung and Blood Institute (NHLBI)-funded POWER trials, which consist of three separate studies that do not share a common intervention protocol but share common goals and measures. Each tests distinct primary care-based interventions, each aimed at reducing weight in primary care patients who were obese and had at least one other cardiovascular disease (CVD) risk factor (Wells 2009). Secondary goals were to report on our experience with the rating system, including the rating procedures used and the reliability of our ratings, since to our knowledge there have not been published reports using the PRECIS criteria since their initial publication by the CONSORT Pragmatic workgroup (Thorpe et al. 2009). Our final goal was to provide a type of concept and criterion validity evaluation by comparing the 10 PRECIS criteria to similar ratings on the extent to which a study was pragmatic versus explanatory on eight other ratings related to efficacy versus practical, generalizable designs.

METHODS

POWER Interventions

The Practice-Based Opportunities for Weight Reduction (POWER) Trials Collaborative Research Group consists of three individual studies: "*Be Fit, Be Well*" (*Washington University/Harvard University*), "*POWER Hopkins*" (*Johns Hopkins University*), and "*POWER-UP*" (*University of Pennsylvania*). All three trials began recruitment in early 2008, and final data collection will occur in the spring of 2011. A total of approximately 1,100 participants were recruited from 15 participating clinics. The trials had common components to facilitate potential cross-site comparisons, but each protocol also incorporated distinct, trial-specific elements including different interventions and different secondary outcome measurements. The common components included most inclusion and exclusion criteria, a common primary outcome (change in weight

from baseline to 24 months), standardized physical measurements, several standard survey measures, and a common analysis plan for the primary analysis of principal outcomes. A single Resource Coordinating Unit provided administrative support to the Collaborative Research Group. Each study was approved by a local Institutional Review Board. NHLBI established a common Protocol Review Committee, which was responsible for approving all the trials, and a single Data and Safety Monitoring Board (DSMB), to monitor the trials (Yeh et al. 2010).

The "*Be Fit, Be Well*" study was conducted in Boston by investigators from Washington University, Harvard School of Public Health, and Kaiser Permanente Colorado (Greaney et al. 2009). This study recruited participants from three Boston area community health centers: participants were predominantly low-income (>85 percent) and from racial/ethnic minority groups (>90 percent). Participants were randomized to one of two arms: the Usual Care group received NHLBI's "Aim for Healthy Weight" brochure and continued their medical care as usual. Participants in the Intervention group set behavior change goals, self-monitored their progress, and received skill training using either a website or a combination of telephone-based interactive voice response system and print materials. They were also assigned a health coach who provided support through 18 counseling phone calls and were invited to attend 12 bi-monthly group sessions led by trained study staff. The health coach was a community health worker who was hired specifically for the research study and trained to deliver the intervention. Additionally, participants received tailored "prescriptions" for weight-related behavior changes signed by their provider, as well as tailored action plans to increase the use of community resources.

"*POWER Hopkins*" at Johns Hopkins University recruited participants from six primary care practices in the Baltimore area. Participants were randomized to one of the three arms. Those assigned to the control condition, self-directed, received written materials, as well as ongoing access to a static web page. Participants assigned to the call-center directed (CCD) intervention received a multi-channel, behavioral intervention with telephone, web, and email contacts, without in-person visits, implemented by trained coaches of Healthways, Inc., a disease management company. Those assigned to the in-person directed (IPD) intervention received a multi-channel behavioral intervention with in-person, group and individual sessions, along with telephone, web, and email contacts. The IPD interventions were delivered by coaches at the Hopkins clinical center. Both active interventions used established behavioral techniques to achieve weight loss (i.e., frequent contact, self-monitoring

of weight and physical activity, use of food records, accountability), and a web-based hub to facilitate communication among counselors, participants, and the primary care provider (PCP), as well as to promote behavior change in participants. At Hopkins, PCPs had a supportive, rather than a primary role, in delivering the interventions. At routinely scheduled visits, the PCP reviewed participants' weight loss reports and encouraged participants in the CCD and IPD interventions to remain active in the weight loss program.

