Monday, September 11, 2017

Improving Peer Review and Scientific Publication

Quality of Reporting

Association of Journal-Level and Study-Level Variables With Proximity of Primary Study Results to Summary Estimates From Meta-analyses in Imaging Journals

Robert Frank,[1] Matthew McInnes,[2,3] Deborah Levine,[4,5] Herbert Kressel,[4,5] Julie Jesurum,[6] William Petrcich,[3] Trevor McGrath,[1] Patrick M. Bossuyt[7]

Objective Are most research findings false but less so in high-impact journals? Variables associated with false results in imaging research are poorly understood. While absolute truth is elusive, high-quality meta-analyses using hierarchical methods produce high-level evidence with minimal random variability in their results, thereby representing a convenient proxy for truth. We evaluated whether journal-level and study-level variables are associated with the distance between primary study results and summary estimates from meta-analyses.

Design In this meta-research study, PubMed was searched for diagnostic accuracy meta-analyses using hierarchical methods, published in imaging journals between January 2005 and April 2016. Data were extracted for each meta-analysis and its included primary studies, including study demographic information, journal Impact Factor, journal cited half-life, Standards for Reporting Diagnostic accuracy studies (STARD) endorsement, citation rate, publication date, sample size, sensitivity, and specificity. Meta-analyses were excluded for failing to report both primary and summary accuracy estimates. Primary studies were divided into 2 groups for each variable assessed; groups were defined based on first publication vs subsequent publications on a topic, publication before vs after STARD introduction, presence vs absence of STARD endorsement, or by median split. The mean absolute deviation of primary study estimates from the corresponding summary estimates for sensitivity and specificity was compared between groups for each variable. Analyses were performed using a model combining a γ distribution for absolute deviations greater than 0 with an estimated probability that the absolute deviation is 0. Means and 95% CIs were obtained using bootstrap resampling. *P* values were calculated using a *t* test. The threshold for significance was defined as $P < .004$ after Bonferroni correction (.05/12) to mitigate bias owing to multiple comparisons.

Results Ninety-eight meta-analyses containing 1458 primary studies met inclusion criteria. There was substantial variability in deviations from the summary estimate between paired groups, but no variable demonstrated a significant association with proximity of primary study diagnostic accuracy estimates to the pooled estimates from their corresponding meta-analyses ($P > .004$ in all comparisons) (**Table 10**).

Conclusions Many variables considered important when selecting imaging diagnostic accuracy literature to guide clinical decisions are not associated with results that are more reflective of the truth as established by meta-analyses. The distance between primary study results and summary estimates of diagnostic accuracy is probably not smaller for studies published in higher versus lower Impact Factor journals.

[1]Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada, robert.frank@uottawa.ca; [2]Department of Radiology, The Ottawa Hospital, University of Ottawa, Ottawa, ON, Canada; [3]Ottawa Hospital Research Institute, Ottawa, ON, Canada; [4]Department of Radiology, Beth Israel Deaconess Medical Center, Boston, MA, USA; [5]*Radiology* Editorial Office, Boston, MA, USA; [6]Harvard University, Boston, MA, USA; [7]Clinical Epidemiology and Biostatistics,

Table 10. Association of Journal-Level and Study-Level Variables With Proximity of Primary Study Results to Summary Estimates From Meta-analyses[a]

Variable	Group Analysis	Difference in Mean Deviation of Primary Estimates From Summary Estimates Between Dichotomized Groups			
		Sensitivity	*P* Value	Specificity	*P* Value
Impact Factor	Above median vs below median	−0.018	.09	−0.013	.11
STARD endorsement	Endorsement vs no endorsement	−0.0057	.60	0.0019	.83
Cited half-life	Above median vs below median	−0.0063	.55	0.0097	.24
Citation rate	Above median vs below median	−0.018	.08	−0.0045	.58
Publication timing (relative to STARD 2003)	Post-STARD vs pre-STARD	0.0059	.55	−0.0082	.38
Publication timing (first published)	First published vs later published	−0.025	.005	0.0077	.48

Abbreviation: STARD, Standards for Reporting Diagnostic accuracy studies.
[a]Statistical significance defined as $P < .004$ (after Bonferroni correction).

Bioinformatics Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands

Conflict of Interest Disclosures: None reported.

Funding/Support: This work was supported in part by stipends from the Ottawa Hospital Department of Radiology and the Undergraduate Research Opportunity Program.

Role of the Funder/Sponsor: The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the abstract.

Discrepancies in Reporting Between Trial Publications and Clinical Trial Registries in High-Impact Journals

Sarah Daisy Kosa,[1,2] Lawrence Mbuagbaw,[1] Victoria Borg Debono,[1] Mohit Bhandari,[1] Brittany B. Dennis,[1] Gabrielle Ene,[4] Alvin Leenus,[3] Daniel Shi,[3] Michael Thabane,[1] Thuva Vanniyasingam,[1] Chenglin Ye,[5] Elgene Yranon,[1] Shiyuan Zhang,[1] Lehana Thabane[1]

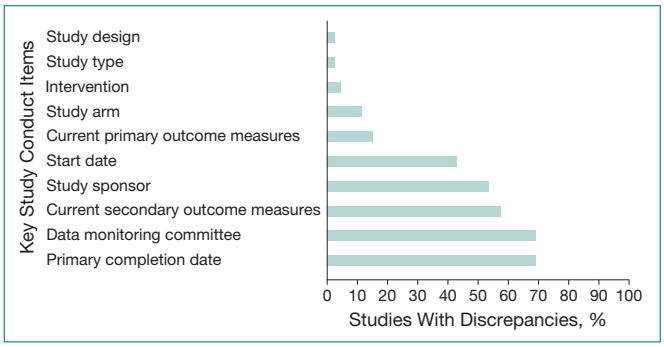
Objective It is currently unclear the extent to which key information mandatory for clinical trial registries is reported in published manuscripts. To address this gap in the literature, the primary objective of this study was to examine the percentage of studies where there are discrepancies in reporting of key study conduct items between the clinical trial registry and the manuscript.

Design We searched PubMed for all randomized clinical trials (RCTs) published between 2012 and 2015 in the top 5 general medicine journals (based on the 2014 impact factor as published by Thomson Reuters), which all required registration of the RCT for publication; 200 full-text publications (50 from each year) were randomly selected for data extraction. Key study conduct items were extracted by 2 independent reviewers for each year. When an item was reported differently or not reported at all in either source, this was considered a discrepancy in reporting between the registry and the full-text publication. Descriptive statistics were calculated to summarize the percentage of studies with discrepancies between the registry and the published manuscript in reporting of key study conduct items. The items of interest were design (ie, randomized control trial, cohort study, case control study, case series), type (ie, retrospective, prospective), intervention, arms, start and end dates (based on month and year where available), use of data monitoring committee, and sponsor, as well as primary and secondary outcome measures.

Results In the sample of 200 RCTs, there were relatively few studies with discrepancies in study design (n=6 [3%]), study type (n=6 [3%]), intervention (n=10 [5%]), and study arm (n=24 [12%]) (**Figure 2**). Only 30 studies (15%) had discrepancies in their primary outcomes. However, there were often discrepancies in study start date (n=86 [43%]), study sponsor (n=108 [54%]), and secondary outcome measures (n=116 [58%]). Almost 70% of studies had discrepancies regarding the use of a data monitoring committee and primary completion date reporting.

Conclusions We identified discrepancies in reporting between publications and clinical trial registries. These findings are limited by only being based on a subset of RCTs in the included journals and may not be generalizable to all RCTs within that journal, other disciplines, journals in other languages, or lower-impact journals. Further measures are needed to improve reporting given the potential threats to the quality and integrity of scientific research.

Figure 2. Percentage of Studies With Discrepancies in Reporting Between Trial Publications and Clinical Trial Registries in 200 High-Impact Journals



[1]Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, Ontario, Canada, mbuagblc@mcmaster.ca; [2]Toronto General Hospital-University Health Network, Toronto, Ontario, Canada; [3]Princess Margaret Hospital-University Health Network, Toronto, Ontario, Canada; [4]Bachelor of Health Sciences Program, McMaster University, Hamilton, Ontario, Canada; [5]Oncology Biostatistics, Genentech, South San Francisco, CA, USA

Conflict of Interest Disclosures: None reported.

Methodological and Reporting Quality of Systematic Reviews Underpinning Clinical Practice Guidelines

Cole Wayant,[1] Matt Vassar[1]

Objective This study summarizes the findings of 3 separate studies conducted simultaneously to determine the methodological and reporting quality of systematic reviews (SRs) underpinning clinical practice guidelines (CPGs) in pediatric obesity, opioid use disorder, and ST-elevated myocardial infarction.

Design A search of guideline clearinghouse and professional organization websites was conducted for guidelines published by national or professional organizations. We included all reviews cited by authors of CPG, including Cochrane reviews, and removed duplicates prior to data extraction. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-analyses) and AMSTAR (A Measurement Tool to Assess Systematic Reviews) instruments were used to score SRs and meta-analyses cited in CPGs. PRISMA and AMSTAR are validated tools for measuring reporting quality and methodological quality, respectively.

