Draft Methods Report

Number XX

Selective Outcome Reporting as a Source of Bias in Reviews of Comparative Effectiveness

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
www.ahrq.gov

This information is distributed solely for the purposes of predissemination peer review. It has not been formally disseminated by the Agency for Healthcare Research and Quality. It does not represent and should not be construed to represent an Agency for Healthcare Research and Quality or Department of Health and Human Services (AHRQ) determination or policy.

Contract No.

Prepared by:

Investigators

AHRQ Publication No. xx-EHCxxx-EF
<Month Year>

Statement of Funding and Purpose

This report is based on research conducted by an Evidence-based Practice Center (EPC) under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No. xxx-xxxx-xxxxx). The findings and conclusions in this document are those of the authors, who are responsible for its contents; the findings and conclusions do not necessarily represent the views of AHRQ. Therefore, no statement in this report should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services.

The information in this report is intended to help health care decision makers—patients and clinicians, health system leaders, and policymakers, among others—make well-informed decisions and thereby improve the quality of health care services. This report is not intended to be a substitute for the application of clinical judgment. Anyone who makes decisions concerning the provision of clinical care should consider this report in the same way as any medical reference and in conjunction with all other pertinent information, i.e., in the context of available resources and circumstances presented by individual patients.

This report may be used, in whole or in part, as the basis for development of clinical practice guidelines and other quality enhancement tools, or as a basis for reimbursement and coverage policies. AHRQ or U.S. Department of Health and Human Services endorsement of such derivative products may not be stated or implied.

Public Domain Notice

This document is in the public domain and may be used and reprinted without permission except those copyrighted materials noted for which further reproduction is prohibited without the specific permission of copyright holders.

Disclaimer Regarding 508-Compliance

Persons using assistive technology may not be able to fully access information in this report. For assistance, contact info@ahrq.gov.

Financial Disclosure Statement

None of the investigators has any affiliations or financial involvement that conflicts with the material presented in this report.

Suggested citation: Author I, Author I. Comparative Effectiveness of Title. Comparative Effectiveness Review No. xx. (Prepared by the xxx Evidence-based Practice Center under Contract No. xxx-xxxxx-xxxx.) AHRQ Publication No. xx-EHCxxx. Rockville, MD: Agency for Healthcare Research and Quality. <Month Year>. Available at: www.effectivehealthcare.ahrq.gov/reports/final.cfm.

Preface

The Agency for Healthcare Research and Quality (AHRQ), through its Evidence-based Practice Centers (EPCs), sponsors the development of evidence reports and technology assessments to assist public- and private-sector organizations in their efforts to improve the quality of health care in the United States. The reports and assessments provide organizations with comprehensive, science-based information on common, costly medical conditions and new health care technologies and strategies. The EPCs systematically review the relevant scientific literature on topics assigned to them by AHRQ and conduct additional analyses when appropriate prior to developing their reports and assessments.

To improve the scientific rigor of these evidence reports, AHRQ supports empiric research by the EPCs to help understand or improve complex methodologic issues in systematic reviews. These methods research projects are intended to contribute to the research base in and be used to improve the science of systematic reviews. They are not intended to be guidance to the EPC program, although may be considered by EPCs along with other scientific research when determining EPC program methods guidance.

AHRQ expects that the EPC evidence reports and technology assessments will inform individual health plans, providers, and purchasers as well as the health care system as a whole by providing important information to help improve health care quality. The reports undergo peer review prior to their release as a final report.

We welcome comments on this Methods Research Project. They may be sent by mail to the Task Order Officer named below at: Agency for Healthcare Research and Quality, 540 Gaither Road, Rockville, MD 20850, or by e-mail to epc@ahrq.hhs.gov.

Carolyn M. Clancy, MD
Director
Agency for Healthcare Research and Quality

Jean Slutsky, PA, MSPH
Director, Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Stephanie Chang, MD, MPH
Director
Evidence-based Practice Program
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Mary Nix, MS, MT(ASCP)SBB
Task Order Officer
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Acknowledgments

Peer Reviewers

Selective Outcome Reporting as a Source of Bias in Reviews of Comparative Effectiveness

Structured Abstract

Objectives: The objectives of this exploratory study were to: 1) describe the frequency of selective outcome reporting (SOR) and selective analysis reporting (SAR) within primary studies included in reviews of comparative effectiveness for outcomes of benefit; 2) explore potential predictors of the proportion of randomized controlled trials (RCTs) with SOR and SAR; and 3) assess the reliability and validity of the Outcome Reporting Bias in Trials (ORBIT) classification system for missing or incomplete outcome reporting and the ORBIT assessment of the risk of bias associated with each type category of SOR.

Data Sources and Methods: We identified comparative effectiveness reviews (CERs) funded by the Agency for Healthcare Quality and Research that included drug-drug comparisons, and selected three for further exploration. Within each CER we then specified one outcome that fulfilled explicit criteria (the "index outcome"). Randomized controlled trials (RCTs) examining the index outcome were then identified, and trial registries were searched for study registration and protocols. Focusing on the RCTs that were listed in a trial registry, we then used registry information, along with the methods section of the publication, to determine the frequency of SOR and SAR (Objective #1). For Objective #2, we examined prespecified predictors of the existence of SOR and SAR in RCTs. Lastly, using the ORBIT classification of SOR, we attempted to examine the inter-rater reliability of ORBIT and its validity, comparing SOR assessments using information contained within the publication (as ORBIT was intended) to assessments of SOR using the additional information in trial registries (Objective #3).

Results: Trials published in or after 2005 and contributing to meta-analyses were not consistently listed in trial registries. Publications of trials did not consistently report the fact of trial registration, and reporting of results in ClincialTrials.gov was inconsistent and the reasons could not be determined. We did not find the ORBIT classification tool (designed for the assessment of SOR within trial publications), particularly useful. ORBIT was difficult to implement; the nine classes were difficult to reliably distinguish. ORBIT classes did not describe the types of SOR and SAR that we most frequently encountered: the addition of outcomes measures, subgroups, and other analyses to published results that were not prespecified in the publication's methods section or listed in the trials registry. Trial registries were of little use in identifying SOR unless trial results were listed in the registry. Registries were of no use in identifying SAR.

Conclusions: Rates of SOR are high among RCTs in the three CERs that we examined, and most frequently outcomes are added to the results in publications that were not listed in the publications' methods section of specified in trial registries. Types of SOR that correspond to the ORBIT classification system were uncommon in our cohort. We were not able to identify SAR with existing tools. We identified numerous challenges in searching for, and characterizing, SOR and SAR among a cohort of RCTs that were included in three CERs. Existing tools were suboptimal: ORBIT does not encompass the most common type of potential selective reporting of outcomes and analyses. The presentation and content of ClinicalTrials.gov

could be improved to better assist the systematic reviewer in identifying potential SOR and SAR. Much further research is needed to develop efficient, tailored approaches to identifying and characterizing SOR and SAR in trials.

Contents

Background 1
Objectives 3
Methods 5
Objective 1: Frequency of SOR and SAR 5
Cohort of Comparative Effectiveness Reviews 5
Identification of an Index Outcome within Each Included CER 7
Identification of Trial Registration for RCTs Reporting the Index Outcome 8
Data Abstraction 12
Identifying SOR and SAR 12
Objective 2: Predictors of SOR and SAR 16
Objective 3: Assessment of the Reliability and Validity of the ORBIT Classification of SOR 18
Data Syntheses and Analyses 18
Results 20
Cohort of CERs and the choice of index outcome for each CER 20
Objective 1: Frequency of SOR and SAR 21
Objective 2: Predictors of SOR 33
Objective 3: Reliability and Validity of ORBIT 33
Discussion 34
Strengths and Limitations of this Study 34
Challenges Encountered Developing our Study Protocol 35
Challenges and Recommendations Regarding Methods and Available Tools for Exploring SOR and SAR 36
ORBIT 36
Challenges Using the ORBIT Classification 36
Limitations of the ORBIT Classification System 36
Recommendations for Future Research on the ORBIT Classification System 37
Trial Registries 38
Problems Encountered Using Trial Registries 38
Recommended Changes in ClinicalTrials.gov 39
Box 41
Suggestions for Identifying and Characterizing SOR and SAR 42
Suggestions for Systematic Reviewers 42
Conclusions 44
References 45
Abbreviations and Acronyms 47

Tables

Table 1. The Outcome Reporting Bias In Trials (ORBIT) study classification system for missing or incomplete outcome reporting in reports of randomized trials[2] 3
Table 2. Inclusion criteria for the comparative effectiveness reviews examined in this report..... 7
Table 3. Types of SOR and SAR classified by the mechanism used for detection 14
Table 4. Index outcome for each comparative effectiveness review 21
Table 5. Characteristics of trials that reported the index outcomes 24
Table 6. SOR and SAR for A1c and fracture outcomes based on the publication only (judgment #1) 25
Table 7. SOR and SAR for A1c and fracture outcomes based on the publication and on the registry (judgment #2 and #3) 26
Table 8. Potential predictors of SOR and SAR for A1c and fracture outcomes..... 27
Table 10. Timing of study registration 28
Table 12. Time (in minutes) to complete review of ClinicalTrials.gov..... 33

Figures

Figure 1. Overview of report methodology 6
Figure 2. Process for identifying an index outcome within each included CER 8
Figure 3. The process for identifying trial registration for each RCT of the index outcome 11
Figure 4. Judgments on the presence of selective outcome and analysis reporting..... 16
Figure 5. Example of study and registry chronology[30] (see text for details) 30
Figure 6. Example of study and registry chronology[31] (see text for details) 31
Figure 7. Example of study and registry chronology[32] (see text for details) 32

Appendixes

Appendix A. Tables and Figures

Background

This project set about to examine the frequency and effect of reporting biases, specifically selective outcome reporting (SOR) and selective analysis reporting (SAR), in reviews of comparative effectiveness. Existing data suggest that SOR and SAR may have an important impact on the reported effects and conclusions in randomized controlled trials.[1] The bias introduced by SOR and SAR in primary research impacts conclusions in systematic reviews, although data on this effect are sparse.[2]

Reporting biases arise when the dissemination of research findings is influenced by the nature and direction of the results,[3] and can arise from processes acting within a study or at the level of the whole study. Within studies, researchers may report their findings selectively – choosing to report selected outcomes and analyses based on the results. Reporting bias can thus result from selective outcome reporting (SOR), wherein only a subset of the original outcomes measured and analyzed in a study are fully reported based on the magnitude of the treatment effect or the statistical significance of selected outcomes.[4] Kirkham and colleagues describe three main types of SOR:[2] selective reporting of an entire study outcome (i.e., analyzed outcomes are not reported); selective reporting of a specific outcome (e.g., selected followup intervals), and incomplete reporting of a specific outcome (e.g., incomplete reporting of non-significant p-values, such as $p>0.05$). SOR can result in outcome reporting bias (ORB), which is the bias produced from selection for publication of a subset of the outcomes originally reported, on the basis of the results.[4]

Reporting bias arising from within-study processes can also result from the selection of analyses for reporting (SAR), which can lead to analysis reporting bias (ARB). Examples of SAR include selective reporting of data on subgroups, presentation of adjusted versus unadjusted analyses, selection of as-treated versus intention-to-treat analyses, selective approaches to the handling of missing data, choosing to analyze continuously measured variables categorically (outcomes or predictors in adjusted models), and choice of cut-point values to define categorical variables.[5]

ORB and ARB differ conceptually from other biases considered by review authors. An outcome measure judged to be at risk of selection, performance, detection, or attrition bias, may have a reported effect that deviates from the 'truth', and that biased estimate then becomes embedded in the systematic review incorporating that study. Selectively omitted or incompletely reported outcomes in a primary study judged to be at risk of ORB or ARB cannot be included in the review because the required information is not reported. In other words, the primary study transmits bias to the systematic review because of the absence of information.