"*POWER-UP*" at the University of Pennsylvania (UP) recruited participants from six primary care practices within the Penn Medicine system. Individuals were randomized to a control group or one of two interventions: Usual Care; Brief Lifestyle Counseling; or Enhanced Brief Lifestyle Counseling. Participants in the Usual Care condition received educational materials (i.e., NHLBI's "Aim for a Healthy Weight") that were distributed at quarterly visits with a PCP. Those in the Brief Lifestyle Counseling intervention received the same PCP visits, plus 26 brief counseling sessions with an auxiliary health care provider (e.g., a medical assistant). Participants in the Enhanced Brief Lifestyle Counseling condition received the same treatment as those in the Brief Lifestyle Counseling group, plus the choice of adjunctive meal replacement products or pharmacotherapy (i.e., orlistat or sibutramine, until sibutramine was removed from the market in October, 2010). PCPs and auxiliary health care providers were trained and certified in implementing the protocol. Extensive attention was devoted to educating PCPs about the use of pharmacotherapy. At the start of the intervention, physicians reviewed with participants the potential benefits and risks of both meal replacements and weight loss medications and asked participants to choose the approach they preferred (i.e., meal replacements versus medications). Participants did not begin their adjunctive treatment until the third visit with their lifestyle coach, to allow time to consider their choices.

The overall POWER program, as well as the individual studies, is described in more detail elsewhere (Derbas et al. 2009; Greaney et al. 2009; Jerome et al. 2009; Wells 2009).

Raters and Rating Procedures

We invited nine raters, including at least two from each participating POWER site, and two independent raters not associated with any of the sites, to utilize a four-step process to rate the three POWER projects on each of the 10 PRECIS dimensions. Raters were experienced researchers, six of whom were involved in one of the POWER studies. Most had M.D. or Ph.D. degrees and at least

moderate experience in clinical trials. We also purposefully included three raters who were not associated with any of the interventions or any part of the POWER study. As details for precisely how to train assessors and to rate the 10 PRECIS dimensions were not clear to us from the literature, we collaboratively developed the following procedures.

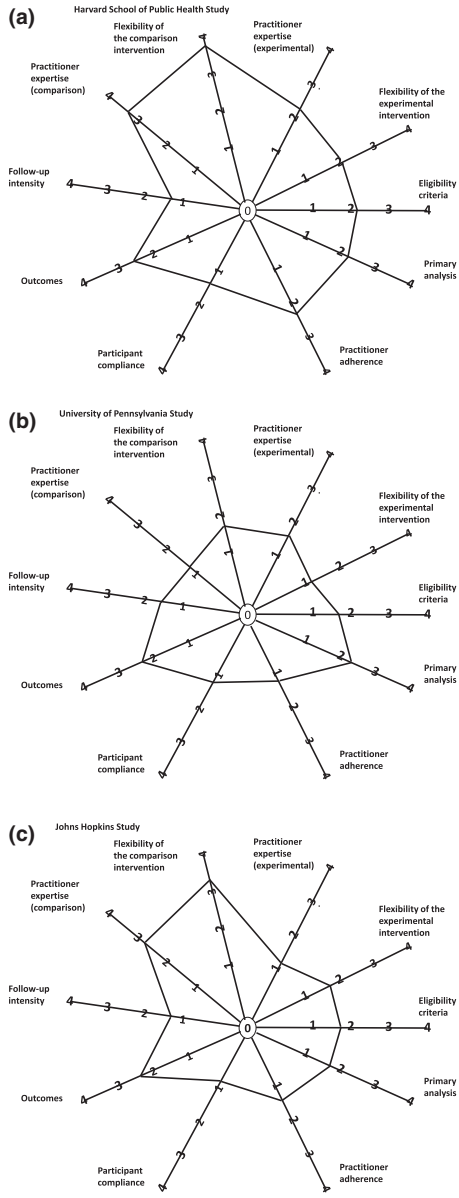
First, all reviewers read the article on the PRECIS criteria by Thorpe et al. (2009) and reviewed the slide presentation on PRECIS by Sackett (http://ww.support-collaboration.org/precis.pdf). Each rater then reviewed the centrally available protocol materials on each intervention and read a background description of each project that appeared in Obesity and Weight Management (Derbas et al. 2009; Greaney et al. 2009; Jerome et al. 2009; Wells 2009). Any questions the rater had were answered by a contact person at each site, who was not involved in the ratings. Raters then independently rated each of the three projects on the 10 dimensions using the rating form in Table S1. All three projects were scored using a "0–4" scale as described in the website above from Dr. Sackett and colleagues, with "0" being completely explanatory and "4" being completely pragmatic.