Results The mean PRISMA total scores for the pediatric obesity, opioid use disorder, and ST-elevated myocardial infarction SRs across all CPGs were 16.9, 20.8, and 20.8, respectively. The mean AMSTAR total scores were 4.4, 8.8, and 6.1, respectively. Consistently underreported items on the PRISMA checklist were items 5 (protocol registration), 8 (search strategy), 15 (risk of bias for cumulative evidence), and 22 (risk of bias across studies). Consistently underreported items on the AMSTAR checklist were items 4 (duplicate extraction/validation), 5 (list of included/excluded studies), 8 (quality of evidence assessments), 10 (publication bias assessments), and 11 (conflict of interest disclosure). Altogether, our study included 150 SRs and 29 CPGs, with only 9 CPGs assigning grades to their recommendations. The 150 SRs were cited a total of 308 times: 95 times as a direct evidence for graded recommendations, 21 times as direct evidence for nongraded recommendations, 189 times as supporting evidence, and 3 times for unclear reasons.

Conclusions These investigations into CPGs in pediatric obesity, opioid use disorder, and ST-elevated myocardial infarction revealed a consistent lack of overall methodological and reporting quality in the included SRs as well as heterogeneity in the use of grading scales, or lack thereof. Because SRs are considered by most to be level 1A evidence, an apparent lack of quality may impair clinical decision making and hinder the practice of evidence-based medicine. Items such as PRISMA items 15 and 22 and AMSTAR items 10 and 11 are of particular concern because these items ensure that bias assessments are performed and conflicts of interests are disclosed.

[1]Oklahoma State University Center for Health Sciences, Department of Analytical and Institutional Research, Tulsa, OK, USA, cole.wayant@okstate.edu

Conflict of Interest Disclosures: None reported.

Additional Contributors: We acknowledge the work done by the members of the team that conducted the investigations into opioid use disorder (Andrew Ross and Justin Rankin) and ST-elevated myocardial infarction (Jared Scott and Ben Howard) as well as all of the faculty and staff at the Oklahoma State University Center for Health Sciences and Oklahoma State University Medical Center that assisted in the process of completing the investigations.

Optimism Bias in Contemporary National Clinical Trial Network Phase 3 Trials

Kaveh Zakeri,[1] Sonal S. Noticewala,[1] Lucas K. Vitzthum,[1] Elena Sojourner,[1] Loren K. Mell[1]

Objective Overestimation of treatment effect sizes—termed *optimism bias*—in research protocols can lead to underpowered clinical trials that fail to demonstrate clinical benefits. We compared hypothesized vs. observed treatment effects to determine if there is evidence of optimism bias in contemporary NCTN phase III trials.

Design We queried PubMed for National Cancer Institute (NCI)–sponsored phase III randomized cooperative group clinical trials from January 2007 to January 2017. We identified 185 published trials. Trials with missing protocols (n = 56), equivalence or noninferiority trials (n = 5), trials that accrued less than 40% of their intended sample size (n = 14), and trials that pooled their data with other studies (n = 2) were excluded. For trials reporting time-to-event outcomes with hazard ratios (HRs) (n = 81), we compared the proposed effect size from the sample size calculation in the research protocol with the observed effect size in the published article to calculate the ratio of observed-to-proposed HRs overall and for trials that did or did not report statistically significant effect on primary end points. All HRs were standardized for a reduction in adverse events such that HRs less than 1 indicated a benefit to therapy. We also compared findings with those previously reported for NCI trials conducted from 1955 to 2006 and tabulated studies that provided a reference, evidence, or other specific rationale for their proposed effect size in the research protocol.

Results Data on 98,200 patients from 108 clinical trials were evaluated. The most common cancers were breast, gynecologic, gastrointestinal, brain, and genitourinary malignant neoplasms. The most common primary end point was overall survival (40.7%). The median ratio of observed-to-proposed HRs was 1.26 (range: 0.33-2.34). The median ratio of observed-to-proposed HRs among trials that observed a statistically significant effect on the primary end point was 1.09 (range: 0.33-1.29) vs 1.30 (range: 0.86-2.34) for trials that did not, compared with 1.34 and 1.86, respectively, for NCI trials conducted from 1955 to 2006. Twenty-four trials (22.2%) observed a statistically significant effect on the primary end point favoring the experimental treatment, compared with 24.6% previously reported. The majority of trials (76.9%) provided no rationale for the magnitude of the proposed treatment effect.

Conclusions Although most NCI-sponsored clinical trials conducted between 2007 and 2017 failed to establish statistically significant benefits of new therapies, the magnitude of optimism bias appears to have decreased compared with that in trials conducted between 1955 and 2006. Better rationalization of proposed effect sizes is needed in clinical trial protocols.

[1]Department of Radiation Medicine and Applied Sciences, University of California, San Diego, La Jolla, CA, USA, kzakeri@ucsd.edu

Conflict of Interest Disclosures: None reported.

Quality of the Scientific Literature

Scientific Quality in a Series of Comparative Effectiveness Research Studies

Harold Sox,[1,2] Evan Mayo-Wilson,[1,2] Kelly Vander Ley,[1,2] Marina Broitman,[1,2] David Hickam,[1,2] Steven Clauser,[1,2] Yen-Pin Chiang,[1,2] Evelyn Whitlock[1,2]

Objective Markers of high-quality comparative effectiveness research (CER) studies are largely unknown but could be valuable to funders and future applicants for CER funding.

Our long-term objective is to identify variables associated with CER scientific quality and impact. The objective of this preliminary report is to describe the frequency of measures of CER study quality.

Design This is a case series of CER studies funded during the first funding cycle (2013) of the Patient-Centered Outcomes Research Institute (PCORI). Awardees are required to submit a final research report (FRR), which undergoes external peer review and is published on the PCORI website when the principal investigator (PI) meets revision requirements. We are using the original application to investigate study and PI-related variables potentially associated with study quality, and are assessing study quality of the peer-reviewed report using US Preventive Services Task Force (USPSTF) criteria (good, fair, poor) and adherence to PCORI methodology standards. When the case series is complete we will study associations between markers of study quality and publication outcomes (citations in published articles, systematic reviews, and practice guidelines; Altmetric scores) and between direct study quality measures and those outcomes.

Results Among 98 FRRs received by early June 2017, 5 have completed peer review. Candidate PI-based variables potentially associated with study quality include number of research awards from the National Institutes of Health, Agency for Healthcare Research and Quality, Centers for Disease Control and Prevention, or the Department of Veterans Affairs (mean, 1 [range, 0-5]; median, 0); number of studies in major journals as first or last author (mean, 2.8 [0-8]; median, 0), the PI's H factor (mean, 23 [range, 16-41]; median, 22), and years since he or she was granted the highest academic degree (19.2 [range, 16-21]; median, 20). All 5 studies were of "fair" quality according to USPSTF grading criteria. Each of PCORI's 5 cross-cutting Methodology Standards (which had not been published when the studies in this report were funded) comprise several component standards (range, 7-17), and rates of meeting the standards varied from 15% (standard for managing missing data) to 34% (standard for formulating research questions). We expect to complete peer review and to report on 20 more research reports.

Conclusions With short-term follow-up on a series of approximately 300 CER studies funded by PCORI through July 2019, this study may eventually provide measures of specific methodological shortcomings and variables associated with CER quality and impact.

[1]Patient-Centered Outcomes Research Institute, hsox@pcori.org, Johns Hopkins University, Baltimore, MD, USA; [2]Oregon Health Sciences University, Portland, OR, USA

Conflict of Interest Disclosures: None reported.

Pitfalls in the Use of Statistical Methods in Systematic Reviews of Therapeutic Interventions: A Cross-sectional Study

Matthew J. Page,[1] Douglas G. Altman,[2] Larissa Shamseer,[3,4] Joanne E. McKenzie,[1] Nadera Ahmadzai,[5] Dianna Wolfe,[5] Fatemeh Yazdi,[5] Ferrán Catalá-López,[5,6] Andrea C. Tricco,[7,8] David Moher[3,4]

Objective Researchers have identified several problems in the application of statistical methods in published systematic reviews (SRs). However, these evaluations have been narrow in scope, focusing only on one particular method (such as sensitivity analyses) or restricting inclusion to Cochrane SRs, which make up only 15% of all SRs of biomedical research. We aimed to investigate the application and interpretation of various statistical methods in a cross-section of SRs of therapeutic interventions, without restriction by journal, clinical condition, or specialty.

Design We selected articles from a database of SRs we assembled previously. These articles consisted of a random sample of 300 SRs addressing various questions (therapeutic, diagnostic, or etiologic) that were indexed in MEDLINE in February 2014. In the current study, we included only those SRs that focused on a therapeutic question, reported at least 1 meta-analysis, and were written in English. We collected data on 61 prespecified items that characterized how well random-effects meta-analysis models, subgroup analyses, sensitivity analyses, and funnel plots were applied and interpreted. Data were extracted from articles and online appendices by a single reviewer, with a 20% random sample extracted in duplicate.