In the design stage of a study, outcomes can be selected based on anticipated results, and these selected outcomes can be then specified in the study protocol. By definition this is not SOR as the selection of outcomes is not based on actual results; however this approach to design of primary studies can ultimately lead to biased results and conclusions in systematic reviews.

Publication bias, whereby an entire study is not published because of the nature or direction of the results,[6, 7] is also an important issue for systematic reviewers. Statistically significant results are more likely to be published: a systematic review of data from five inception cohorts of registered clinical trials found that studies with 'positive' findings were approximately twice as likely to be published as studies with 'negative' or 'null' results.[8] In addition, positive findings are more likely to be published rapidly,[9, 10] in English, more than once, in high impact journals, and to be cited by others.[11, 12] Publication bias is not, however, examined in this report where we

focus exclusively on the less well studied and recognized issues of within-study selective reporting, specifically SOR and SAR.

The high prevalence of SOR and SAR among primary studies is well documented. This research has been done almost exclusively in randomized, controlled trials (RCTs), usually by comparing study protocols submitted to regulatory or funding agencies with published outcomes.[1, 11, 13] In a systematic review of five such cohorts (four of which contained only RCTs), Dwan and colleagues[13] reported that changes in pre-specified outcomes occurred that were not documented in protocol amendments in 40 to 62 percent of studies, where there was at least one primary outcome that was changed, introduced, or omitted between the protocol and the publication. In addition, statistically significant outcomes had a higher odds of being fully reported compared to non-significant outcomes (range of odds ratios [OR], 2.2 to 4.7), suggesting ORB as well as SOR.

Data are sparse, however, on the prevalence of SOR and SAR among studies included in systematic reviews, and little is known about the effects of selective reporting on effect estimates and conclusions in such reviews. We are aware of only one study that has investigated potential bias in effect estimates in systematic reviews resulting from SOR in included primary studies. Kirkham and colleagues[2] compared effect estimates reported in meta-analyses to estimates obtained with sensitivity analyses estimating the same effects without SOR (using the maximum bias bound approach[14]) for a sample of new systematic reviews published in the Cochrane Library. Of 81 reviews with a single meta-analysis of the review primary outcome, 52 (64 percent) included one or more RCTs with a high suspicion of ORB. Of 25 reviews that could be assessed, the median percentage change in treatment effect between the reported effect and the estimated effect without SOR was 39 percent (interquartile range, 18 to 67 percent). Of 42 meta-analyses with statistically significant results, 19 percent became non-significant after adjustment for ORB and 26 percent overestimated the treatment effect by 20 percent or more. The maximum bias in a pooled effect is estimated using the assumption that only studies with a treatment effect above (or below) a threshold effect are published. The upper bound of the bias is estimated over all possible selection mechanisms (the possibility that study selection depends on study outcomes). The only assumption made is that the chance that a study is selected cannot get smaller as a study gets larger (has increased precision). This provides a worst case estimate of effect for a given range of missing studies.

Kirkham and colleagues[2] developed a classification system for SOR, called Outcome Reporting Bias in Trials (ORBIT) (Table 1). This nine-category assessment tool is based on information in the trial publication(s) only and not on other information such as that contained in trial registries. This system focuses on outcomes that are missing or incompletely reported in reports of RCTs, and differentiates types of selective outcome reporting based on the assessor's certainty about whether the outcome was measured and analyzed, and the potential reasons for missing data.

Table 1. The Outcome Reporting Bias In Trials (ORBIT) study classification system for missing or incomplete outcome reporting in reports of randomized trials[2]

	Description	Level of Reporting	Risk of Bias
	Clear that the outcome was measured and analyzed		
A	Trial report states that outcome was analyzed but only reports that result was not significant (typically stating p>0.05)	Partial	High risk
B	Trial report states that outcome was analyzed but only reports that result was significant (typically stating p<0.05)	Partial	No risk
C	Trial report states that outcome was analyzed but insufficient data were presented for the trial to be included in meta-analysis or to be considered to be fully tabulated	Partial	Low risk
D	Trial report states that outcome was analyzed but no results reported	None	High risk
	Clear that the outcome was measured		
E	Clear that outcome was measured but not necessarily analyzed. Judgment says likely to have been analyzed but not reported because of non-significant results	None	High risk
F	Clear that outcome was measured but not necessarily analyzed. Judgment says unlikely to have been analyzed but not reported because of non-significant results	None	Low risk
	Unclear whether the outcome was measured		
G	Not mentioned but clinical judgment says likely to have been measured and analyzed but not reported on the basis of non-significant results	None	High risk
H	Not mentioned but clinical judgment says unlikely to have been measured at all	None	Low risk
	Clear that the outcome was not measured		
I	Clear that outcome was not measured	NA	No risk

Abbreviations: NA, not applicable.
This table presents the categories of ORBIT that Kirkham and colleagues proposed[2], including their categorization of level of reporting their assessment of the risk for bias. Reproduced with permission.

Objectives

Systematic reviewers should assess the risk of all potential biases in primary studies included in their reviews. Given that there are emerging data suggesting the presence of SOR and SAR in both RCTs and in nonrandomized studies, systematic reviewers need to consider the potential existence of missing outcomes or analyses among the primary studies included in a review. In addition, review authors need to consider how SOR and SAR might affect the direction, magnitude, and precision of pooled effect estimates, as well as the conclusions about both benefits and harms in systematic reviews.

There are no data that we are aware of on the effects of SOR and SAR in reviews of comparative effectiveness and it is possible that SOR and SAR have different frequencies and implications across various types of systematic reviews, interventions, and outcomes. For example, the availability of protocols may vary among types of interventions and studies (e.g., effectiveness versus efficacy). In addition, some of the characteristics of comparative effectiveness research may affect the frequency and impact of SOR and SAR: CERs are more likely to include subjective measures of patient-important outcomes (e.g., symptoms, quality of life), head-to-head rather than placebo controlled studies, and heterogeneity of populations and interventions. Evidence on selective reporting across various study designs and outcomes may

assist in the interpretation of summary effect measures and conclusions in reviews of comparative effectiveness.

The objectives of this task order were therefore to:

1. describe the frequency of SOR and SAR within primary studies included in reviews of comparative effectiveness for outcomes of benefit;
2. explore potential predictors of SOR and SAR in RCTs; and
3. assess the reliability and validity of the Outcome Reporting Bias in Trials (ORBIT)[2] classification system for missing or incomplete outcome reporting and the ORBIT assessment of the risk of bias associated with each type category of SOR.

The ultimate goal of this work is to develop guidance for the systematic reviewer on how to efficiently identify SOR and SAR in primary studies and to understand the potential biases introduced into conclusions in systematic reviews, including reviews of comparative effectiveness (CERs). This project will generate additional evidence to inform assessments of the risk of bias and strength of evidence for an outcome across studies in a body of evidence on comparative effectiveness, and to assess the degree and direction of potential biases. These data will, in turn, inform specific guidance for systematic reviewers on when to look for SOR and SAR, how to most efficiently identify them, and how to quantify the resultant risk of bias and potential impact on conclusions.

Methods

Figure 1 provides an overview of the methods for Objectives #1 through #3. Among CERs funded by the Agency for Healthcare Quality and Research (AHRQ), we identified those that fulfilled our inclusion criteria. Within each CER we then specified one outcome that fulfilled explicit criteria (the "index" outcome). RCTs examining the index outcome were then identified, and trial registries were searched for study registration and protocols. Focusing on RCTs that were listed in a trial registry, we then used registry information, along with the methods section of the publication and judgment about what outcomes should have been reported in a given study, to determine the frequency of SOR and SAR (Objective #1). For Objective #2, we examined prespecified predictors of the existence of SOR and SAR in RCTs. Lastly, using the ORBIT classification of SOR,[2] we attempted to examine the inter-rater reliability of ORBIT and its validity, comparing SOR assessments using information contained within the publication (as ORBIT was intended) to assessments of SOR using the additional information in trial registries (Objective #3).

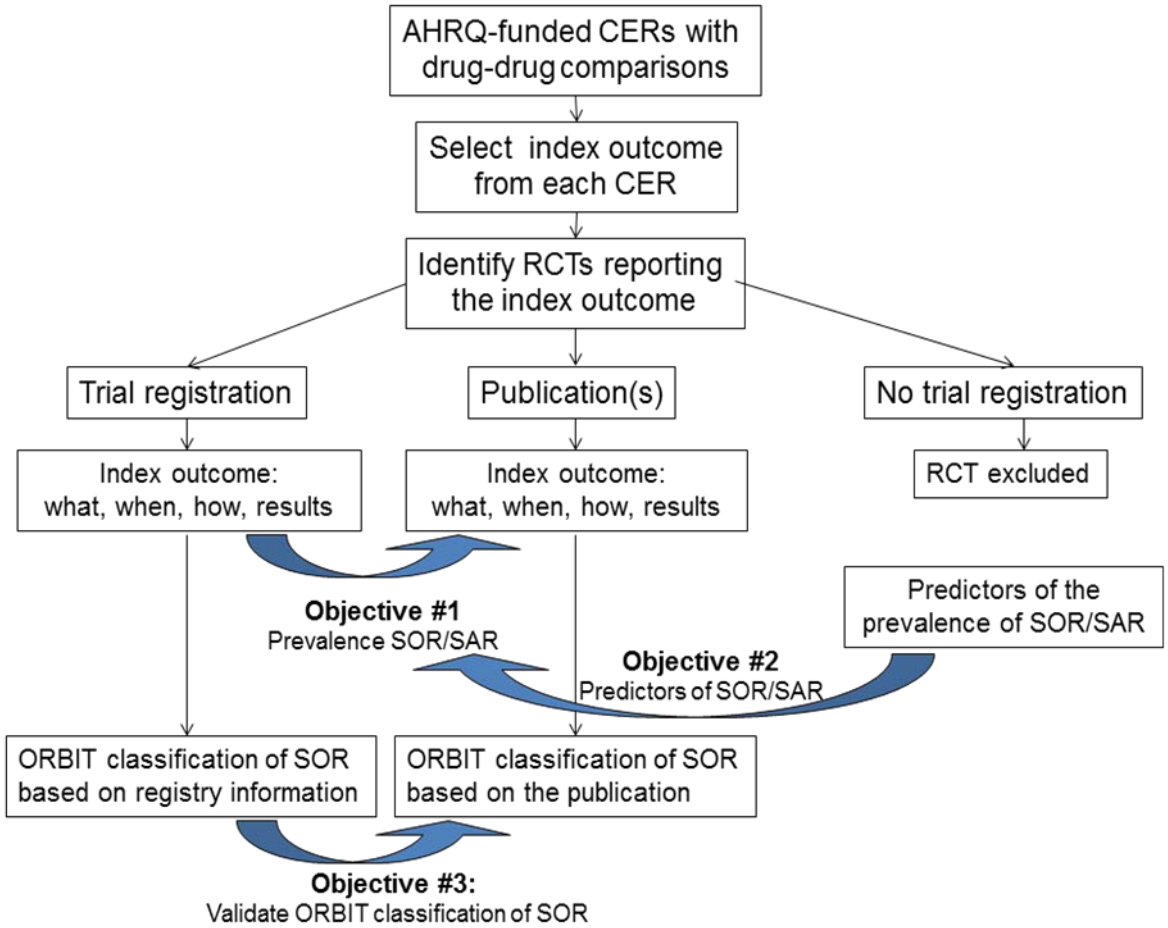
Objective 1: Frequency of SOR and SAR

Cohort of Comparative Effectiveness Reviews

To explore SOR and SAR, we selected a small cohort of CERs that fulfilled the criteria outlined in Table 2. We desired a cohort that was relevant to the AHRQ Effective Healthcare Program, was feasible to examine with the resources allocated to this project, would provide results applicable to future AHRQ CERs, would contribute to the existing methodological literature on SOR and SAR, and would facilitate the development of specific guidance on the detection and implications of SOR and SAR for systematic reviewers.