Analyses

Mean and median scores were calculated for each study on every dimension, and the mean was plotted on the PRECIS "spoke and wheel" diagrams in Figure 1, unless the resulting distributions were extremely skewed, in which case medians were used. This diagram is intended to convey, at a glance, how pragmatic versus explanatory a trial is by the distance of the marks on each dimension from the centroid: the further away from the center, the more pragmatic the trial is on that dimension. One-way ANOVAs were calculated to compare the three projects on overall composite scores, calculated by averaging the 10 component ratings (as well as the eight components for the additional items). Intraclass correlation coefficients (continuity model) were used to evaluate inter-rater reliability (McGraw and Wong 1996). The composite scores from each rater for each project were used in these analyses.

Parallel procedures were followed with eight additional ratings, which were developed by the study team based on their experience in translational research. These additional ratings were developed during this project to address concepts and issues related to relevance to practice and policy which were felt to be missing or not adequately covered by the PRECIS ratings. As shown in Table 3, these items addressed issues such as representativeness of participants and settings, inclusion of cost estimates, reporting on context and

Figure 1: PRagmatic Explanatory Continuum Indicator Summary (PRECIS) "Spoke and Wheel" Diagrams: (a) Harvard School of Public Health Study; (b) University of Pennsylvania Study; (c) Johns Hopkins University Study



level of engagement with the primary care practices. As these elements were added following and as a result of the initial training, raters did not have the same level of training on these items as on the original 10 PRECIS items.

RESULTS

Reliability

As can be seen in Tables 1 and 2, there was reasonable variability on almost all ratings. In general, there was moderate to high consistency among ratings of the individual PRECIS items; the ICC for individual items was 0.72 and for averages 0.96. The overall kappa inter-rater reliability on the composite PRECIS score was $r = 0.88$. Inspection of the actual ratings revealed that agreement varied across the individual items and was by far the lowest on PRECIS item #10 for Analysis. A random selection of 10 pairs of scores revealed that 76 percent of the scores of raters agreed within 1 point on the 5-point scale. Table S1 presents the raw data table, with raters de-identified (for readers who wish to inspect the raw data).

Additional Items

Reliability results for the additional eight items related to external validity were similar to those for the original PRECIS ratings. There was moderate variability across measures, ICCs were moderate: they were 0.45 for single items and 0.83 for averages. Inter-rater reliability on the composite eight-item scale was = 0.71. Correlation of the PRECIS composite with the external validity composite across raters were $r = 0.69, 0.57$, and 0.70 for the three sites.

Table 1: PRECIS Summary Scores by Site: Mean (Standard Deviation, Median)

Site	*No. of Raters*	*Overall PRECIS Score Mean (SD, Median)*	*PRECIS Score by Own Site Mean (SD, Median)*	*PRECIS Score Other Site Raters Mean (SD, Median)*
Harvard	9	2.36 (0.27, 2.3)	2.43 (0.15, 2.4)	2.32 (0.31, 2.2)
U of Penn	9	1.58 (0.50, 1.7)	1.95 (0.35, 2.0)	1.47 (0.5, 1.6)
Johns Hopkins U	9	1.82 (0.40, 1.8)	2.20 (0.42, 2.2)	1.71 (0.35, 1.8)
Overall		1.92 (0.5, 1.9)		