Results Among 110 SRs, 78 (71%) were non-Cochrane SRs and 55 (50%) investigated a pharmacological intervention. The SRs presented a median of 13 (interquartile range, 5-27) meta-analyses. Among the 110 primary meta-analyses in each SR, 62 (56%) used the random-effects model but only 5 of 62 (8%) interpreted the pooled result correctly (that is, as the average of the intervention effects across all studies). Subgroup analyses were reported in 42 of 110 SRs (38%), but findings were not interpreted with respect to a test for interaction in 29 of 42 cases (69%), and the issue of potential confounding in the subgroup analyses was not raised in any SR. Sensitivity analyses were reported in 51 of 110 SRs (46%), without any rationale in 37 of 51 cases (73%). Authors of 37 of 110 SRs (34%) reported that visual inspection of a funnel plot led them to not suspect publication bias. However, in 28 of 37 cases (76%), fewer than 10 studies of varying size were included in the plot.

Conclusions There is scope for improvement in the application and interpretation of statistical analyses in SRs of therapeutic interventions. Guidelines such as PRISMA may need to be extended to provide more specific statistical guidance.

[1]School of Public Health and Preventive Medicine, Monash University, Melbourne, Victoria, Australia, matthew.page@monash.edu; [2]UK EQUATOR Centre, Centre for Statistics in Medicine, NDORMS, University of Oxford, Oxford, UK; [3]Centre

for Journalology, Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, Ontario, Canada; [4]School of Epidemiology, Public Health, and Preventive Medicine, Faculty of Medicine, University of Ottawa, Ottawa, Ontario, Canada; [5]Knowledge Synthesis Group, Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, Ontario, Canada; [6]Department of Medicine, University of Valencia/INCLIVA Health Research Institute and CIBERSAM, Valencia, Spain; [7]Knowledge Translation Program, Li Ka Shing Knowledge Institute, St Michael's Hospital, Toronto, Ontario, Canada; [8]Epidemiology Division, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada

Conflict of Interest Disclosures: Douglas G. Altman and David Moher are Peer Review Congress Advisory Board Members but were not involved in the review or decision for this abstract.

Introducing Reporting Guidelines and Checklists for Contributors to *Radiology*: Results of an Author and Reviewer Survey

Marc Dewey,[1] Deborah Levine,[2,3] Patrick M. Bossuyt,[4] Herbert Y. Kressel[5,6]

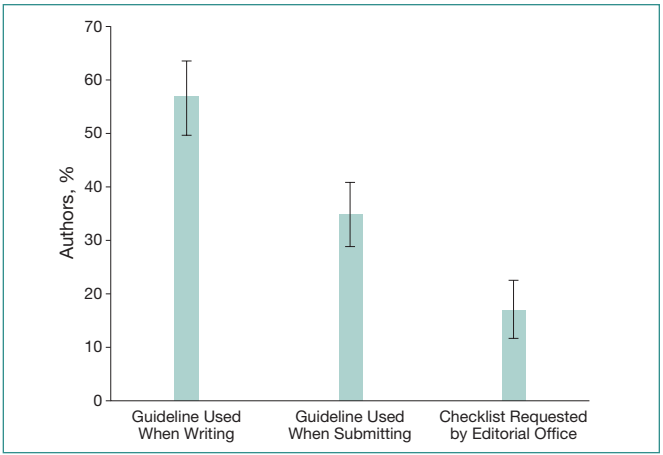
Objective Numerous reporting guidelines have been developed to make study reports more informative, but it is uncertain whether they are perceived as useful by authors and reviewers. We surveyed the use and perceived value of reporting guidelines after an initiative begun in January 2016 that required authors to submit appropriate guideline checklists along with their manuscripts prior to peer review.

Design Cohort study of authors of original research submissions to *Radiology* between July 5, 2016, and June 1, 2017, and of reviewers who had performed reviews since January 2016. Authors were asked to complete an anonymized online survey within 2 weeks of manuscript submission but before the editorial decision was made. Reviewers were surveyed with similar questions from May 17, 2017, until June 1, 2017.

Results A total of 831 of 1391 authors (59.7%) completed the survey within a mean (SD) of 1.5 (2.7) days (range, 0-17 days) of the request. Consistent with the types of studies submitted to *Radiology*, most authors used STROBE (447 of 829 authors [53.9%]) or STARD (313 authors [37.8%]) and only a small minority used CONSORT (40 authors [4.8%]) or PRISMA (29 authors [3.5%]). Only 120 of 821 authors (14.6%) used the guideline and checklist when designing the study, more so for PRISMA users (16 of 29 [55%]), less so for STARD users (52 of 310 [16.8%]; $P < .001$) and STROBE users (46 of 443 [10.4%]; $P < .001$). The guidelines were used by 189 of 821 authors (23.0%) when writing the manuscript; these authors more often reported an impact on the final manuscript (107 of 189 [56.6%]) compared with those who used the guideline when submitting the manuscript (95 of 272 [34.9%]; $P < .001$) or when the checklist was requested by the editorial office (41 of 240 [17.1%]; $P < .001$) (**Figure 3**). Filling out the checklist was considered very useful by 256 of 819 authors (31.3%), somewhat useful by 390 (47.6%), not very useful by 122 (14.9%), and not at all useful by 51 (6.2%). The response rate of reviewers was 32.1% (259 of 808 reviewers). The checklist was used by 200 of 259 reviewers (77.2%) some or all of the time, and 119 of 199 (59.8%) said it affected their reviews. Having the checklist for review was considered very useful by 28 of 256 reviewers (10.9%), somewhat useful by 106 (41.4%), not very useful by 82 (32.0%), and not at all useful by 40 (15.6%).

Conclusions Almost 4 of 5 authors and half of the reviewers judged the guideline checklists to be useful or very useful. Using the guidelines while writing the manuscript was associated with greater impact on the final manuscript.

Figure 3. Impact on Manuscripts Depending on When Guideline and Checklist Were Used



Significantly more authors who used the guidelines and checklist when writing the manuscript reported an impact on the final manuscript (107 of 189 authors [56.6%]) compared with those who used the guideline when submitting the manuscript (95 of 272 authors [34.9%]; $P < .001$) or when the checklist was requested by the editorial office (41 of 240 authors [17.1%]; $P < .001$). Error bars show 95% CIs.

[1]Department of Radiology, Charité–Universitätsmedizin Berlin, Berlin, Germany, marc.dewey@charite.de; [2]Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA; [3]Senior Deputy Editor, *Radiology*, Boston, MA, USA; [4]Amsterdam Public Health Research Institute, Department of Clinical Epidemiology, Biostatistics, and Bioinformatics, University of Amsterdam, Amsterdam, the Netherlands; [5]Department of Radiology, Harvard Medical School, Boston, MA, USA; [6]Editor, *Radiology*, Boston, MA, USA

Conflict of Interest Disclosures: Marc Dewey was an Associate Editor of *Radiology* at the time of initiation of this study and is a consultant to the editor now. Debbie Levine is Senior Deputy Editor of *Radiology*. Patrick Bossuyt is the lead senior author of the STARD guidelines and checklist. Herbert Y. Kressel is the Editor in Chief of *Radiology* and an author of the STARD 2015 guidelines.

Funding/Support: This study was supported by the Young Leaders Club program of the International Society of Strategic Studies in Radiology and the Heisenberg Program of the German Research Foundation.

Role of the Funder/Sponsor: The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the abstract.

Reported Use of Standard Reporting Guidelines Among *JNCI* Authors, Editorial Outcome, and Reviewer Ratings Related to Adherence to Guidelines and Clarity of Presentation

Jeannine Botos[1]

Objective A study was conducted to examine associations between author-reported use of standard reporting guidelines (SRGs) to prepare *JNCI* submissions with editorial decisions and reviewer ratings for adherence to reporting guidelines and clarity of presentation.

Design At submission authors were asked if they used SRGs to prepare their manuscript and, if so, which one(s). Reviewers rated (poor, fair, good, very good, outstanding, not applicable) adherence to reporting guidelines and clarity of presentation. This information was collected using a customized Editorial Manager Enterprise Analytics Report for submissions with first or final decisions that were submitted between November 1, 2015 and April 30, 2017. All manuscript types that would benefit from the use of SRGs were included (ie, Articles, Brief Communications, Reviews, MiniReviews, Systematic Reviews, and Meta-analyses). Each peer-reviewed submission received 1 to 3 ratings per question and all ratings were included in the analyses. Numerical values were given to each answer (SRG use, 1; no SRG use, 0) or reviewer rating (not applicable, 0; fair, 1; poor, 2; good, 3; very good, 4; and outstanding, 5), and scores were compared using 2-sided *t* tests.