We piloted our methods using the report "Oral Diabetes Medications for Adults with Type 2 Diabetes. An Update" by Bennett and colleagues of the Johns Hopkins Evidence-based Practice Center (EPC).[15] We selected this report because: 1) it was published recently (March, 2011) and therefore likely used current AHRQ methods; 2) diabetes mellitus has a high burden of illness; and 3) this review included a large number of RCTs (n=104) reporting on an outcome that has relatively standardized measurement techniques (glycated hemoglobin, A1c). After developing and piloting our methods, we then examined two other AHRQ CERs focusing on drug-drug comparisons and with a large number of RCTs examining the index outcome that we had identified.[16, 17]

Figure 1. Overview of report methodology



Abbreviations: AHRQ, Agency for Healthcare Quality and Research; CER, comparative effectiveness review; O/C, outcome; RCTs, randomized controlled trials; SAR, selective analysis reporting; SOR, selective outcome reporting.

Figure 1. Overview of methods. Among CERs funded by the Agency for Healthcare Quality and Research (AHRQ), we identified those that fulfilled our inclusion criteria. Within each CER we then specified one outcome that fulfilled explicit criteria (the "index" outcome). RCTs examining the index outcome were then identified, and trial registries were searched for study registration and protocols. Focusing on RCTs that were listed in a trial registry, we then used registry information, along with the methods section of the publication and judgment about what outcomes should have been reported in a given study, to determine the frequency of SOR and SAR (Objective #1). For Objective #2, we examined prespecified predictors of the existence of SOR and SAR in RCTs. Lastly, using the ORBIT classification of SOR,[2] we attempted to examine the inter-rater reliability of ORBIT and its validity, comparing SOR assessments using information contained within the publication (as ORBIT was intended) to assessments of SOR using the additional information in trial registries (Objective #3).

Table 2. Inclusion criteria for the comparative effectiveness reviews examined in this report

General Criteria	Specific Criteria	Rationale	Number of Reports Fulfilling the Inclusion Criteria*
Report sponsor	AHRQ	Data gathered on AHRQ-funded reports are applicable to future such reports and to the development of guidance on methods for AHRQ systematic review programs in general.	NA
Type of review	Comparative effectiveness review	Reviews of comparative effectiveness are used for clinical and public health decision-making, thus understanding SOR and SAR for this type of systematic review is important. Technical briefs were excluded as they examine devices or diagnostic tests and include a very small number of RCTs.	26
Review status	Final	Final reports were examined in order to avoid reports that may be revised in the near future, and to include reports that have been reviewed and incorporated input from peer reviewers and the AHRQ Associate Editor.	25
Focus of the CER	The focus must be pharmacotherapy with drug-drug or drug-placebo comparators.	Methodology of drug reviews is more advanced in general than reviews of devices, procedures, prognosis, diagnosis, or epidemiology. RCTs are likely to be included, and some sources of clinical and methodological heterogeneity may be minimized by examining only drug interventions.	15
Design of included studies	The CER must contain RCTs.	Only RCTs were examined as we sought to identify study registration and studies with other designs are much less likely to be registered.	15
Outcomes	Benefits must be examined in the CER	Benefit outcomes may be more likely to be delineated in a trial registry than are specific harms. SOR of harms is also an important issue, but resource constraints did not permit us to explore harms in this project.	15

Abbreviations: AHRQ, Agency for Healthcare Quality and Research; CERs, comparative effectiveness reviews; NA, not applicable; RCT, randomized controlled trial; SAR, selective outcome reporting; SOR, selective outcome reporting.
* Reports were identified on April 12, 2011; the number of AHRQ CERs changes as additional reports are published. On that date, there were 26 CERs available on the AHRQ Website.

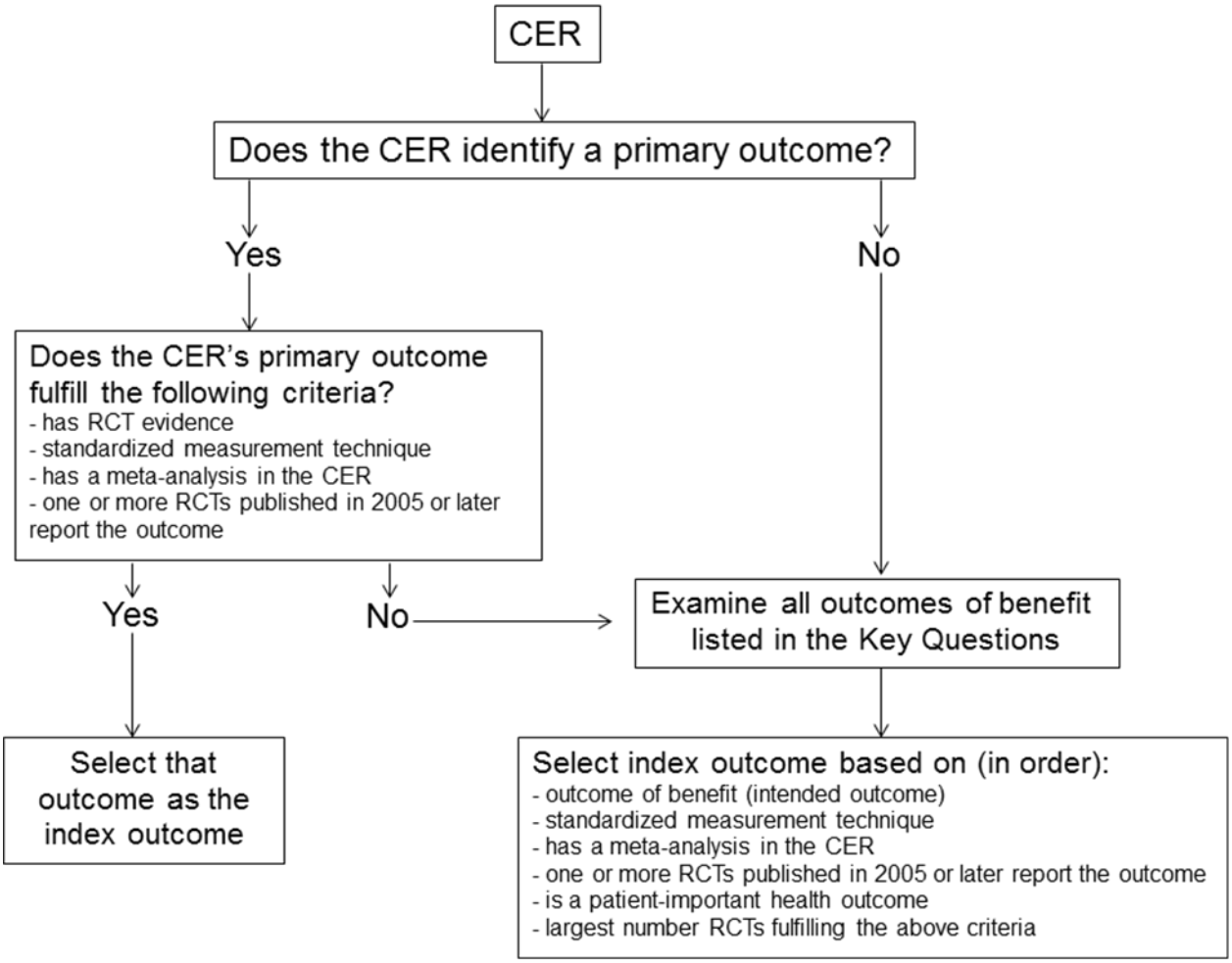
Identification of an Index Outcome within Each Included CER

We identified a single effectiveness or efficacy outcome for each CER according to the criteria in Figure 2. This outcome was either explicitly stated to be the primary outcome of the CER, or, more commonly, when the review did not specify, two authors of this methods report independently selected one outcome based on prespecified criteria. When differences between the two authors occurred, consensus was achieved through discussion.

The outcome selected for each CER must be consistently measured using the same technique and same measurement scale because it would be impossible to compare outcomes measured in different ways across studies and to attribute missing information to SOR or SAR rather than to differences in measurement approach. We also selected outcomes for which a meta-analysis was presented in the CER as initially we planned to compare pooled estimates in the CER to estimates using imputation for missing outcomes (objective 3).

We refer to the selected outcome from each CER as the "index outcome" for that CER. We selected this term to differentiate it from the primary outcome of the report, if such was identified, and from the primary outcomes delineated in studies included within each CER.

Figure 2. Process for identifying an index outcome within each included CER



Abbreviations: CER, comparative effectiveness review; MA, meta-analysis; RCTs, randomized controlled trials; SOR, selective outcome reporting.

Figure 2. Flow diagram depicting our process for identifying a single, index outcome within each included CER. Bullet points delineate the criteria for selecting an outcome for each CER.

Identification of Trial Registration for RCTs Reporting the Index Outcome

In order to identify SOR and SAR for the index outcome in our cohort of CERs and RCTs, we looked for trial registration, in other words, the publication of an agreed-upon set of information about the design, conduct, and administration of a clinical trial.[18] There are a variety of sources for this information.

In the United States, the Food and Drug Administration (FDA) Modernization Act of 1997 called for the creation of ClinicalTrials.gov and mandated registration of all efficacy drug trials for serious or life-threatening diseases and conditions conducted under FDA Investigational New Drug Application regulations.[19] An additional US federal policy in 2007 further expanded trial registration to include all drug trials other than phase 1, regulated by the FDA.[20] Each record in

ClinicalTrials.gov includes summary information on the study protocol, patient recruitment status, and the location of the study site.

The World Health Organization (WHO) initiated a policy in 2006 requiring trial registration of all medical studies that test treatments on patients or healthy volunteers.[18] WHO developed the International Clinical Trials Registry Platform (ICTRP), a global initiative that aims to make information about all clinical trials involving humans publicly available (http://www.who.int/ictrp/network/primary/en/index.html).[18] The ICTRP publishes a Search Portal, which provides access to information about ongoing and completed clinical trials. ICTRP is not a trial registry, but rather provides a single platform for access to trial registration data sets provided by a number of different trial registries, including the following:

Australian New Zealand Clinical Trials Registry
Brazilian Clinical Trials Registry (ReBec)
Chinese Clinical Trial Registry
Clinical Research Information Service - Republic of Korea
Clinical Trials Registry - India
ClinicalTrials.gov
Cuban Public Registry of Clinical Trials
German Clinical Trials Register
International Standard Randomized Controlled Trial Number (ISRCTN) Registry
Iranian Registry of Clinical Trials
Japan Primary Registries Network
Pan African Clinical Trial Registry
Sri Lanka Clinical Trials Registry
The Netherlands National Trial Register

For the purposes of registration, the ICTRP defines a clinical trial as any research study that prospectively assigns human participants or groups of humans to one or more health-related interventions in order to evaluate the effects on health outcomes.[18]

The International Committee of Medical Journal Editors (ICMJE) established a policy effective July, 2005, that requires prospective trial registration as a condition of publication and delineates the criteria for an acceptable registry.[21] The ICMJE requirement led to a marked increase in registration in ClinicalTrials.gov in 2005.[22] ClinicalTrials.gov is the largest registry accepted by the ICMJE,[22] however, the ICMJE also accepts registration in any of the primary registries that participate in the WHO platform. ICMJE journals accept "retrospective registration" (registration occurring after subject enrollment started) of trials that began before July 1, 2005. After that date, however, ICMJE considers publication of trials only if registration occurred before the first patient was enrolled ("prospective registration").[23]

Trial registration has further evolved to sometimes include study results. The US FDA Amendments Act of 2007, effective September, 2008, requires that clinical trial results be made publicly available on the internet in a database of both registry and results, although some exceptions are permitted.[24]

There are thus a multitude of study registries internationally, and policies and registry requirements have evolved over the last decade and are continuing to evolve. Through discussion among our workgroup members, we devised an approach for identifying trial registration that

optimized the sensitivity of our search (minimized missing registrations), but was still feasible given available resources. Our approach is outlined in Figure 3, and discussed in detail below.