Table 2: PRECIS Scores by Site and Dimension

PRECIS Dimension	*Harvard School of Public Health Mean (SD, Median)*	*University of Pennsylvania Mean (SD, Median)*	*Johns Hopkins University Mean (SD, Median)*	*Rating Across Projects Mean (SD, Median)*
Participant eligibility	2.1 (0.8, 2)	1.6 (0.5, 2)	1.8 (0.4, 2)	1.8 (0.6, 2)
Experimental intervention				
Flexibility	2.0 (0.7, 2)	1.1 (1.1, 1)	1.7 (0.7, 2)	1.6 (0.9, 2)
Practitioner expertise	2.2 (0.8, 2)	1.6 (1.1, 1)	1.1 (0.6, 1)	1.6 (0.9, 1)
Comparison intervention				
Flexibility	3.8 (0.4, 4)	1.8 (0.8, 2)	3.2 (1.1, 4)	2.9 (1.2, 3)
Practitioner expertise	3.4 (0.7, 4)	1.3 (0.9, 2)	2.8 (1.3, 3)	2.5 (1.3, 3)
Follow-up intensity	1.4 (1.0, 2)	1.6 (1.2, 2)	1.3 (1.2, 1)	1.4 (1.1, 2)
Primary trial outcome	2.6 (0.5, 3)	2.3 (0.7, 2)	2.3 (1.0, 2)	2.4 (0.7, 2)
Participant compliance	1.4 (0.9, 1)	1.2 (0.8, 1)	0.9 (0.6, 1)	1.2 (0.8, 1)
Practitioner adherence to study protocol	2.3 (0.9, 2)	1.2 (0.7, 1)	1.4 (0.7, 2)	1.7 (0.9, 2)
Analysis	2.2 (1.4, 2)	2.1 (1.4, 2)	1.7 (1.4, 2)	2.0 (1.4, 2)
Overall composite				
Median rating	2.3	1.7	1.8	1.9
Mean rating	2.4	1.6	1.8	1.9
Range of ratings	2.0–2.7	0.8–2.2	1.3–2.5	0.8–2.7

Note. These overall ratings are average scores.

Comparisons of Scores across Programs and Raters

There were two clear findings from the PRECIS ratings that are summarized in Tables 1 and 2. First, the Harvard School of Public Health (Harvard) trial was rated as more pragmatic than the other two programs. This pattern was seen on 9 of the 10 individual ratings, and the ANOVA comparing the composite PRECIS score across sites was highly significant, despite the small number of observations ($F = 9.01$, $p = .001$, $R^2 = 0.43$; adjusted $R^2 = 0.38$). Tukey post hoc follow-up tests revealed that the Harvard trial was indeed rated as significantly more pragmatic than the other two sites, which did not differ significantly from each other. Differences between the Harvard and the other two trials were especially pronounced on ratings of practitioner adherence and on both measures concerning how pragmatic the comparison condition was (see Figure 1). As might be expected given the common assessment protocol, the smallest differences were observed on items related to follow-up intensity and primary outcome. On an absolute scale, the three projects were rated as moderately pragmatic, being viewed as most pragmatic on comparison

treatment flexibility and comparison treatment expertise. They were rated as most explanatory on ratings of follow-up intensity and patient compliance, as seen in Figure 1.

The second consistent finding was that despite the standard introduction and common set of procedures, raters gave more pragmatic ratings to their own trial than they did to the other trials (see Table 1). This was true across all trials. Because of this and to provide a type of sensitivity analysis, we repeated the analyses comparing the trials after removing the rater from the Harvard team who provided the most pragmatic ratings. The ANOVA with the reduced number of raters did not change conclusions and remained significant ($F = 7.51$, $p < .003$, $R^2 = 0.40$; adjusted $R^2 = 0.34$), as was the follow-up Tukey HSD analysis.