Results Of 2209 submissions included in the analysis, 1144 (51.8%) indicated that at least 1 SRG was used (**Table 11**). The STROBE guidelines were the most common (n = 531, 24.0%). Of the 2068 (93.6%) submissions that were rejected, 1105 (50.1%) indicated using SRGs and 963 (43.6%) did not (mean [SD] scores of rejected vs not rejected, 0.53 [0.50] vs 0.49 [0.50], *P* = .47). Of the 1033 ratings for adherence to reporting guidelines, mean (SD) scores for not rejected vs rejected submissions were 3.2 (1.61) vs 2.9 (1.57) (*P* = .005), and mean (SD) scores for SRG use vs no use were 3.1 (1.48) vs 2.9 (1.70) (*P* = .01). Of the 1036 ratings for clarity of presentation, mean (SD) scores for not rejected vs rejected submissions were 3.6 (1.00) vs 3.1 (1.08) (*P* < .001), whereas mean (SD) scores for SRG use vs no use were 3.3 (1.04) vs 3.3 (1.10) (*P* = .64).

Conclusions Among these *JNCI* submissions, reporting the use of SRGs was not associated with editorial decisions or with reviewer ratings for clarity of presentation. Reviewer ratings for adherence to guidelines and clarity of presentation were associated with editorial decisions after peer review, and ratings for adherence to guidelines were associated with reported use of SRGs.

[1]*JNCI*, Oxford University Press, New York, NY, USA, jeannine.botos@oup.com

Conflict of Interest Disclosures: None reported.

Table 11. Reported Use of Standard Reporting Guidelines Among JNCI Authors, Editorial Outcomes, and Reviewer Ratings for Adherence to Guidelines and Clarity of Presentation for Articles, Reviews, Mini-Reviews, Systematic Reviews, Meta-analysis, and Brief Communications[a]

Editorial Decision or Reviewer Question	All, no (%)	Reported Using a SRG				Adherence to Reporting Guidelines					Clarity of Presentation				
		No, no. (%)	Any, no. (%)	Mean Score (SD)	*P*[b]	All, no. (%)	No SRG, no. (%)	Any SRG, no. (%)	Mean Score (SD)	*P*[c]	All, no. (%)	No SRG, no. (%)	Any SRG, no. (%)	Mean Score (SD)	*P*[c]
All submissions	2209 (100)	1065 (48.2)	1144 (51.8)	0.52 (0.5)		1033 (100)	552 (53.4)	481 (46.6)	3.0 (1.6)		1036 (100)	487 (47.0)	549 (53.0)	3.3 (1.1)	
Rejected without peer review	1813 (82.1)	875 (39.6)	938 (42.5)	0.53 (0.5)		NA	NA	NA	NA		NA	NA	NA	NA	
Rejected after peer review	255 (11.5)	88 (4.0)	167 (7.6)	0.53 (0.5)	.68	609 (58.9)	343 (33.2)	266 (25.7)	2.9 (1.6)		608 (58.7)	340 (32.8)	268 (27.6)	3.1 (1.1)	
Not rejected after peer review	141 (6.4)	102 (4.6)	39 (1.8)	0.49 (0.5)	.47	424 (41.0)	209 (20.2)	215 (20.8)	3.2 (1.6)	.004	428 (41.3)	219 (21.1)	209 (20.2)	3.6 (1.0)	P<.001
Adherence to reporting guidelines, mean score (SD)	3.0 (1.6)	2.9 (1.7)	3.1 (1.5)		.01										
Clarity of presentation, mean score (SD)	3.3 (1.1)	3.3 (1.1)	3.3 (1.0)		.64										

Abbreviations: SRG, standard reporting guideline; NA, not applicable.

[a]Authors reported using the following SRGs: Strengthening-Reporting of Observational-Studies in Epidemiology (STROBE); Animal Research: Reporting In Vivo Experiments (ARRIVE); Minimum Information for Publication of Quantitative Real-Time PCR Experiments (MIQE); Consolidated Standards of Reporting Trials (CONSORT); REporting recommendations for tumour MARKer prognostic studies (REMARK); Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA); studies of diagnostic accuracy (STARD); Meta-analyses of Observational Studies (MOOSE); Biospecimen reporting for improved study quality (BRISQ); STrengthening the REporting of Genetic Association Studies (STREGA), an extension to STROBE; and Consolidated Health Economic Evaluation Reporting Standards (CHEERS). Some percentages do not add to 100 owing to rounding. Numerical values were given to each answer (SRG use, 1; no SRG use, 0) or reviewer rating (not applicable, 0; fair, 1; poor, 2; good, 3; very good, 4; and outstanding, 5), and mean scores are presented. *P* values were calculated using a 2-sided paired *t* test.

[b]*P* comparing scores for SRG use vs no SRG use.

[c]*P* comparing scores for rejected vs not rejected editorial decisions.

Impact of an Intervention to Improve Compliance With the ARRIVE Guidelines for the Reporting of In Vivo Animal Research

Emily Sena,[1] for the Intervention to Improve Compliance With the ARRIVE Guidelines (IICARus) Collaborative Group

Objective To conduct a randomized controlled trial to determine whether journal-mandated completion of an ARRIVE checklist (requiring authors to state on which page of their manuscript each checklist item is met) improves full compliance with the ARRIVE guidelines.

Design Manuscripts submitted to *PLOS One* between March 2015 and June 2015 determined in the initial screening process to describe in vivo animal research were randomized to either mandatory completion and submission of an ARRIVE checklist or the normal editorial processes, which do not require any checklist submission. The primary outcome was between-group differences in the proportion of studies that comply with the ARRIVE guidelines. We used online randomization with minimization (weighted at 0.75) according to country of origin; this was performed by the journal during technical checks after submission. Authors, academic editors, and peer reviewers were blinded to the study and the allocation. Accepted manuscripts were redacted for information relating to the ARRIVE checklist by an investigator who played no further role in the study to ensure outcome adjudicators were blinded to group allocation. We performed outcome adjudication in duplicate by assessing manuscripts against an operationalized version of the ARRIVE guidelines that consists of 108 items. Discrepancies are being resolved by a third independent reviewer.

Results We randomly assigned 1689 manuscripts, with 844 manuscripts assigned to the control arm and 845 assigned to the intervention arm. Of these, 1299 (76.9%) were sent for review, and of these, 688 (53.0%) were accepted for publication. All 688 manuscripts were dual assessed, and reconciliation of discrepancies is ongoing. Agreement between reviewers was high in relation to questions of the species reported (93%) and measures to reduce the risk of bias (73%-91% for 6 questions) and lowest for reporting the unit of analysis (50%). Data analysis is ongoing. We will present data for between-group differences in the proportion of studies that comply with the ARRIVE guidelines, each of the 38 subcomponents of the ARRIVE checklist, each of the 108 items, and the proportion of submitted manuscripts accepted for publication.

Conclusions Our study will determine the effect of an alteration of editorial policy to include a completed ARRIVE checklist with submissions on compliance with the ARRIVE guidelines in the work when published. These results will inform the future development and further implementation of the ARRIVE guidelines.

[1]Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK, emily.sena@ed.ac.uk

Conflict of Interest Disclosures: The study management committee included a representative from the Public Library of Science (Catriona MacCallum), but other than providing general advice during the design of the study and organizing the provision of PDFs of included manuscripts, they had no role.

Funding/Support: The Medical Research Council, National Centre for the Replacement Refinement and Reduction of Animals in Research, Biotechnology and Biological Sciences Research Council, and Wellcome Trust pooled resources without a normal grant cycle to fund this project.

Role of the Funder/Sponsor: The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the abstract. The funders used their social media streams to publicize the study and recruit outcome assessors. National Centre for the Replacement, Refinement, and Reduction of Animals in Research employees were not allowed to enroll as outcome assessors because of their possible conflict of interest as sponsors of the ARRIVE guidelines.

Group Members: The IICARus Collaborative group includes the following members: *University of Edinburgh, Edinburgh, UK*: Emily Sena, Cadi Irvine, Kaitlyn Hair, Fala Cramond, Paula Grill, Gillian Currie, Alexandra Bannach-Brown, Zsanett Bahor, Daniel-Cosmin Marcu, Monica Dingwall, Victoria Hohendorf, Klara Zsofia Gerlei, Victor Jones, Anthony Shek, David Henshall, Emily Wheater, Edward Christopher, and Malcolm Macleod; *University of Tasmania, Hobart, Tasmania*: David Howells; *University of Nottingham, Nottingham, UK*: Ian Devonshire and Philip Bath; *Public Library of Science, Cambridge, UK*: Catriona MacCallum; *Imperial College London, London, UK*: Rosie Moreland; *Mansoura University, Mansoura, Egypt*: Sarah Antar, Mona Hosh, and Ahmed Nazzal; *University of New South Wales, Kensington, NSW, Australia*: Katrina Blazek; *Animal Sciences Unit, Animal and Plant Health Agency, Addlestone, UK*: Timm Konold; *University of Glasgow, Glasgow, UK*: Terry Quinn and Teja Gregorc; *AstraZeneca, Wilmington, Delaware, USA*: Natasha Karp; *Nuffield Research Placement Student, London, UK*: Privjyot Jheeta and Ryan Cheyne; *GlaxoSmithKline, Middlesex, UK*: Joanne Storey; *University College London, London, UK, and École Normale Supérieure, Paris, France*: Julija Baginskaite; *University Medical Center Utrecht, Utrecht, the Netherlands*: Kamil Laban; *University of Rome Sapienza, Rome, Italy*: Arianna Rinaldi; *Radboud University Nijmegen Medical Centre, Nijmegen, the Netherlands*: Kimberley Wever; *University of Southampton, Southampton, UK*: Savannah Lynn; *Federal University of Rio de Janeiro, Rio de Janeiro, Brazil*: Evandro Araújo De-Souza; *University of Birmingham, Birmingham, UK*: Leigh O'Connor; *Hospital Research Center of the Sacred Heart of Montreal, Montreal, QC, Canada*: Emmanuel Charbonney; *National Cancer Institute, Milano, Italy*: Marco Cascella; *Federal University of Santa Catarina, Florianópolis, Brazil*: Cilene Lino de Oliveira; *University of Geneva, Geneva, Switzerland*: Zeinab Ammar; *British American Tobacco, London, UK*: Sarah Corke; *Ministry of Health, Cairo, Egypt*: Mahmoud Warda; *Vita-Salute San Raffaele University, Milan, Italy*: Paolo Roncon; *University of Hertfordshire, Hertfordshire, UK*: Daniel Baker; *University of Veterinary Medicine Hanover, Hanover, Germany*: Jennifer Freymann.