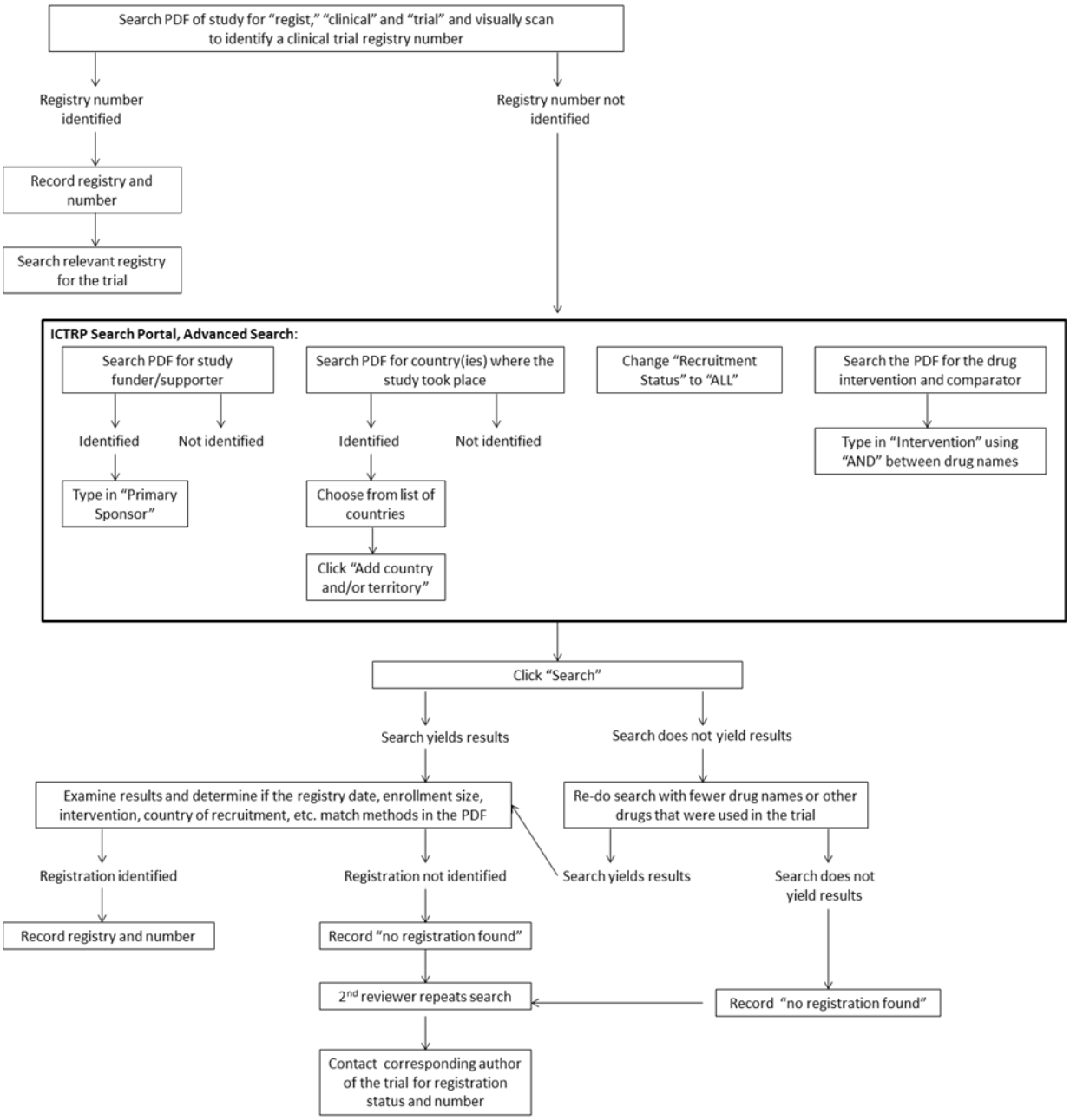
We confined our search for trial protocols on studies published from January, 2005 forward because of the effective date of the ICMJE requirement for registration of published trials (July, 2005). Although ClinicalTrials.gov went online in February, 2000,[25] registration was infrequent initially, and we wanted to focus our search efforts on a cohort of RCTs with a reasonable likelihood of registration. In addition, Zarin and colleagues[22] reported that the quality and completeness of trial records has improved since October, 2005.

We searched only ICTRP for trial registration and protocols, as this platform encompasses ClincialTrials.gov and a number of other registries, as noted above. If an RCT provided information on trial registration, particularly the National Clinical Trials (NCT) number for ClinicalTrials.gov, we were able to quickly identify the study within ICTRP. Frequently, however, study publications provided no indication that the study was listed in a registry encompassed by ICTRP. If there was no indication of trial registration in the publication, we then searched for registration using the advanced search tool within ICTRP. We used the study funder, country or countries where subject recruitment took place, and the generic names of the intervention and comparator drugs. If this approach did not reveal the study registration, we then searched on the study drug name only (generic and brand names) as we encountered instances where this less restrictive search identified relevant registrations when the more specific search did not.

This process of searching for trial registrations was repeated for all studies without an identified protocol by a second, independent searcher. If we still did not identify a study registration, we then contacted the corresponding author indicated in the trial publication by e-mail, and asked if the study was registered, and if so, what the trial registration number was.

For all studies with registration in ICTRP, we looked for results within the relevant registry. ClinicalTrials.gov is the only registry that contains results, however, and these are linked to the summary page for each trial in the registry, facilitating the identification and review of available results.

Figure 3. The process for identifying trial registration for each RCT of the index outcome



Abbreviations: ICTRP, International Clinical Trials Registry Platform; PDF, portable document format; WHO, World Health Organization.

Figure 3. See text for a detailed description of the process for identifying trial registration.

Data Abstraction

Data from the primary publication and any companion papers cited in the CER as well as information from the trial registration were abstracted into a standardized template in Excel (Microsoft Office Excel 2007, Microsoft, Redmond, WA). This template was piloted by study staff at Oregon Health & Science University and by members of the workgroup, using the CER on oral hypoglycemic agents.[15] The template was revised as needed during the pilot phase. Abstracted data were reviewed by a second team member to ensure accuracy. Abstracted data included information from the publication (author affiliations and disclosures, study funder, interventions, relevant outcomes, analytic approach, and analysis set) and registry (dates of registration and recruitment, proposed primary and secondary outcomes and any changes among those, relevant results (if any).

Identifying SOR and SAR

In order to determine the proportion of trials with SOR and/or SAR within a CER, we created a matrix of all the outcomes related to the designated index outcomes. This approach was similar to that used by Kirkham and colleagues in their Table 1.[2] Each row in the matrix represented an RCT that had an identifiable trial registration, and each column contained different approaches to presenting and/or measuring the index outcome across the included RCTs and trial registration information. Specifically, each column in the matrix contained outcomes:

1. mentioned in the methods section of the publication;
2. reported in the results section of the publication;
    a. including outcomes reported in different ways (e.g., categorical (with various definitions of categories)) or continuous; and
3. reported in the trial registry (either registration information or results).

This exploration and data display enabled us to depict the universe of potential approaches to presenting the index outcome for each CER, and to identify missing outcomes in each trial report or in the trial's registry information. We did not seek clinical input to discern other relevant outcomes as we felt that the list of outcomes that we identified in the CER and trial registry adequately represented the universe of potential outcomes.

In our exploration of SOR and SAR within each included trial, we started with the definition of SOR and SAR delineated by Kirkham and colleagues.[2] We soon found, however, that there were many additional facets of potential reporting biases related to selection of outcomes and analyses that were not encompassed by this definition. We therefore developed a list of additional potential types of SOR/SAR, with categorizations and examples listed in Table 3.

We used a similar approach to identify SAR: we noted the planned analyses in the methods section of publications and any information in the trial registry, and compared those to the reported analyses within and across trials. We did not expect that information in trial registries would be as detailed for SAR as for SOR, nonetheless, we sought information on SAR in a similar manner. We did not include information that was provided in the results section of the registry (when available) on how analyses had been performed, as that information was added to the registry with the results, and thus could not be considered to have been conceived *a priori*.

Since this work was exploratory, we did not develop a specific list of discrepant or altered analyses that we would consider SAR. Rather, we sought to describe any change in analytic

approach between the methods section of the publication or the registry summary information, and results reported in the publication. SAR includes, for example, selective reporting of data on subgroups, presentation of adjusted versus unadjusted analyses, selection different analytic sets (e.g., as-treated versus intention-to-treat), selective approaches to the handling of missing data, choosing to analyze continuously measured variables categorically or changing measurement scales, and choice of cut-point values to define categorical variables.[5] We sought cases where there was a change from the originally specified analytic approach, rather than situations where analyses appropriately tested different models to achieve optimal fit to the data.

In order to determine definitively whether we considered SOR/SAR to exist in each study, we developed and applied three judgments about SOR/SAR to each trial. These judgments are outlined in Table 3 and depicted in Figure 4. Judgment #1 was based on the comparison of the methods and results sections of the publication; judgment #2 on the methods in the trial registry compared to the results in the publication; and judgment #3 on the results in the registry (if any) versus those in the publication. Each judgment was made independently by two authors (SLN and HKH), and consensus was achieved through discussion.

Table 3. Types of SOR and SAR classified by the approach used for detection

Judgment #1	Judgment #2	Judgment #3*	Comments
Publication methods versus publication results	**Trial registry summary versus publication results**	**Trial registry results versus publication**	
		Differences in the numerical values reported in the publication and the registry	
Change in primary and/or secondary outcome designation for the index outcome between methods and results	Change in primary and/or secondary outcome designation for the index outcome between the registry and the publication		"Yes" if there was a switch in primary versus secondary designation of the index outcome between the registry and the publication.
Change in the index outcome between methods and results	Change in the index outcome between registry and publication		This includes a change in the specificity of the description of the within the registry outcome.
Outcome listed in methods but not reported in results	Outcome listed in the registry but NR in the publication	Outcomes reported in the registry results but not in the publication	Typical SOR per Kirkham and colleagues[2]
Outcome in methods inadequately reported in results	Outcome listed in the registry inadequately reported in publication	Outcomes reported in the registry results but not adequately reported in the publication	Typical SOR per Kirkham and colleagues[2]
	Change in followup interval between registry and publication		Includes any discrepancy between the followup interval specified in the registry and in the publication.
Outcome reported in results but not reported in methods	Outcome reported in publication but not listed in the registry	Outcomes reported in the publication but not in the registry results	Type of SOR that differs from the classes depicted by Kirkham and colleagues[2]
Change in analyses between methods and results	Change in analyses between registry and publication	Change in analyses between registry and publication	
Subgroups reported in results that were not described in the methods section or vice versa	Subgroup specified in the registry but NR in the publication(s) or vice versa	Subgroup reported in the registry but NR in the publication, or vice versa.	
Summary: judgment #1: SOR and/or SAR are deemed to be present if there are discrepancies in description of the index outcome between the publication methods and results section.	**Summary: judgment #2:** SOR and/or SAR are deemed to be present if there are discrepancies in description for the index outcome between the trial registry summary information and the results section of the publication.	**Summary: judgment #3:** SOR and/or SAR are deemed to be present if there are discrepancies in the results presented for the index outcome between the trial registry results and the results presented in the publication. This includes both the publication of a subset of the results in the registry and the publication of additional results not presented in the registry.	