Added External Validity Ratings

The results on the added eight external validity ratings were parallel in many ways to those on the PRECIS ratings. In summary, ratings of the Harvard trial were higher than those of the other trials, although the ANOVA did not reach significance. The effect size was moderate ($R^2 = 0.14$) and

Table 3: Additional Practical Feasibility Item Ratings

PRECIS Dimension	*Harvard School of Public Health Mean (SD, Median)*	*University of Pennsylvania Mean (SD, Median)*	*Johns Hopkins University Mean (SD, Median)*	*Rating across Projects Mean (SD, Median)*
Participant representativeness	3.0 (0.6, 3)	2.2 (0.8, 2)	1.9 (0.7, 2)	2.3 (0.8, 2)
Setting representativeness	2.8 (1.5, 3)	1.3 (1.0, 1)	2 (1.0, 2)	2.1 (1.3, 2)
Context and setting	2.0 (1.4, 2)	1.3 (0.8, 1.5)	0.9 (1.1, 1)	1.4 (1.2, 1)
Community/setting engagement	2.0 (0.9, 2)	0.3 (0.8, 0)	0.7 (1.3, 0)	1.0 (1.2, 0)
Adaptation/change	1.8 (1.0, 2)	1.3 (1.4, 1)	1.1 (0.7, 1)	1.4 (1.0, 1)
Sustainability	1.8 (1.0, 2)	1.7 (1.0, 2)	0.9 (0.9, 1)	1.4 (1.0, 2)
Costs/feasibility of treatment	2.2 (1.3, 2)	2.2 (1.5, 2.5)	1.6 (0.8, 2)	2.0 (1.2, 2)
Comparison condition(s)	2.2 (1.6, 2)	2.2 (0.4, 2)	2.6 (1.1, 2)	2.3 (1.1, 2)
Overall composite				
Median rating	2.32	1.56	1.37	1.60
Mean rating	1.91	1.56	1.44	1.63
Range of ratings	0.2–3.1	0.4–2.5	0.5–2.5	0.2–3.1

the Harvard trial was rated as more pragmatic than both other trials on six of the eight items. As can be seen in Table 3, the items on which the Harvard trial was rated most different from the other trials were Engagement with the Community Setting and Reporting on Context. As with the PRECIS ratings, raters tended to rate their own trial as more pragmatic than they did other trials.

Overall, the items rated as most pragmatic were Participant Representativeness and Comparison Condition, and the study characteristic rated as most explanatory was Community Setting Engagement.

DISCUSSION

The primary purpose of this study was to apply the PRECIS criteria to three separate trials that have different interventions but share several common features as part of a NHLBI funded obesity reduction research program. Each trial tested distinct primary care–based interventions, all aimed at improving weight and health in primary care patients who were obese and had at least one other CVD risk factor (Wells 2009). The trials differed on the extent to which they were rated as being pragmatic versus explanatory, but overall they were seen as being midway between explanatory and pragmatic. When results from the three trials are available, it will be interesting to compare the projects on factors such as reach, participation rate (Abrams et al. 1996; Glasgow et al. 2004; Glasgow 2008), sustainability, implementation, and magnitude of weight loss. Long term, it will be of even greater interest to see whether one or more of the interventions is more likely to be widely adopted, as one of the purposes of the PRECIS ratings is to more objectively evaluate how close research studies are to real-world conditions.

As might be expected given the trial management and DSMB structure, the POWER studies tended to be more explanatory on the PRECIS factors related to evaluation and also more similar along those dimensions. The POWER protocols required standard measures and assessments intervals. The Harvard trial was rated as more pragmatic primarily because its intervention and the comparison conditions tended to be more flexible, and there was less close monitoring of practitioner adherence to protocol. As the developers of the PRECIS criteria emphasized, being pragmatic versus explanatory is neither good nor bad in an absolute sense, and no trial is completely pragmatic or completely explanatory (Thorpe et al.

2009) (see also Figure 1). Reporting PRECIS information should assist reviewers, potential adoptees, and policy makers to better judge the applicability of these interventions. We hope that this type of more transparent reporting of study and intervention details, and publication of PRECIS "spoke and wheel diagrams" (as shown in Figure 1), will help advance the field and suggest high-priority issues for further research. Although not addressed in this study, the PRECIS and related criteria may also be of use when planning intervention studies.

A possible reason for the differences between the trials in how pragmatic they were rated may have been related to the health systems in which they were conducted. The Harvard trial was conducted in community health centers, whereas the other two studies were conducted at primary care clinics associated with academic medical centers. Many aspects of health care are determined or strongly influenced by the setting or medical organization, including former and informal policies, and other cultural factors. It may be that being planned and implemented within these different cultures contributed to the differential ratings.