Trial Registration

Association of Trial Registration With Reporting of Clinical Trials: Comparison of Protocols, Registries, and Published Articles

An-Wen Chan,[1,2] Annukka Pello,[3] Jessica Kitchen,[1] Anna Axentiev,[1] Jorma Virtanen,[4] Annie Liu,[1] Elina Hemminki[5]

Objective To evaluate adherence to trial registration and its association with subsequent publication and selective reporting of primary outcomes in an unselected cohort of clinical trials.

Design This was an inception cohort study of all initiated clinical trial protocols approved in 2002 (n=135) and 2007 (n=113) by the research ethics committee for the region of Helsinki and Uusimaa, Finland. We identified registry records and articles published up to February 2017 using keywords to search trial registries, PubMed, EMBASE, and Google. Trial characteristics (approval year, funding, sample size, intervention type, number of arms and centers) and outcomes were abstracted from each protocol, registry record, and publication. Using descriptive statistics and multivariable logistic regression, we determined the rates and predictors of registration and publication; the proportion of trials with discrepant primary outcomes in the protocol compared with the registry and publication; and the association between registration and subsequent publication without discrepant primary outcomes. Discrepancies were defined as (1) a new primary outcome being reported that was not specified as primary in the protocol; or (2) a protocol-defined primary outcome being omitted or downgraded (reported as secondary or unspecified) in the registry or published article.

Results Registration rates increased from 0% (0 of 135) for trials approved in 2002 to 61% (69 of 113) in 2007. Overall, 130 of 248 of all trials (52%) were published (publication years 2003 through 2016); 16 of 69 registered trials (23%) had discrepancies in primary outcomes defined in the registry compared with the protocol, while 24 of 116 published trials (21%) had discrepancies in primary outcomes between the published article and the protocol. Among trials approved in 2007, trial registration was significantly associated with subsequent publication (68% of registered trials vs 39% of unregistered trials; adjusted odds ratio [aOR], 4.5; 95% CI, 1.1-18). Registered trials were also significantly more likely than unregistered trials to be subsequently published with the same primary outcomes defined in the published article compared with the protocol (64% vs 25%; aOR, 5.8; 95% CI, 1.4-24).

Conclusions Clinical trials are not only often unregistered and unpublished but also discrepant in the reporting of primary outcomes across different information sources. These major deficiencies impair transparency and facilitate the biased reporting of trial results, which can be mitigated through adherence to trial registration. Journal editors, legislators, funding agencies, regulators, research ethics committees, and sponsors should implement and enforce policies mandating registration and public access to full protocols for all clinical trials.

[1]Women's College Research Institute, Women's College Hospital, Toronto, Ontario, Canada, anwen.chan@utoronto.ca; [2]Department of Medicine, University of Toronto, Ontario, Canada; [3]Faculty of Medicine, University of Helsinki, Helsinki, Finland; [4]Faculty of Medicine, University of Oulu, Oulu, Finland; [5]THL (National Institute for Health and Welfare), Helsinki, Finland

Conflict of Interest Disclosures: None reported.

Funding/Support: This project was supported by the Canadian Institutes of Health Research Dissemination Events (grants MET 117434 and MET 133851) and the Academy of Finland (grant No. 28356).

Role of the Funder/Sponsor: The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the abstract.

Impact of FDAAA on Registration, Results Reporting, and Publication of Neuropsychiatric Clinical Trials Supporting FDA New Drug Approval, 2005-2014

Constance X. Zou,[1] Jessica E. Becker,[2,3,4] Adam T. Phillips,[5] Harlan M. Krumholz,[6,7,8,9] Jennifer E. Miller,[10] Joseph S. Ross[7,8,9]

Objective Selective publication and reporting of clinical trial results undermines evidence-based medicine. The 2007 Food and Drug Administration Amendments Act (FDAAA) mandates, with few exceptions, the registration and reporting of results of all non–phase I clinical trials on ClinicalTrials.gov for approved products. The objective of this study was to determine whether efficacy trials supporting US Food and Drug Administration (FDA) approval of new drugs used for neurological and psychiatric conditions that were completed after FDAAA was enacted were more likely to have been registered, have their results reported, and be published in journals than those completed pre-FDAAA.

Design We conducted a retrospective observational study of efficacy trials reviewed by the FDA as part of any new neuropsychiatric drugs approved between 2005 and 2014. In January 2017, for each trial, we searched ClinicalTrials.gov for the registration record and for reported results, and we searched MEDLINE-indexed journals using PubMed for corresponding publications. In addition, published findings were validated against FDA interpretations described in regulatory medical review documents. Trials were considered FDAAA applicable if they were initiated after September 27, 2007, or were still ongoing as of December 26, 2007. The rates of trial registration, results reporting, publication, and publication-FDA agreement were compared between pre-FDAAA and post-FDAAA trials using Fisher exact test.

Results Between 2005 and 2014, the FDA approved 37 new neuropsychiatric drugs on the basis of 142 efficacy trials, of which 41 were FDAAA applicable. Post-FDAAA trials were significantly more likely to be registered (100% vs 64%; $P < .001$) and to report results (100% vs 10%; $P < .001$) than

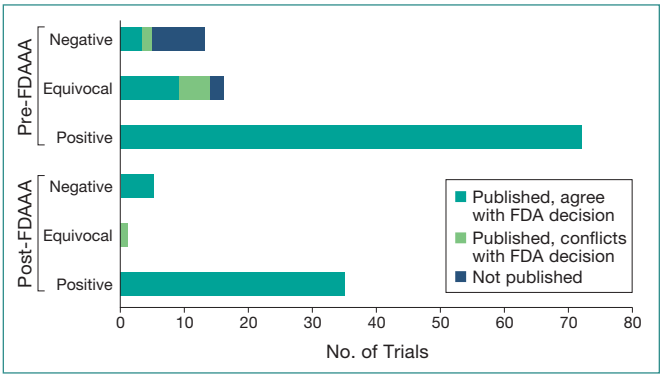
pre-FDAAA trials, but post-FDAAA trials were not significantly more likely to have been published (100% vs 90%; *P* = .06) nor to have been published with findings in agreement with the FDA's interpretation (98% vs 93%; *P* = .28) (**Figure 4**). Subgroup analyses suggest that the changes in overall publication rate were primarily the consequence of publishing negative trials, as all pre-FDAAA and post-FDAAA positive trials were published (72 of 72 and 35 of 35, respectively), whereas 38% (5 of 13) of pre-FDAAA negative trials were published vs 100% (5 of 5) of post-FDAAA negative trials.

Conclusions After FDAAA was enacted, all efficacy trials reviewed by the FDA as part of new drug applications for neuropsychiatric drugs were registered, with the results reported and published. Moreover, nearly all were published with interpretations that agreed with the FDA's interpretation. While our study was limited by searching for registration status only on ClinicalTrials.gov, our findings suggest that by mitigating selective publication and reporting of clinical trial results, FDAAA improved the availability of evidence for physicians and patients to make informed decisions regarding the care of neuropsychiatric illnesses.