Abbreviations: NR, not reported; SAR, selective analysis reporting; SOR, selective outcome reporting

* Judgment #3 is not applicable if there are no results in the registry or no differences in results between the publication and registry

Figure 4. Judgments on the presence of selective outcome and analysis reporting

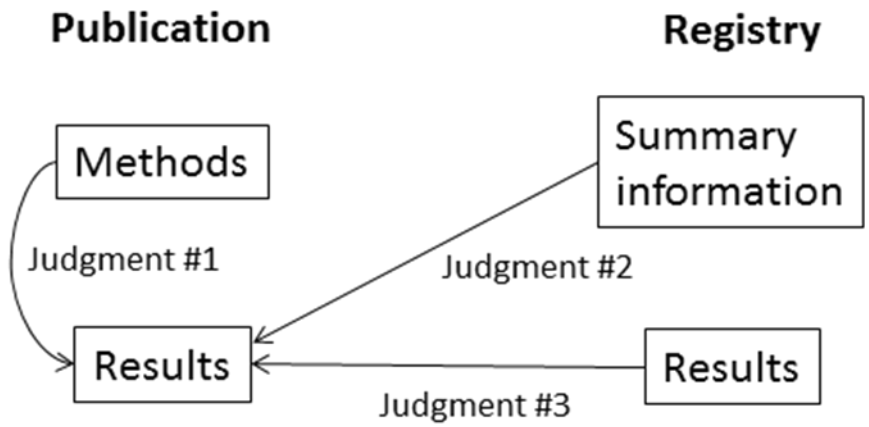


Figure 4. Schematic of the judgments on the presence of selective outcome and analysis reporting (SOR and SAR). Judgment #1 compares the methods and results sections of the trial publication. Judgment #2 compares summary information in the registry to the results reported in the publication for that trial. Judgment #3 compares the results reported in the registry to the results reported in the publication. Discrepancies in one of these comparisons constitute an assessment of "yes" for that judgment, and are indicative or SOR and/or SAR.

Objective 2: Predictors of SOR and SAR

We explored potential predictors of SOR and SAR, using the following pre-specified, study-level, independent variables. Because this is an exploratory analysis with essentially no existing evidence base, we selected a variety of diverse predictors based on our hypotheses. First, we predicted that funding and other conflicts of interest on the part of the funder and authors might influence the frequency of reporting biases. We based this prediction on existing data that suggest that financial and other conflicts of interest can affect the data presented and the conclusions in primary studies.[26-29] Second, we sought to examine the possibility that SOR and SAR might vary with the characteristics of the registry and the timing of trial registration. We theorized that the level of detail required in a registry and the timing of registration (ideally before subject recruitment starts) might impact the frequency of SOR. Third, we explored the situation where trial authors changed outcomes in the registry between initial registration and the last update of outcomes in the registry. Since this was not captured in our definitions of SOR and SAR, we hypothesized that such changes might relate to the selection of outcomes based on the results and might therefore correlate SOR and SAR. We therefore examined the following independent variables:

1. Characteristics of the funder and authors
    a. One or more study funders or sponsors manufactured one or more of the intervention drugs (categorical variable, yes/no).
    b. One or more authors were employees of a company making one of the intervention drugs (categorical variable, yes/no; percentage of authors who were employed by the company making one of the intervention drugs (proportions).

c. Assistance in writing the manuscript was provided by study funder (categorical yes/no). When the source of funding for authorship was unclear or not reported, the response was considered "no".

2. Trial registration
    a. Registry in which study was registered (categorical variable, ClinicalTrials.gov versus other registry)
    b. Timing of trial registration with respect to first subject recruitment (categorical variable, before or the same month, versus after or unclear).
    c. Studies with results versus studies with no results in the registry. Here we only examined judgments #1 (publication methods versus results) and #2 (registry summary versus publication results) as the dependent variable. (Note that only ClinicalTrials.gov provided results.)
3. Changes in the index outcome within the registry
    a. Studies that changed the index outcome between "original" and "current" within the trial registry (categorical variable, yes/no). The following were considered "yes" responses:
        i. change in designation between primary and secondary outcome;
        ii. change in specificity (e.g., "glycemic control" was the original outcome and "change in A1c between baseline and week 24" was the current outcome in the trial registry); or
        iii. addition of an outcome to the registry (e.g., "percentage of patients who achieved A1c < 7.0 percent at followup" was added to the registry where "change in A1c" was the only original A1c-related outcome).

We examined change in the index outcome (3, above) as a predictor of judgments #1, #2, #3 (our definition of SOR), as changes between the "original" and "current" outcomes listed in trials registries were not part of our definition of SOR. Our criteria for SOR involved only a comparison of the "current" outcome listed in the registry to what was published in the publication (judgment #2: registry summary versus publication results). Thus a change in outcome from "original" to "current" might be a predictor of SOR as we defined it. In other words, if trial authors made changes to the primary and secondary outcomes listed in the registry, we explored the correlation of such changes with SOR.

The dependent variables for this analysis were:

1. Presence of SOR and/or SAR detected by examining the publication only ("judgment #1": publication methods versus results)
2. Presence of SOR/SAR detected by comparing registry information to the results reported in the publication ("judgment #2": registry summary versus publication results)
3. Presence of SOR/SAR detected by comparing the results in the registry (if any) to the results reported in the publication ("judgment #3": registry results versus publication results).

We combined data from the reports on oral hypoglycemic agents[15] and osteoporosis[16] as we considered it likely that the potential predictors of our judgments on SOR/SAR would be similar between the two reports. For the report on lipid modifying agents,[17] the dependent outcomes for

all RCTs were judged to have the same type of SOR, precluding statistical analysis to explore predictors. In order to examine the relationships between the timing of initial study registration, specification of final primary and secondary outcomes, and publication, we explored the chronology of these events in detail by developing a timeline for each RCT for the CER on oral hypoglycemic agents,[15] and osteoporosis[16] based on information from both the study publication and the trial registry.

Objective 3: Assessment of the Reliability and Validity of the ORBIT Classification of SOR

We initially proposed examining the inter-rater reliability and validity of the ORBIT classification system, in addition to the feasibility of its implementation. For our assessment of inter-rater reliability, each index outcome within each included RCT was classified by two independent assessors as fully reported, partially reported, or not reported, using the ORBIT categories A through I.[2] The assessors are coauthors on this report (SLN and HKH), have formal training in epidemiology, and are experienced in systematic review methods including the assessment of risk of bias in primary research studies. After documentation of the two independent assessments, the two assessors achieved consensus through discussion. The assessors initially piloted the ORBIT assessments by independently rating four included RCTs, followed by discussion and consensus on the ORBIT classification. For the RCTs that were not part of the training exercise, we planned to calculate a kappa statistic as well as percentage agreement between assessors for each of the nine ORBIT categories and overall. We were unable to calculate kappa statistics, however, for reasons described in the results section.

We planned to assess the validity of the ORBIT classification system by comparing the ORBIT classification of SOR using the trial publication (the manner in which ORBIT was intended[2]) to the ORBIT classification achieved using additional information from the trial registry. We planned to examine the percentage agreement in the classification between these two approaches. This also proved infeasible as discussed in the results section.

We examined one aspect of the feasibility of using ORBIT by recording the time in minutes it took each assessor to complete the ORBIT classification using the trial publication and registry information, and then the time it took to examine trial results in the registry (if any) and to make any further assessments of ORBIT.

Data Syntheses and Analyses

This was an exploratory study, and so we selected a convenience sample of CERs and RCTs to examine, and no sample size calculations were performed. The three CERs that we examined involved different types of outcomes, so we described each of the three cohorts of RCTs separately.

Descriptive statistics were used to present our findings on SOR and SAR among the RCTs included in this study. In addition, we examined the association between potential predictors as categorical variables and the dependent categorical variable of the presence or absence of SOR and/or SAR as indicated by judgments #1, #2, and #3 as outlined above. Due to the small sample size, the association between the presence of SOR/SAR and study-level characteristics was explored using Fisher's exact test and exact logistic regression. Since there was no significant difference between SOR/SAR and the two CERs (diabetes and osteoporotic fractures), the analysis was first conducted by combining data from the diabetes and fracture studies to assess

overall association. Then the analysis was conducted by examining RCTs in the two CERs separately to look at association for each individual CER.

Results

Cohort of CERs and the choice of index outcome for each CER

Table 4 indicates the 15 CERs that fulfilled our inclusion criteria (Table 2). Table 4 also presents the index outcome for each of the CERs, with the number of RCTs in each CER for that outcome. We selected three CERs to explore in detail. [15-17] In order to have a reasonable number of trials to examine the frequency and predictors of SOR and SAR, we selected the three CERs with the largest number of trials examining the index outcome. Resource constraints did not permit us to examine a larger number of CERs in this exploratory work. The rationale for selecting A1c for the index outcome on the oral hypoglycemic agent report[15] was that there were a large number of RCTs that examined this intermediate outcome, it was the first outcome addressed in key question 1 (although it was not labeled as the primary outcome of the review), and there were a large number of meta-analyses. Mortality was chosen for the CER on lipid agents[17] because there was a meta-analysis of that outcome in the report, and we wanted to focus on a patient-important, health outcome. We had anticipated that decreased mortality would be reported as an outcome of benefit in this report, however examination of the RCTs reporting this outcome revealed that (increased) mortality was always considered an adverse event. For the CER on medications for fracture prevention in women with low bone density or osteoporosis,[16] a large number of trials reported fractures as an outcome, it is a patient-important outcome, and there were meta-analyses reported in the CER.

Table 4. Index outcome for each comparative effectiveness review

Comparative Effectiveness Review	Index Outcome	Number of Included RCTs
1. Oral Diabetes Medications for Adults with Type 2 Diabetes. An Update	A1c	24
2. Effectiveness of Recombinant Human Growth Hormone (rhGH) in the Treatment of Patients With Cystic Fibrosis	Pulmonary function testing, including both forced expiratory volume in one second (FEV1) and forced vital capacity (FVC)	1
3. Comparative Effectiveness of In-Hospital Use of Recombinant Factor VIIa for Off-Label Indications vs. Usual Care	Mortality	4
4. Comparative Effectiveness of Angiotensin Converting Enzyme Inhibitors or Angiotensin II Receptor Blockers Added to Standard Medical Therapy for Treating Stable Ischemic Heart Disease	Mortality	1
5. Comparative Effectiveness of Medications To Reduce Risk of Primary Breast Cancer in Women	Invasive and non-invasive breast cancer	5
6. * Comparative Effectiveness of Lipid-Modifying Agents	All-cause mortality	14
7. * Comparative Effectiveness, Safety, and Indications of Insulin Analogues in Premixed Formulations for Adults With Type 2 Diabetes	A1c	11
8. * Comparative Effectiveness of Treatments To Prevent Fractures in Men and Women With Low Bone Density or Osteoporosis	Fracture reduction, including both vertebral and extremity/hip/etc	12
9. Comparative Effectiveness of Drug Therapy for Rheumatoid Arthritis and Psoriatic Arthritis in Adults	American College of Rheumatology 50	1
10. Comparative Effectiveness of Angiotensin-Converting Enzyme Inhibitors (ACEIs) and Angiotensin II Receptor Antagonists (ARBs) for Treating Essential Hypertension	Blood pressure control	2
11. Comparative Effectiveness of Second-Generation Antidepressants in the Pharmacologic Treatment of Adult Depression	Response in treating depressive symptoms (according to various depression rating scales)	0
12. Efficacy and Comparative Effectiveness of Off-Label Use of Atypical Antipsychotics	Obsessive-compulsive disorder: Yale-Brown Obsessive-Compulsive Scale	1
13. Comparative Effectiveness and Safety of Analgesics for Osteoarthritis	Pain relief	0
14. Comparative Effectiveness of Epoetin and Darbepoetin for Managing Anemia in Patients Undergoing Cancer Treatment	Hemoglobin levels	2
15. Comparative Effectiveness of Management Strategies for Gastroesophageal Reflux Disease	Complete symptom relief at 4 and 8 weeks; time to complete resolution of symptoms	1

Abbreviations: RCTs, randomized controlled trials

*Comparative Effectiveness Reviews selected for further evaluation of SOR and SAR

Objective 1: Frequency of SOR and SAR

Table 5 summarizes the characteristics of the RCTs for the index outcomes of the three CERs examined in this report. Registered trials variably reported trial registration numbers in the publications (28.6, 66.7, and 75 percent for lipid modifying agents, osteoporosis and oral hypoglycemic agents CERs, respectively). The majority of registered studies were listed in

ClinicalTrials.gov, although this varied across the three CERs. The percentage of trials with results posted in the registry also varied: 8.3, 14.3, and 62.5 percent for osteoporosis, lipid modifying agents, and oral hypoglycemic agents CERs, respectively .