Secondary goals were to report on our experience with the PRECIS rating system, including the rating procedures used and the reliability of our ratings. To our knowledge there have not been any published reports using the PRECIS criteria since their initial publication by the CONSORT Pragmatic Workgroup or prior publications on the psychometric characteristics of the PRECIS criteria (Thorpe et al. 2009). Results related to this goal were more mixed. Some raters were skeptical of the ability to reliably rate the various criteria, and several conference calls were held among investigators to discuss and attempt to specify precisely what was meant by each dimension, and to develop and refine the eventual rating instruments and instructions. For example, raters from the Johns Hopkins and Penn trials felt that they used very flexible eligibility criteria for the POWER trial (i.e., "when in doubt, enroll the candidate"), compared with criteria used for most of their weight loss trials. Nonetheless, mean ratings for both sites left the trials appearing more explanatory than pragmatic. A larger sample of ratings and further exploration of rating behavior, obtainable through use of qualitative methods such as cognitive interviews, would further clarify the measurement properties of the PRECIS scales.

It was also challenging to find the most appropriate reliability index to evaluate inter-rater reliability given the 5-point rating scale, the small number of trials rated, and the relatively large number of raters. Overall, it appears that different raters were moderately consistent in their rating of individual

PRECIS items and the added external validity ratings. The reliability scores might have been higher if our training and scoring procedures were more rigorous and included scoring of and feedback on sample intervention study protocols prior to rating actual studies, or if the criteria had been more clear for some raters. Alternatively, raters could have questioned study implementation staff more thoroughly to complete the ratings, but few did so. Our goal, however, was to develop a relatively low-intensity training procedure that would be more generalizable and more likely to be adopted by most reviewers conducting systematic reviews. Despite moderate agreement at the individual items level, inter-rater reliability on the composite scores was high. Raters were able to distinguish one study as more pragmatic and to differentiate components of the three studies that were more explanatory versus pragmatic.

In retrospect, it was a good idea to have both independent raters and raters from each of the POWER study sites rate both their own and other sites. This revealed a type of potential bias: namely to rate one's own site as more pragmatic (or conversely other sites as more explanatory). NHLBI funded the POWER studies to adapt approaches shown to induce and sustain weight loss in efficacy studies and test their effectiveness in routine clinical practice. Raters from the study sites might have believed it was more desirable to be more pragmatic in this research program, and this may have influenced their ratings. An alternative explanation is that raters inherently had more information about their own site on the many decisions that are made daily in the conduct of a trial, which are not communicated in written protocols or publications.

Future research is indicated to investigate whether the more pragmatic ratings of one's own site are a generalizable finding or for some reason unique to this study. We also calculated a composite PRECIS score to summarize the individual ratings. It is not clear whether the PRECIS developers would concur with such a strategy. On one hand, it seems appropriate to have an index to summarize the overall extent to which a study is explanatory versus pragmatic. On the other hand, quite different study designs could possibly receive identical composite ratings, and one could argue that it is better to report all 10 individual ratings and focus attention on the pattern of results or the summary spoke and wheel diagrams, as shown in Figure 1.

Our final goal was to provide a type of concept and criterion validity evaluation by comparing the 10 PRECIS criteria and the composite score to similar ratings on the extent to which a study is pragmatic versus explanatory on eight other ratings related to efficacy versus practical, generalizable designs. The overall composite scores for the two summary scores were moderately related (r's = 0.57–0.70) but still distinct. The additional items

were created for this study to fill what the present authors saw as gaps or remaining questions often asked by potential adopters in clinical and public health settings or policy makers that were not addressed by the original 10 PRECIS items. It remains to be seen whether the added items will be of value to program planners, those conducting systematic reviews, or to practitioners considering program adoption. An alternative might be to develop additional items more directly linked to the CONSORT Work Group on Pragmatic Trials recommended reporting criteria (Zwarenstein et al. 2008). The ultimate evaluation of the utility of both the PRECIS and added external validity ratings, which will not be possible for some time, would be the extent to which such scores predict eventual success of programs and their adoption, implementation, and sustainability in real-world settings. As a reviewer noted, if one is only using the PRECIS criteria to provide checks in the planning stage of a trial, some unreliability in the PRECIS criteria may not be problematic.