[1]Yale School of Medicine, New Haven, CT, USA, constance.zou@yale.edu; [2]Department of Psychiatry, Massachusetts General Hospital, Boston, MA, USA; [3]McLean Hospital, Belmont, MA, USA; [4]Harvard Medical School, Boston, MA, USA; [5]Department of Internal Medicine, Yale School of Medicine, New Haven, CT, USA; [6]Section of Cardiovascular Medicine, Department of Internal Medicine, Yale School of Medicine, New Haven, CT, USA; [7]Center for Outcomes Research and Evaluation, Yale-New Haven Hospital, New Haven, CT, USA; [8]Robert Wood Johnson Foundation Clinical Scholars Program, Department of Internal Medicine, Yale School of Medicine, New Haven, CT, USA; [9]Department of Health Policy and Management, Yale School of Public Health, New Haven, CT, USA; [10]Division of Medical Ethics, Department of Population Health, NYU School of Medicine, Bioethics International, New York, NY, USA

Conflict of Interest Disclosures: Constance Zou has received a fellowship through the Yale School of Medicine from the National Heart, Lung, and Blood Institute. Jennifer Miller has received research support through New York University from the Laura and John Arnold Foundation to support the Good Pharma Scorecard. Harlan Krumholz and Joseph Ross have received research support through Yale University from Johnson and Johnson to develop methods of clinical trial data sharing, from Medtronic and the US Food and Drug Administration (FDA) to develop methods for postmarket surveillance of medical devices, and from the US Centers of Medicare and Medicaid Services to develop and maintain performance measures that are used for public reporting. Harlan Krumholz also has received compensation as a member of the Scientific Advisory Board for United Healthcare. Joseph Ross has received research support through Yale University from the FDA to establish a Center for Excellence in Regulatory Science and Innovation at Yale University and the Mayo Clinic, from the Blue Cross Blue Shield Association to better understand medical technology evaluation, and from the Laura and John Arnold Foundation to support the Collaboration on Research Integrity and Transparency at Yale University.

Figure 4. Pre- and Post-FDAAA Efficacy Trials Supporting Neuropsychiatric Drugs First Approved Between 2005 and 2014: Publication Status and Published Conclusion Concordance With FDA Decision



FDA indicates the US Food and Drug Administration; FDAAA, the 2007 Food and Drug Administration Amendments Act.

Evaluation of the ClinicalTrials.gov Results Database and Its Relationship to the Peer-Reviewed Literature

Deborah A. Zarin,[1] Tony Tse,[1] Rebecca J. Williams,[1] Thiyagu Rajakannan,[1] Kevin M. Fain[1]

Objective As of February 22, 2017, ClinicalTrials.gov contained summary results for 24,377 studies and received 160 new submissions weekly. We estimate that US academic medical centers are required to report more than half of their sponsored trials to ClinicalTrials.gov under federal policies. We previously estimated that one-half of registered studies with results posted on ClinicalTrials.gov lacked results publications. It is critical to continue assessing the degree to which this database meets its intended goals. The objective of this study was to assess the potential scientific impact of the ClinicalTrials.gov results database using our 2013 evaluation framework.

Design We analyzed 2 samples of ClinicalTrials.gov results data to assess the impact on the available evidence base.

Results On February 10, 2017, 10,464 of 24,251 posted results (43%) had links to PubMed. Because not all publications are automatically linked and not all linked publications report results, we manually examined a random sample of 100 sets of posted results listing study completion dates in 2014. Of these, 28 had at least 1 results publication prior to results posting, 15 had a results publication after results posting, and we could not identify results publications for 57 studies. We also identified examples of how publications leveraged the information on ClinicalTrials.gov. To further examine the potential impact on selective publication, we evaluated drug-condition-sponsor "families." We identified 329 registered, industry-funded, phase 2 through 4, US trials completed or terminated from 2007 through 2009, representing 88 drugs and 96 unique drug-condition-sponsor families (eg, Amgen-sponsored trials of alendronate for osteoporosis). Ideally, summary results for all trials in all families would be publicly available. As of December 1, 2014, of 329 trials, 109 (33%) had results posted on ClinicalTrials.gov only, 42 (13%) available from PubMed only, 81 (25%) available from both, and 97 (29%) in neither (**Table 12**). Overall, 45 of the 96 drug-condition-sponsor families had results available for all 144 trials from at least 1

Table 12. Study Design Characteristics of 329 ClinicalTrials.gov-Registered Trials in the Drug-Condition-Sponsor "Families" Sample by Results Dissemination Category as of April 27, 2017

Study Design Characteristic	Trials by Dissemination of Results, No. (% by Characteristic)				
	Results Disclosure			Any vs No Results Disclosure on ClinicalTrials.gov and PubMed	
	ClinicalTrials.gov Only (n = 109)	PubMed Only (n = 42)	Both ClinicalTrials.gov and PubMed (n = 81)	Results Disclosure Total (n = 232)	Neither ClinicalTrials.gov nor PubMed (n = 97)
Interventional model					
Parallel assignment (n = 242)	79 (33)	33 (14)	68 (25)	180 (74)	62 (26)
Single group assignment (n = 72)	26 (36)	7 (10)	8 (11)	41 (57)	31 (43)
Crossover assignment (n = 9)	2 (22)	1 (11)	4 (44)	7 (78)	2 (22)
Factorial assignment (n = 4)	2 (50)	0	1 (25)	3 (75)	1 (25)
Missing data (n = 1)	0	1 (50)	0	0 (50)	1 (50)
Masking					
Open label (n = 99)	37 (37)	9 (9)	18 (18)	64 (65)	35 (35)
Double blind (n = 209)	62 (60)	32 (15)	57 (27)	151 (72)	58 (28)
Single blind (n = 11)	6 (55)	0	2 (18)	8 (73)	3 (27)
Missing data (n = 10)	4 (40)	1 (10)	4 (40)	9 (90)	1 (10)
Allocation					
Randomized (n = 246)	75 (30)	34 (14)	70 (28)	179 (73)	67 (27)
Missing data (n = 30)	17 (57)	4 (13)	1 (3)	22 (73)	8 (27)
Nonrandomized (n = 53)	17 (32)	4 (8)	10 (19)	31 (58)	22 (42)
No. of sites					
Multiple (n = 218)	70 (32)	29 (13)	62 (28)	161 (74)	57 (26)
Single (n = 58)	19 (33)	6 (10)	3 (5)	28 (48)	30 (52)
Missing data (n = 53)	20 (38)	7 (13)	16 (30)	43 (81)	10 (19)
No. of participants enrolled					
1-100 (n = 102)	37 (36)	13 (13)	11 (11)	61 (60)	41 (40)
101-500 (n = 150)	52 (35)	21 (14)	32 (21)	105 (70)	45 (30)
>500 (n = 75)	20 (27)	8 (11)	38 (51)	66 (88)	9 (12)
Missing data (n = 2)	0	0	0	0	2 (100)

source, 18 families involving a total of 48 trials had no results available, and 15 families had results disclosed on ClinicalTrials.gov only.

Conclusions Between 33% (109 of 329) and 57% (57 of 100) of completed or terminated ClinicalTrials.gov-registered trials have posted results but no corresponding PubMed-cited results articles. These findings suggest that ClinicalTrials.gov provides a unique source of results for substantial numbers of trials.

[1]ClinicalTrials.gov, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA, dzarin@nih.gov

Conflict of Interest Disclosures: All authors work for ClinicalTrials.gov as full-time employees of the National Library of Medicine. Dr Zarin is a member of the Peer Review Congress Advisory Board but was not involved in the review or decision for this abstract.

Funding/Support: This work was supported by the Intramural Research Program of the National Library of Medicine, National Institutes of Health. The National Library of Medicine has approved this submission.

Funding/Grant Review

Geographic and Gender Bias in Peer Review of Applications Submitted to the Swiss National Science Foundation

João Martins,[1] François Delavy,[1] Anne Jorstad,[1] Matthias Egger[1]

Objective The Swiss National Science Foundation (SNSF), the leading public research funder in Switzerland, relies on external experts to review grant applications. Applicants can propose reviewers, provided there are no obvious conflicts of interests. On average, applications receive 3 reviews, 1 of which is typically from a reviewer proposed by the applicants. We examined whether the source of the review, the gender of the principle applicant and the reviewer, and the country of affiliation of reviewers influenced the scores given to grant applications submitted to the SNSF.

Design Reviewers scored applications from 1 (poor) to 6 (outstanding). We calculated mean scores by source of

reviewers (applicant vs SNSF), country of affiliation of reviewers (Switzerland vs international), and gender of applicants and reviewers. We fit a multivariable linear regression model adjusting for all these variables plus calendar year of submission, discipline (21 disciplines), and applicants' age (5 age classes) and affiliation (4 institution types).

Results Between 2009 and 2015, 36,993 reviewers assessed 12,132 applications for the SNSF. The mean (SD) score of reviewers proposed by applicants (n=8308) was 5.12 (1.01) vs 4.47 (1.25) for reviewers proposed by the SNSF (n=26,594). Mean (SD) scores were 4.19 (1.27) for Swiss experts (n=8399) vs 4.76 (1.19) for international experts (n=26,503); 4.44 (1.25) for female (n=7121) vs 4.67 (1.22) for male (n=27,781) principle applicants; and 4.48 (1.26) for reviews from female (n=6933) vs 4.66 (1.22) from male (n=27,969) reviewers. In adjusted analyses, the gender differences were attenuated, whereas the other differences changed little (**Table 13**). All differences were statistically significant.

Conclusions Applications received higher scores from applicant-proposed reviewers and lower scores from Swiss-based experts. Scores were lower for applications submitted by female applicants. Our results are compatible with a positive bias of reviewers chosen by the applicant, or a negative bias of experts based in Switzerland, and cannot exclude bias against female applicants. Interestingly, female reviewers consistently scored applications lower than male reviewers, independent of the applicant's gender. Panels making funding decisions should be aware of these potential biases. Given the association between scores and source of reviewer, the SNSF no longer accepts reviewers proposed by the applicants.