Tables 6-9 summarize our findings on SOR and SAR, including our three consensus judgments on the presence of SOR/SAR as defined previously in Table 3 and depicted in Figure 4.

The cohort of trials reporting A1c (n=24) (Tables 6-8), reported this outcome in a variety of ways, both as a continuous and as a categorical variable. In none of these trials did the designation of A1c as a primary or secondary outcome change within the publication. On comparison of the methods and results section of the trial publication (judgment #1: publication methods versus results), the outcomes listed in the methods section were always reported in the results, although in two studies (8 percent) the outcome was not fully reported. In three studies, however, A1c outcomes were reported in the results section that were not indicated in the methods section. In addition, in five studies subgroup data were presented on A1c that was not specified in the methods section of the publication.

In 46 percent (n=11) of the trials reporting A1c we judged SOR and/or SAR to be present based on the publication only (judgment #1: publication methods versus results). In seven of those 11 studies, we were able to assign an ORBIT class, but in the remaining four we could not. These latter situations were where outcomes were presented in the results that were not mentioned in the methods section.

When the A1c trial publications were examined along with information in the registry summary (but not the registry results) (judgment #2: registry summary versus publication results), SOR and/or SAR was assessed as present in 20 studies (83 percent). In none of these trials could we apply an ORBIT class, as all entailed the addition of results to the trial publication that were not specified in the registry. Of the 16 trials with results in the registry, two were assessed as having SOR/SAR (judgment #3: registry results versus publication results), and ORBIT classes did not apply as these were both the situation of results added to the publication that were not in the registry.

We report the frequency of prespecified potential predictors of SOR and SAR for A1c trials in Table 8. 92 percent of studies had one or more authors who were an employee of the company sponsoring the study and making one or more of the intervention drugs. Almost half of studies (46 percent) received assistance from the study funder in authoring the publication. Also in 46 percent of studies, was the first subject recruited before the study was registered in the trials registry.

The cohort of trials reporting the outcome of fractures (n=12) is summarized in Tables 9. These studies were much better reported, with only four studies reporting outcomes in the results section that were not mentioned in the methods section of the publication (judgment #1: publication methods versus results). Likewise, there was much better agreement between trial registry outcome information and the reported results, although judgment #2 (registry summary versus publication results) was still positive in 58 percent of studies. ORBIT class did not apply in any of these trials where SOR/SAR was judged to be present. Only one study of the eight registered in ClinicalTrials.gov had results reported in the registry and in that trial results differed between the registry and the publication (judgment #3). Industry employment of the study authors (33 percent) was less common than for the A1c trials (92 percent).

Among the cohort of trials that reported mortality outcomes with lipid agents (n=14), this outcome was always reported as an adverse events (ie., decreased mortality was not reported as

we had anticipated). Eight trials mentioned safety or adverse events in the methods section of the publication, but only one explicitly mentioned death as an outcome in the methods section (judgment #1: publication methods versus results). None of these studies mentioned mortality as a primary or secondary outcome in the registry summary information (judgment #2: registry summary versus publication results), and neither of the two studies reporting results in the registry mentioned mortality (judgment #3: registry results versus publication results). As a result of the homogeneity of reporting across these studies and the fact that none of these studies considered mortality a primary or secondary outcome of either benefit or harm, we did not explore this cohort of studies further.

Table 5. Characteristics of trials that reported the index outcomes

CER	Index outcome	Unique RCTs	RCTs published 2005 to present (n)	Registration identified (n (% 2005-present*; % all years**))	Trials with registry number reported in the publication n (%)	Trials included in ClinicalTrials.gov (n (% of trials registered))	Trials included in ISRCTN Register (n (% of trials registered))	Trials listed in other registries (n (% of trials registered))	Trials with results (n (% of trials registered))
Oral Diabetes Medications for Adults With Type 2 Diabetes	A1c	98	59	24 (40.7; 24.5)	18/24 (75)	23 (95.8)	0 (0.0)	1 (4.2) UMIN Clinical Trials Registry	15 (62.5)
Comparative Effectiveness of Lipid-Modifying Agents	All-cause mortality	24	19	14 (73.7; 58.3)	4/14 (28.6)	11 (78.6)	2 (14.3)	1 (7.1%) Cochrane Renal Group	2 (14.3)
Comparative Effectiveness of Treatments To Prevent Fractures in Men and Women With Low Bone Density or Osteoporosis	Fracture reduction, including both vertebral and extremity, hip, other sites)	177	29	12 (41.4; 6.8)	8/12 (66.7)	8 (66.7)	2 (16.7)	2 (16.7) Australia New Zealand Clinical Trials Registry	1 (8.3)

Abbreviations: CER, comparative effectiveness review; ISRCTN, International Standard Randomised Controlled Trial Number; n, number; RCTs, randomized controlled trial

* Percentage of trials reporting the index outcome published in or after 2005 that were registered

** Percentage of all trials reporting the index outcome that were registered

Table 6. SOR and SAR for A1c and fracture outcomes based on the publication only (judgment #1)

	Change in A1c from baseline	Subjects with A1c <7.0%	Subjects with A1c <6.5%	Subjects with a decrease in A1c > specific value	Other A1c O/C	Change in primary and/or secondary O/C	O/C in methods not reported in results	O/C in methods not fully reported in results	O/C in results not listed in methods	Change in analyses between methods and results	Subgroup results not described in methods	SOR/ SAR present: Judgment #1	ORBIT class
No (%)	0 (0)	8 (33)	12 (50)	23 (96)	22 (92)	24 (100)	24 (100)	22 (92)	21 (88)	23 (96)	19 (79)	13 (54)	No SOR/ SAR: 13
Yes (%)	24 (100)	16 (67)	12 (50)	1 (4)	0 (0)	0 (0)	0 (0)	2 (8)	3 (13)	1 (4)	5 (21)	11 (46)	*A: 2; B:2; D: 3; E: 1; NA: 4
Fracture outcomes (n=12)													
	Total	**Hip**	**Vertebral**	**Non-vertebral**	**Wrist**	**Other**							
No (%)	1 (8)	7 (58)	6 (50)	7 (58)	8 (67)	7 (58)	12 (100)	12 (100)	12 (100)	8 (67)	12 (100)	12 (100)	8 (67)	No SOR/ SAR: 8
Yes (%)	11 (92)	5 (42)	6 (50)	5 (42)	4 (33)	5 (42)	0 (0)	0 (0)	0 (0)	4 (33)	0 (0)	0 (0)	4 (33)	ORBIT NA: 4

Abbreviations: NA, not applicable; O/C, outcome; ORBIT, Outcome Reporting Bias in Trials; SAR, selective analysis reporting; SOR, selective outcome reporting.
A, B, D, E are ORBIT classes.
*One of the 11 publications with SOR/SAR was assigned two ORBIT classes.
Data are number of trials (percentage of all trials where n=24 [A1c] and n=12 [fractures]) unless otherwise specified.

Table 7. SOR and SAR for A1c and fracture outcomes based on the publication and on the registry (judgment #2 and #3)

	Registry without registry results							With registry results	
	Change in O/C between registry and publication	**Change in status of O/C* between the registry and the publication**	**Change in followup interval between registry and publication**	**O/C not reported in publication but in registry**	**O/C reported in publication but not reported in registry**	**SOR/SAR present: judgment #2** (registry summary versus publication results)	**ORBIT class (judgment #2)**	**SOR/SAR present: judgment #3** (registry results versus publication results)	**ORBIT class (judgment #3)**
A1c outcomes (n=24)									
No (%)	23 (96)	24 (100)	21 (88)	24 (100)	4 (17)	4 (17)	No SOR/SAR: 4	No results: 8	No SOR/SAR: 14
Yes (%)	1 (4)	0 (0)	2 (8)	0 (0)	20 (83)	20 (83)	ORBIT NA: 20	2/16 (13)	ORBIT NA: 2
Fracture outcomes (n=12)									
No (%)	12 (100)	12 (100)	9 (75)	12 (100)	5 (42)	5 (42)	No SOR/SAR: 5	No results: 11	No SOR/SAR: 11
Yes (%)	0 (0)	0 (0)	3 (25)	0 (0)	7 (58)	7 (58)	ORBIT NA: 7	1 (83)	ORBIT NA: 1

Abbreviations: NA, not applicable; O/C, outcome; ORBIT, Outcome Reporting Bias in Trials; SAR, selective analysis reporting; SOR, selective outcome reporting.
*Change in designation of the index outcome as primary or secondary.
Data are number of trials (percentage of all trials where n=24 [A1c] and n=12 [fractures]) unless otherwise specified.

Table 8. Potential predictors of SOR and SAR for A1c and fracture outcomes

	Predictors						
	Author industry affiliation	**Study drug made by study sponsor***	**Assistance authoring the publication by study funder**	**Trial registered in ClinicalTrials.gov**	**First subject recruited before trial registered**	**Registry reports results**	**Change in A1c measure between original and current O/C listed in registry**
A1c outcomes (n=24)							
No (%)	2 (8)	0 (0)	13 (54)	1 (4)	13 (54)	8 (33)	19 (79)
Yes (%)	22 (92)	23 (96)	11 (46)	23 (96)	11 (46)	16 (66)	5 (21)
Fracture outcomes (n=12)							
No (%)	8 (67)	6 (50)	10 (83)	4 (33)	11 (92)	11 (92)	11 (92)
Yes (%)	4 (33)	6 (50)	2 (17)	8 (67)	0 (0)	1 (8)	1 (8)

Abbreviations: O/C, outcome; SAR, selective analysis reporting; SOR, selective outcome reporting.
*The trial sponsor was not reported in one of the A1c trials.
Data are number of trials (percentage of all trials where n=24 [A1c] and n=12 [fractures]) unless otherwise specified.

Table 9. Reporting of mortality in lipid studies (n=14)

	Trials with Adverse Events Mentioned in the Methods Section of the Publication	**Trials with Mortality Mentioned in the Methods Section**	**Trials with Adverse Events Mentioned in the Registry Summary**	**Trials with Results in the Registry**	**Trials with Mortality Reported in the Registry**
Yes (%)	6 (43)	1 (7)	0 (0)	2 (14)	0 (0)
No (%)	8 (57)	13 (93)	14 (100)	12 (86)	14 (100)

Data are number of trials (percentage of all trials where n=14).

Timing of study registration

The timing of trial registration occurred at various times during the course of included trials (Table 10). A minority of studies was registered before or in the same month that subject recruitment started. Studies commenced before 2005 (the year ICMJE recommendations were implemented) reasonably could have been registered later in the course of the study, although ClinicalTrials.gov was launched in 2000.