This study has several limitations, including the small number of different trials rated. Rating only three trials, all of which were part of a collaborative research program, likely decreased the potential variability in PRECIS scores. Since results of the POWER trials are not yet available, it is not possible to conclude how ultimately useful such scores will be. Despite this, our experiences suggest that it is feasible to operationalize the PRECIS rating criteria (and the external validity items) and to rate intervention projects with a modest amount of training and time commitment. We recommend that applications of the PRECIS (and related scales such as the external validity rating items) be completed by (or at least include) independent raters not associated with the studies being rated. The PRECIS criteria were developed to help with trial design, and we conclude that they were useful for this purpose. In addition, we encourage health program planners, researchers, and health decision makers to become familiar with, understand, and use the PRECIS criteria and similar systems and scales to enhance transparency and aid in program design and adoption decisions.

We offer the following hypotheses that can be tested in future research. Studies, and especially the intervention programs, that are rated as more pragmatic will:

1. Be more likely to be broadly adopted (because of their generally greater flexibility and perceived relevance to practitioners), and

2. Produce lower effect sizes on standard outcome measures (because of potentially greater variability in more pragmatic studies and especially when conducted in low-resource settings).

In summary, we encourage other researchers to use and report on the utility of the PRECIS ratings, and use them to enhance the transparency of reporting, as well as in systematic reviews to characterize the literature on a given topic.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: Healthways, Inc. developed the website for both interventions for the Johns Hopkins study, in collaboration with Hopkins investigators, and provided coaches for the CCD intervention. Healthways also provided some research funding to supplement NIH support. Johns Hopkins University has an institutional consulting agreement with Healthways, Inc. Under this institutional agreement, the University is entitled to fees for consulting services. Those faculty investigators who participate in the consulting services receive a portion of the University fees, either as research support or salary supplement as determined by their supervisors. The terms of this arrangement are managed by the Johns Hopkins University in accordance with its conflict of interest policies. The opinions expressed do not necessarily reflect those of the NIH.

Funding Source: National Heart, Lung, Blood Institute (NHLBI)—Grant #5 U01 HL087071-01, U01 HL 087085-01, and U01 HL 087072-01. Representatives from the funding agency (NHLBI) did not participate in the ratings, analyses, or results. They commented only on the write-up, introduction, and discussion.

Disclosures: None.

REFERENCES

Abrams, D. B., C. T. Orleans, R. S. Niaura, M. G. Goldstein, J. O. Prochaska, and W. Velicer. 1996. "Integrating Individual and Public Health Perspectives for Treatment of Tobacco Dependence under Managed Health Care: A Combined Stepped Care and Matching Model." *Annals of Behavioral Medicine* 18: 290–304.

Derbas, J., M. Vetter, S. Volger, Z. Khan, E. Panigrahi, A. G. Tsa, et al. 2009. "Improving Weight Management in Primary Care Practice: A Possible Role for Auxiliary

Health Professionals Collaborating with Primary Care Physicians." *Obesity and Weight Management* 5: 222–8.

Glasgow, R. E. 2008. "What Types of Evidence Are Most Needed to Advance Behavioral Medicine?" *Annals of Behavioral Medicine* 35: 19–25.

Glasgow, R. E., L. M. Klesges, D. A. Dzewaltowski, S. S. Bull, and P. Estabrooks. (2004). "The Future of Health Behavior Change Research: What Is Needed to Improve Translation of Research into Health Promotion Practice?" *Annals of Behavioral Medicine* 27: 3–12; PMID 14979358.

Glasgow, R. E., K. W. Davidson, P. L. Dobkin, J. Ockene, and B. Spring. (2006). "Practical Behavioral Trials to Advance Evidence-Based Behavioral Medicine." *Annals of Behavioral Medicine* 31: 5–13; PMID 16472033.