[1]Swiss National Science Foundation, Bern, Switzerland, joao.martins@snf.ch

Conflict of Interest Disclosures: The authors are employees of the Swiss National Science Foundation.

Table 13. Unadjusted and Adjusted Differences in Scores Assigned by Reviewers of Grant Applications Submitted to the Swiss National Science Foundation

Variable	Difference (95% CI)[a]	
	Unadjusted	Adjusted
Source of reviewer		
Applicant vs SNSF	0.65 (0.62 to 0.68)	0.52 (0.49 to 0.55)
Affiliation of reviewer		
Switzerland vs international	−0.56 (−0.59 to −0.53)	−0.50 (−0.53 to −0.47)
Gender of applicant		
Female vs male	−0.23 (−0.19 to −0.26)	−0.09 (−0.06 to −0.12)
Gender of reviewer		
Female vs male	−0.17 (−0.13 to −0.21)	−0.07 (−0.03 to −0.10)

[a]All unadjusted *P* values from *t* tests <.001. Adjusted results from a linear regression model adjusted for calendar year of submission, discipline, and applicants' age and affiliation; all adjusted *P* values <.001.

Stakeholder Perceptions of Peer Review at the National Institutes of Health Center for Scientific Review

Mary Ann Guadagno,[1] Richard K. Nakamura[1]

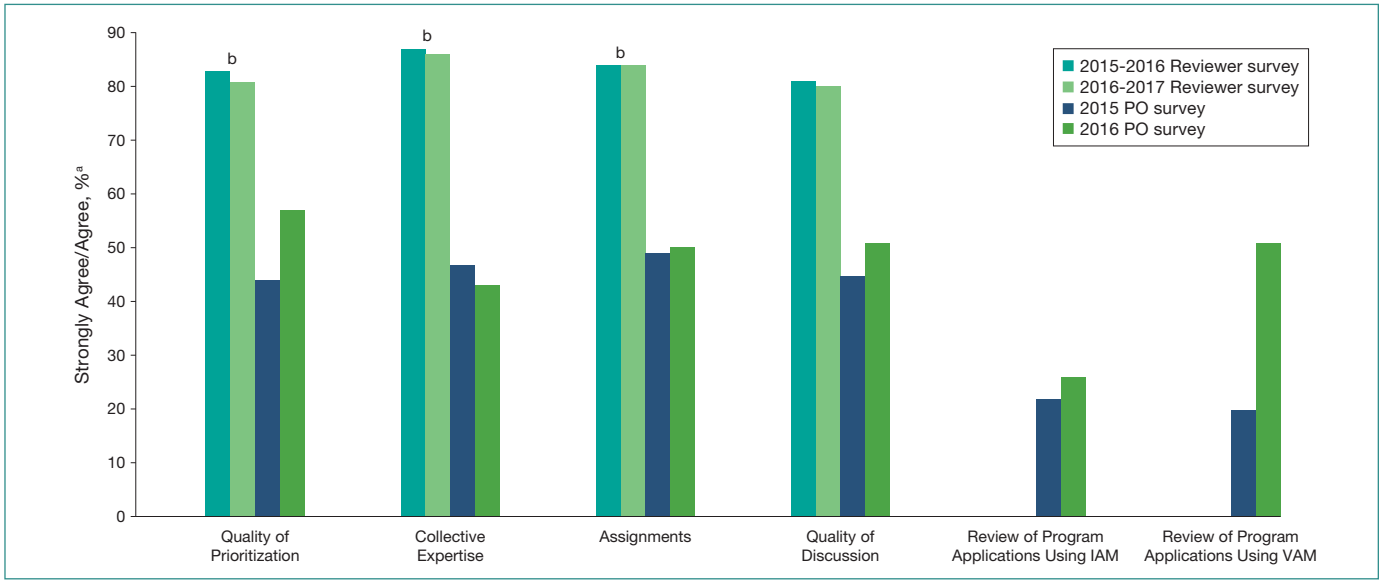
Objective To identify best practices for the successful peer review of grant applications and areas for improvement at the National Institutes of Health (NIH) Center for Scientific Review (CSR), the following questions guided an evaluation study: (1) to what extent are current CSR practices for peer review optimal for achieving its mission? and (2) what are the areas of success and improvement in the quality of peer review?

Design Pilot assessments were conducted to develop a short "Quick Feedback" survey instrument with four 7-point Likert-type scale statements ranging from "strongly agree" to "strongly disagree," measuring key features of peer review, and an open text box for comments. During 1 grant cycle between 2015-2016 and 2016-2017, 2 surveys were sent to 10,262 and 10,228 reviewers, respectively, in all CSR study sections. In 2015, a survey was sent to 916 NIH Program Officers (POs), and a replication survey was sent to POs in 2016 to 905 POs. During 2015, 27 focus groups were conducted with 4 stakeholder groups, and 10 personal interviews were completed with NIH Institute Directors. Focus group participants were selected from NIH databases to ensure diversity. Interrater reliability between coders was 95.8%.

Results The 2015-2016 reviewer survey yielded a response rate of 47.1% (4832 of 10,262), and the 2016-2017 reviewer survey yielded a response rate of 47.0% (4807 of 10,228). The 2015 PO survey had a response rate of 38.0% (348 of 916), and the 2016 replication PO survey yielded a response rate of 37.0% (335 of 905). Nonrespondents were not substantially different from respondents. "Quick Feedback" surveys with reviewers in both years reported a high level of satisfaction with the peer review process. More than 80% of reviewers indicated they either "strongly agreed" or "agreed" that panels were doing a good job in terms of scoring and discussion and CSR did a good job relative to the quality of the rosters and assignments (**Figure 5**). Program Officers were less favorable than reviewers in both years, with only 43% to 57% of POs responding favorably. Program Officers' dissatisfaction with review meetings focused on insufficient reviewer expertise in general and technical and logistical challenges at meetings more specifically. Focus group results supported these findings. Areas for improvement included reducing the burden of peer review for all stakeholders, technical and logistical issues during meetings, need for clearer communication, and more guidance on preparing applications.

Conclusions A comprehensive evaluation using systematic surveys, focus groups, and interviews has resulted in useful suggestions for improving best practices for peer review by stakeholders in real time. Areas of success and suggestions for

Figure 5. Overall Center for Scientific Review Quick Feedback Favorable Responses



The 2015-2016 reviewer survey yielded a response rate of 47.1% (4832 of 10,262), and the 2016-2017 reviewer survey yielded a response rate of 47.0% (4807 of 10,228). The 2015 Program Officer (PO) survey had a response rate of 38.0% (348 of 916), and the 2016 replication PO survey had a response rate of 37.0% (335 of 905). IAM indicates Internet-assisted meeting; VAM, video-assisted meeting.
[a]Strongly agree or agree refers to a 1 or 2, respectively, as assessed on a 7-point Likert-type scale.
[b]IAM reviewers not included in 2016.

improvements by stakeholders are being addressed by leadership.

[1]Center for Scientific Review, National Institutes of Health, Bethesda, MD, USA, guadagma@csr.nih.gov

Conflict of Interest Disclosures: Both authors are federal employees at the National Institutes of Health. Survey research was conducted as part of their federal employment responsibilities.

Funding/Support: Focus groups and personal interviews were funded by the NIH Evaluation Set-Aside Program (14-5725 CSR), administered by the Office of Program Evaluation and Performance of the National Institutes of Health.

Role of the Funder/Sponsor: The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the abstract.

Testing of 2 Application Ranking Approaches at the National Institutes of Health Center for Scientific Review

Richard K. Nakamura,[1] Amy L. Rubinstein,[1] Adrian P. Vancea,[1] Mary Ann Guadagno[1]

Objective The National Institutes of Health (NIH) is a US agency that distributes approximately $20 billion each year for research awards based on a rigorous peer review that provides a merit score for each application. Final scores are based on the mean of scores from reviewers and then ranked via percentile. In 2009, the NIH changed its scoring system from a 40-point scale to a 9-point scale. There have been concerns that this new scale, which is functionally cut in half for the 50% of applications that are considered competitive, is not sufficient to express a study section's judgment of relative merit. The question guiding these pilot studies was whether alternative methods of prioritizing applications could reduce the number of tied scores or increase ranking dispersal.