Table 9. Timing of study registration

Oral Diabetes Medications for Adults With Type 2 Diabetes[15] (n=24)	Comparative Effectiveness of Lipid-Modifying Agents[17] (n=14)	Fractures in Men and Women With Low Bone Density or Osteoporosis[16] (n=12)	Time of Study Registration
11 (41.7%)	4 (28.6%)	0 (0.0%)	Before recruitment started or in the same month that recruitment started
16 (66.7%)	4 (28.6%)	5 (41.7%)	Before primary completion date/primary completion date for recruitment/ before date recruitment completed (from article)
18 (75.0%)	7 (50.0%)	7 (58.3%)	Before study completion date
19 (79.2%)	10 (71.4%)	8 (66.7%)	Before publication submitted
21 (87.5%)	10 (71.4%)	9 (75.0%)	Before publication accepted
21 (87.5%)	10 (71.4%)	10 (83.3%)	Before article first published online
21 (87.5%)	11 (78.6%)	10 (83.3%)	Before publication printed
24 (100.0%)	14 (100.0%)	12 (100.0%)	Before 2012

Timing of Changes in Outcomes in the Registry

We examined the proportion of trials with a change in the index outcome between the "original" outcome and the "current outcome" in the trial summary in the registry, and the timing of those changes (Table 11). The outcome of mortality was never mentioned among the outcomes listed in the trial summaries for the CER on lipid management, so there was no documentation of any changes. 41.7 percent of trials in the CER on oral hypoglycemic agents had a change in some aspect of A1c measurement, whereas 16.7 percent of trials in the CER on fracture prevention had a change in our index outcome indicated in the registry.

Table 11. Timing of changes in the registry of the index outcome

Oral Diabetes Medications for Adults With Type 2 Diabetes[15] (n=24)	Fractures in Men and Women With Low Bone Density or Osteoporosis[16] (n=12)	Time of Change in Index Outcome
0	0	Before recruitment started or in the same month that recruitment started
0	0	Before the primary completion date or the primary completion date for recruitment or before date that recruitment was completed (from trial publication)
0	0	Before study completion date
1	0	Before publication submitted to publishing journal
2	0	Before publication accepted
2	0	Before article first published online
3	0	Before publication printed
10	2	Before 2012
10/24 (41.7%)	2/12 (16.7%)	Total number of trials with no change in the primary outcome

Note: Comparative Effectiveness of Lipid-Modifying Agents[17] was not included because the index outcome of mortality was never listed as an outcome in the registry summary or in the publication methods section.
Data are the number of trials.

We illustrate the chronology of trial registration and changes in outcomes with three examples (Figures 5, 6, 7). Dates above the horizontal line refer to dates from the publication; dates below the line are from the trial registry. Figure 5 provides an example of a reasonable study-registration chronology, wherein the study was registered at the time it was started and results were posted. Ideally, however, results would have been posted earlier, at the time of publication. Figures 6 and 7 illustrate suboptimal chronologies. In Figure 6, the trial was registered after the study was submitted for publication, and the results submitted 2 years later. In Figure 7 the study was also registered after publication and results were never reported.

Figure 5. Example of study and registry chronology[30] (see text for details)

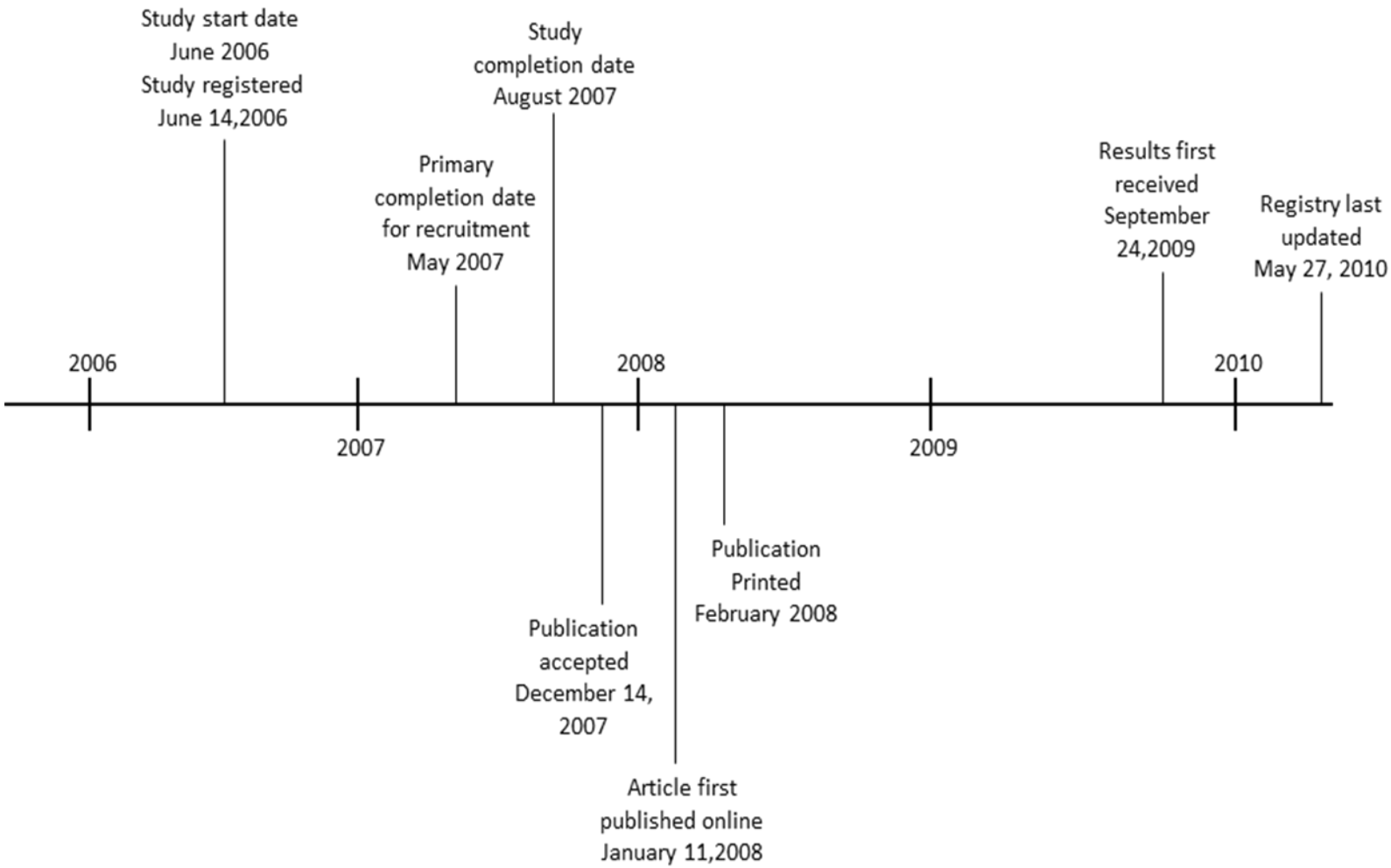


Figure 6. Example of study and registry chronology[31] (see text for details)

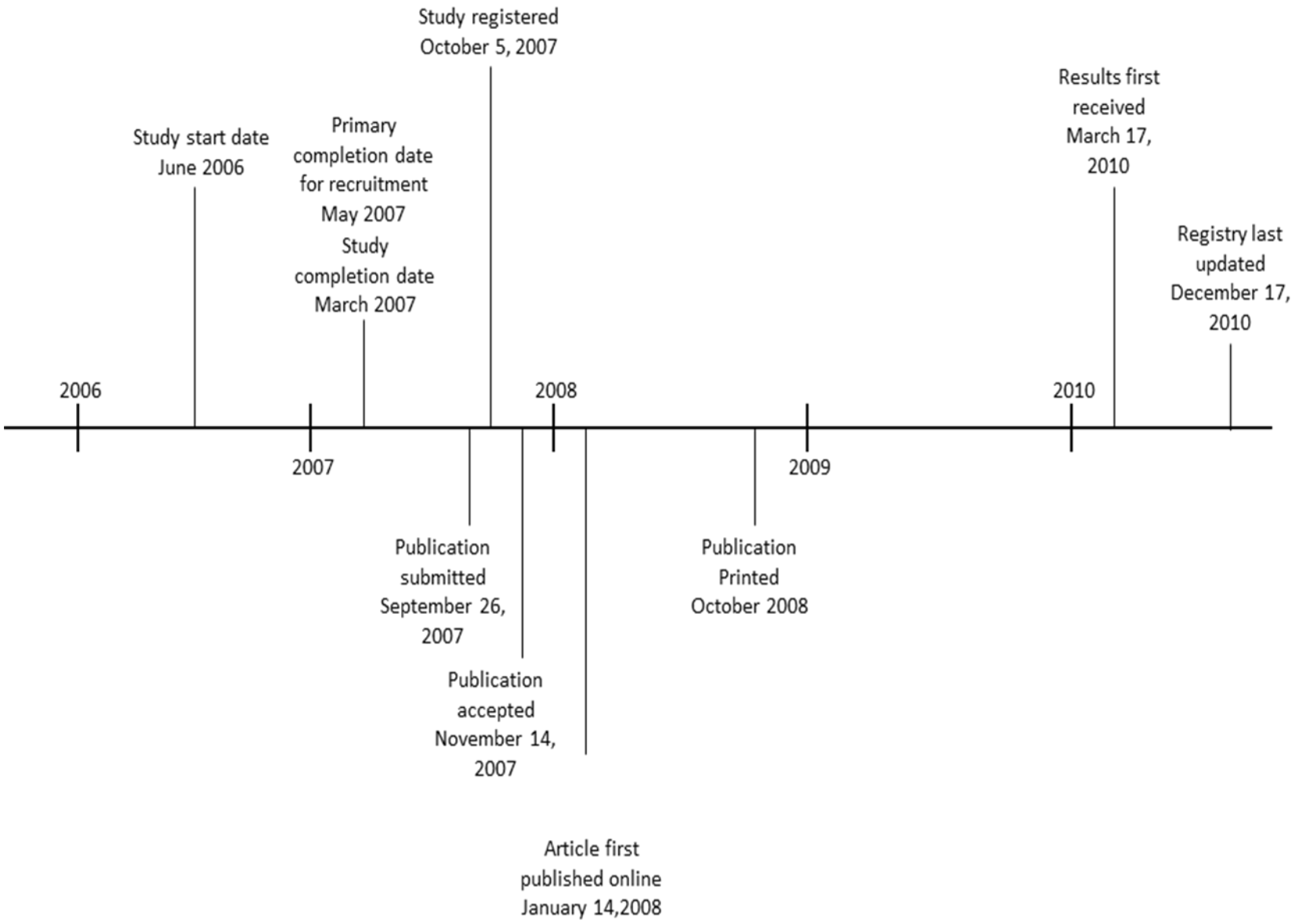
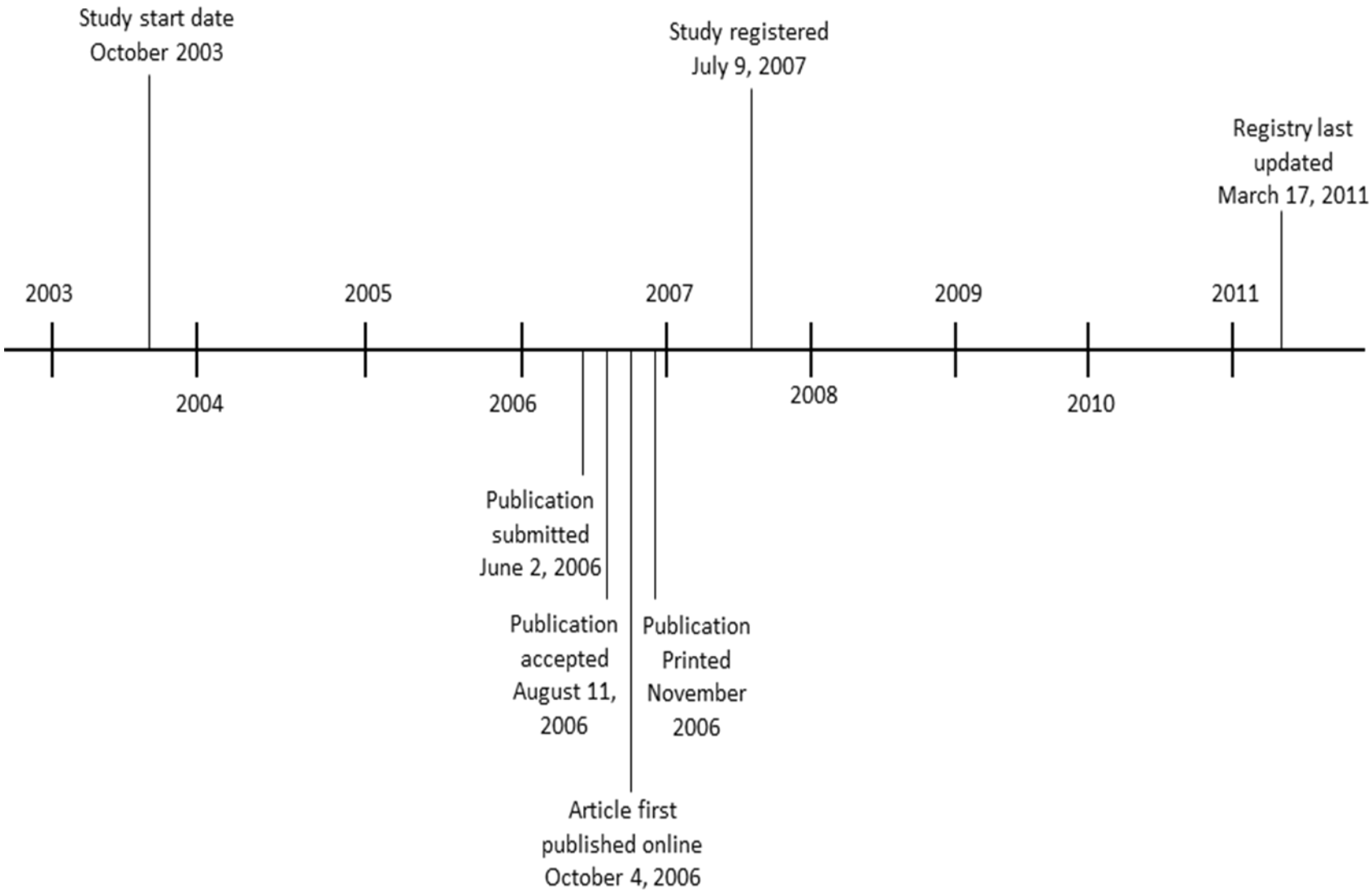