Greaney, M. L., L. M. Quintiliani, E. T. Warner, D. K. King, K. M. Emmons, G. A. Colditz, R. E. Glasgow, and G. G. Bennett 2009. "Weight Management among Patients at Community Health Centers: The "Be Fit, Be Well" Study." *Obesity and Weight Management* 5: 218–24.

Institute of Medicine, Committee on Quality Health Care in America. 2003. *Crossing the Quality Chasm: A New Health System for the 21st Century.* Washington, DC: National Academies Press.

Jerome, G. J., Y. Hsin-Chieh, A. Dalcin, J. Reynolds, M. E. Gauvey-Kern, J. Charleston, N. Durkin, and L. J. Appel 2009. "Treatment of Obesity in Primary Care Practice: The Practice Based Opportunities for Weight Reduction (POWER) Trial at Johns Hopkins." *Obesity and Weight Management* 5: 216–21.

McGlynn, E. A., S. M. Asch, J. Adams, J. Keesey, J. Hicks, A. DeCristofaro and E. A. Kerr 2003. "The Quality of Health Care Delivered to Adults in the United States." *New England Journal of Medicine* 348: 2635–45.

McGraw, K. O., and S. P. Wong. 1996. "Forming Inferences about Some Intraclass Correlation Coefficients." *Psychological Methods* 1: 390.

Mercer, S. M., B. J. DeVinney, L. J. Fine, and L. W. Green. 2007. "Study Designs for Effectiveness and Tranlsation Research: Identifying Trade-Offs." *American Journal of Preventive Medicine* 33: 139–54.

Rothwell, P. M. 2005. "External Validity of Randomised Controlled Trials: To Whom Do the Results of This Trial Apply?" *Lancet* 365: 82–93.

Simons-Morton, D. G., K. J. Calfas, B. Oldenburg, and N. W. Burton. 1998. "Effects of Interventions in Health Care Settings on Physical Activity or Cardiorespiratory Fitness." *American Journal of Preventive Medicine* 15: 413–30.

The Dartmouth Institute for Health Policy & Clinical Practice. (2010). Standards for Quality Improvement Reporting Excellence (SQUIRE) [accessed October 19, 2011]. Available at http://www.squire-statement.org.

Thorpe, K. E., M. Zwarenstein, A. D. Oxman, S. Treweek, C. D. Furberg, D. G. Altman, S. Tunis, E. Bergel, I. Harvey, D. J. Magid, and K. Chalkidou 2009. "A Pragmatic-Explanatory Continuum Indicator Summary (PRECIS): A Tool to Help Trial Designers." *CMAJ* 180: E47–57.

Tunis, S. R., D. B. Stryer, and C. M. Clancey. (2003). "Practical Clinical Trials: Increasing the Value of Clinical Research for Decision Making in Clinical and Health

Policy." *Journal of the American Medical Association* 290: 1624–32; PMID 14506122.

Wells, B. (2009). "Weight Loss in Obese Adults with Cardiovascular Risk Factors: Three Randomized Control Trials to Assess Interventions in Clinical Practice." *Obesity and Weight Management* 5: 207–9, doi: 10.1089/obe.2009.0504.

Yeh, H. C., J. M. Clark, K. E. Emmons, R. H. Moore, G. G. Bennett, E. T. Warner, D. B. Sarwer, G. J. Jerome, E. R. Miller, S. Volger, T. A. Louis, B. Wells, T. A. Wadden, G. A. Colditz, and L. J. Appel 2010. "Independent but Coordinated Trials: Insights from the Practice-Based Opportunities for Weight Reduction Trials Collaborative Research Group." *Clinical Trials* 7: 322–32.

Zerhouni, E. A. 2005. "Translational and Clinical Science – Time for a New Vision." *New England Journal of Medicine* 353: 1621–3.

Zwarenstein, M., S. Treweek, J. J. Gagnier, D. G. Altman, S. Tunis, B. Haynes, A. D. Oxman, and D. Moher 2008. "Improving the Reporting of Pragmatic Trials: An Extension of the CONSORT Statement." *British Medical Journal* 337: a2390, doi: 10.1136/bmj

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article:

Appendix SA1: Author Matrix.
Table S1: Power Precis Rating.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.