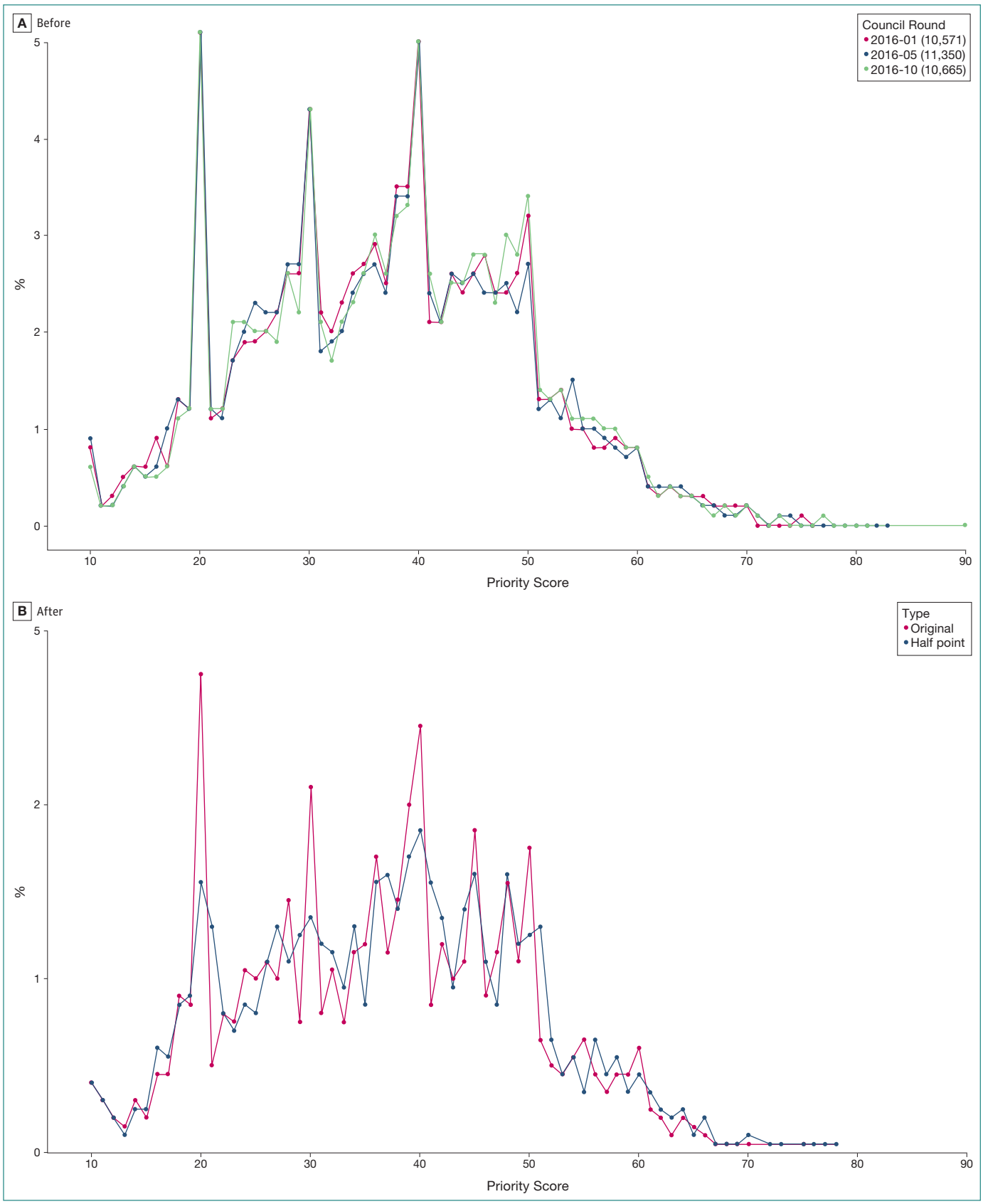
Design The Center for Scientific Review has been testing alternate scoring systems, including (A) postmeeting ranking of the top scoring applications, in which reviewers rank-order the 10 best applications at the end of a review, and (B) giving reviewers the option of adding or subtracting a half point during final scoring of applications following discussion. These alternatives were compared against standard scoring in real study sections to see if they could improve prioritization of applications and reduce the number of tied scores. Reviewer opinions of the ranking systems were assessed, including surveys for alternate B.

Results (A) Postmeeting ranking of applications were applied to 836 applications across 32 study sections; these often produced rankings inconsistent with scores given to applications by reviewers. The best 2 or 3 scored applications were generally agreed on, but increased disagreement among reviewers was observed with poorer average scores. Most reviewers liked the ranking system, but there was more hesitation about recommending adoption of this practice over the current scoring system. (B) Making available a half point to add or subtract freely in final voting was applied to 1371 applications across 39 study sections; the half-point system helped to spread scores and halved the number of ties (**Figure 6**). It was also recommended for adoption by 72% of reviewers in postmeeting surveys.

Conclusions Initial results of the half-point scoring system have been interpreted favorably. The Center for Scientific Review will conduct a full test of the half-point scoring system under real review conditions.

[1]Center for Scientific Review, National Institutes of Health, Bethesda, MD, USA, rnakamur@mail.nih.gov

Figure 6. Distribution of National Institutes of Health Grant Application Scores by Percent Before and After Use of the Half-Point Option



A, Distribution of final scores for grant applications as a percent of all scores (of 32,586 applications). Each application received scores from many reviewers that were multiplied by 10 and averaged to the nearest unit. Possible final scores for each application ranged from 10 to 90. Dates refer to the cycle of review and the number is the quantity of applications with scores. In January 2016, 10,571 applications received scores; in May 2016, 11,350 applications received scores; and in October 2016, 10,665 applications received scores. B, Comparison of distribution of original average scores (red) with scores for which reviewers were allowed to add or subtract a half point (blue). Average scores are rounded to the nearest digit to establish ranking. Original refers to the proportion of scores at each possible score level under normal whole digit scoring. Half point refers to the proportion of scores at each possible score level when reviewers used whole digits plus or minus 1 half point.

Conflict of Interest Disclosures: All authors are employees of the National Institutes of Health.

Funding/Support: Study funded by the National Institutes of Health.

Role of the Funder/Sponsor: The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the abstract.

Scientist, Patient, and Stakeholder Roles in Research Application Review: Analysis of the Patient-Centered Outcomes Research Institute (PCORI) Approach to Research Funding

Laura P. Forsythe,[1] Lori B. Frank,[1] A. Tsahai Tafari,[1] Sarah S. Cohen,[2] Michael Lauer,[1,3] Steve Clauser,[1] Christine Goertz,[1,4] and Suzanne Schrandt[1]

Objective Scientific review of funding applications was established to fund rigorous, high-impact research. The Patient-Centered Outcomes Research Institute (PCORI) uses unique review criteria and includes patients and other healthcare stakeholders as reviewers. This study assesses the relative importance of each criterion and the associations of different reviewer types' ratings with final scores and funding outcomes.

Design This study is a retrospective, cross-sectional analysis of PCORI Merit Review administrative data for 5 funding cycles from 2013 through 2015. Before a panel discussion, patients and other stakeholders were required to score each application overall and on 3 criteria: potential to improve care and outcomes, patient-centeredness, and engagement. Scientist reviewers also scored impact of condition and technical merit. Scores ranged from 1 (exceptional) to 9 (poor). All reviewers provided postdiscussion overall scores. Funding decisions were made by the PCORI Board of Governors based on Merit Review, portfolio balance, and programmatic fit. Linear regression models stratified by reviewer type (ie, scientist, patient, or other stakeholder) tested associations of postdiscussion overall scores with prediscussion criteria scores. Associations between funding decisions and prediscussion criteria scores were tested using logistic regression. All models adjusted for funding program, review cycle, and principal investigator characteristics (ie, National Institutes of Health funding, clinical degree[s] of applicants, and years of experience of applicants).

Results A total of 535 reviewers (254 scientists, 139 patients, and 142 stakeholders) reviewed 1312 applications; 663 (50.5%) were discussed and 121 (9.2%) were funded. Prediscussion mean (SD) overall scores were higher (ie, worse) for scientist reviewers (4.9 [2.1]) than patient reviewers (4.2 [2.2]) and stakeholder reviewers (4.2 [2.1]) ($P < .001$). The mean overall score postdiscussion was 28.0 for funded applications and 50.1 for unfunded applications. All reviewer types changed their overall score through panel discussion for more than half of the applications. Score agreement across reviewer types was greater postdiscussion. For all reviewer types, postdiscussion review scores were positively associated with at least 1 prediscussion criterion score from each of the 3 reviewer types (**Table 14**). The strongest association with postdiscussion overall scores for all reviewer types was scientists' ratings of technical merit. More favorable prediscussion ratings by each reviewer type for the potential to improve care and outcomes and scientist reviewers' ratings of technical merit and patient-centeredness were associated with greater likelihood of funding.

Conclusions Scientist, patient, and stakeholder views of applications converged following discussion. Technical merit is critical to funding success, but patient and stakeholder ratings of other criteria also relate to application disposition. Results suggest that research application review can incorporate nonscientist perspectives in scoring and funding outcomes.

[1]Patient-Centered Outcomes Research Institute, Washington, DC, USA, lforsythe@pcori.org; [2]EpidStat Institute, Ann Arbor, MI, USA; [3]National Institutes of Health, Bethesda, MD, USA; [4]Palmer Center for Chiropractic Research, Davenport, IA, USA

Conflict of Interest Disclosures: None reported.

Funding/Support: This work was funded by the Patient-Centered Outcomes Research Institute (PCORI).

Role of the Funder/Sponsor: Members of the PCORI staff, Board of Governors, and Methodology Committee designed and conducted the study and reported the results.