Figure 7. Example of study and registry chronology[32] (see text for details)



Efforts Taken to Identify SOR and SAR

The time it took the two assessors to examine the trial registry for SOR and SAR is reported in Table 12. There was considerable variation across studies: examination of the summary information in the registry took as few as 3 minutes and as long as 25. Examination of trials with results in the registry took the two assessors 21 and 23 minutes on average for the studies examining A1c. We did not collect data for the trials on mortality in lipid agents.

Table 10. Time (in minutes) to complete review of ClinicalTrials.gov

Report (n=number of studies reviewed)	Mean without results (range) Number of studies timed		Mean with results (range) Number of studies timed	
	Reviewer 1	Reviewer 2	Reviewer 1	Reviewer 2
Diabetes (n=24)	13 (5 to 23) n=23	10 (3 to 25) n=18	23 (16 to 38) n=14	21 (9 to 34) n=12
Lipids (n=14)	NA	NA	NA	NA
Osteoporosis (n=12)	10 (4 to 15) n=11	7 (3 to 15) n=12	17 (17 to 17) n=1*	19 (19 to 19) n=1*

Abbreviations: NA, not applicable; n, number of trials.
*The osteoporosis review only had one trial with results reported.

Objective 2: Predictors of SOR

We explored the relationships between various prespecified characteristics of study authors, trial registration, and changes in the index outcome within the registry, and our assessment of SOR and SAR as summarized by judgments #1, #2, and #3 for the CERs on oral hypoglycemic agents and osteoporosis. We analyzed both CERs separately and combined, and found no significant relationships (all p-values >0.05) (Appendix Table A1 and A2). There was not sufficient variation in the judgments of SOR and SAR for the CER on lipid modifying agents and mortality to perform predictor analyses for that CER.

Objective 3: Reliability and Validity of ORBIT

Our proposed determination of the inter-rater reliability and validity of the ORBIT classification system proved infeasible. We were unable to determine inter-rater reliability for the ORBIT classifications of the trials that we examined for two reasons. First, very few trials were assessed as having an ORBIT class (Table 6, 7, 8). Second, the two assessors found that many more than the initially proposed pilot of four trials were required to understand the types of SOR and SAR that were present among the trials. Each additional trial appeared somewhat different in its presentation of SOR and SAR, and there were a multitude of nuances. The two assessors therefore evolved their categorization of SOR and SAR (i.e., judgments #1,#2, #3) through their consensus process, and developed a categorization of SOR and SAR as they proceeded with the consensus exercise. It was thus impossible to determine inter-rater reliability with such a process.

Our proposed assessment of the validity of ORBIT also proved infeasible. In addition to the fact that the two assessors noted that very few studies permitted designation of an ORBIT category, the assessors also found that information in trial registries did not enable validation of the ORBIT classification system. Only ORBIT classes G and H were applicable to information from sources outside of the trial publication, and we did not assign any trials to those two classes.

Discussion

Our work identifying and characterizing SOR and SAR in a small cohort of trials included in recent AHRQ-funded CERs suggests that outcomes are frequently added to the results in publications that were not listed in the publication's methods section or specified in trial registries. Types of SOR that correspond to those in the ORBIT classification system were uncommon in our cohort. Trial registries frequently revealed inadequate specification of study outcomes, including vague descriptions of outcomes, multiple changes in outcomes between the original and current outcomes listed in the registry, and trials occurring after commencement of subject recruitment. It was therefore difficult to determine if there were significant differences in outcomes between the registry and publication. Neither the study publication nor the registry provided useful information for identifying SAR. In addition, exploration of trial registries was time consuming. The identification of trial registry numbers took multiple steps and trial registries did not have an optimal interface for exploring SOR or SAR.

Our work provides valuable experiences and lessons upon which to start building operational guidance for exploring SOR and SAR in trials and for incorporating those findings into systematic reviews.

Strengths and Limitations of this Study

We explored SOR and SAR using ORBIT in a real-world cohort of RCTs included in AHRQ CERs. The challenges encountered and lessons learned contribute to knowledge directly applicable to future CERs produced by AHRQ and other sponsors and authors. We are not aware of published literature exploring the use of ORBIT, nor attempts to examine the reliability and validity of this classification system. We purport that our suggestions for revisions of existing tools and our recommendations for future research are applicable to work on reporting biases in biomedical studies, well beyond the cohort of trials and CERs examined in this report.

There are a number of limitations to our study. First, we examined only three CERs and a small sample of trials. Thus our findings may not be applicable to systematic reviews of other types of interventions and outcomes, and our findings on SOR and SAR at the trial level may not be generalizable to a wider variety of trials. In addition, our small sample size limited our power to detect significant predictors of SOR.

Second, we faced limitations inherent in trial registration: incomplete trial registration after 2005 and the infrequent registration prior to that year. Thus we could not examine complete cohorts of trials included in meta-analyses within CERs.

Third, studies listed in trials registries may not be representative of all trials published. Perhaps registered trials differ in important reporting and quality characteristics from trials that are not registered. If registered trials are of higher quality, they might have lower rates of SOR and SAR, and thus our cohort may underestimate the frequency of SOR/SAR across trials included in a CER, both registered and unregistered.

Fourth, information in various trial registries differed. Most importantly, much more information is provided in ClinicalTrials.gov than in other registries. Thus all registered trials do not contain the same information and level of detail. For example, as far as we are aware, only ClinicalTrials.gov tracks all changes made to the registration information. Thus comparing and synthesizing the frequency of SOR and SAR across different registries is problematic.

Fifth, we did not examine studies excluded from CERs because they did not report data on our outcome of interest (our "index outcome"). Such studies are excluded by authors of CERs

prior to the synthesis phase of the systematic review, and we did not seek out such studies. These studies may have exhibited SOR leading to exclusion of the study from a CER, and thus our estimates would underestimate the frequency of SOR.

Sixth, during the consensus process between the two assessors of SOR/SAR, our thinking and definitions of SOR/SAR evolved as each assessment brought nuances and new issues, making calculations of inter-rater reliability meaningless in this exploratory work. We were thus unable to calculate inter-rater reliability for ORBIT assessments as we had planned.

And finally, our study does not examine the broader issue of bias in study design and choice of outcomes based on anticipated findings by the trialists. This manipulation of study design is not technically SOR or SAR, which are defined in terms of the selective reporting of outcomes based on results. However, such design issues are a critical source of bias, although beyond the scope of this report.

Challenges Encountered Developing our Study Protocol

The methodology of this project changed significantly as it proceeded. We had initially proposed exploring SOR and SAR in nonrandomized studies (NRS), to develop an ORBIT-like classification system for NRS, and to quantify the effects of SOR/SAR on effect estimates in meta-analyses in CERs. These initial goals proved infeasible, however, and were changed after lengthy discussions among the coauthors of this report. In the absence of registry information or any other prespecified characteristics on NRS, we did not feel that we were able to explore the frequency of SOR/SAR in these study designs. We felt that determination of SOR/SAR based purely on information in the publication would not be useful.

We therefore evolved our objectives to focus on RCTs, with an assessment of the proportion of trials included in CERs with SOR or SAR, determination of the inter-rater reliability of the ORBIT classification system, and an examination of ORBIT's validity when used to assess trial publications (as the tool was intended) when compared with ORBIT assessment using trial registry information. ORBIT proved inadequate for our assessment of the types of SOR and SAR that we identified in included RCTs, and assessment of ORBIT's inter-rater reliability and validity proved infeasible as discussed above.

We also explored quantifying the effect of SOR/SAR in RCTs by comparing the effect estimate from meta-analyses in CERs to estimates obtained using imputed data in studies with missing outcome data (i.e., SOR), using an approach such as that of Copas and Jackson.[14] After discussion among the coauthors of this report, we decided that such analyses would not be useful in view of the multitude of assumptions made in such imputations, and the difficulty determining if SOR/SAR existed in individual studies. In addition, studies contributing to pooled estimates frequently included those published both before and after 2005. The former studies were rarely listed in trial registries and more recent studies were not always registered. The coauthors of this report felt that we should focus on describing SOR/SAR and exploring ORBIT, rather than on problematic quantitative estimates.

We had initially proposed examining SOR and SAR separately. The distinction between the two was often unclear, however. For example, an outcome such as A1c could be analyzed in several ways, including as a continuous measure (with absolute change in percent) or as categorical outcome with various thresholds defining the categories. These presentations of A1c could be considered different outcomes or different analyses. Thus the distinction between SOR and SAR is somewhat arbitrary.

Challenges and Recommendations Regarding Methods and Available Tools for Exploring SOR and SAR

As a result of our exploration of SOR and SAR in a small cohort of RCTs and our use of trial registries to identify SOR/SAR and to assign an ORBIT class, we formulated a number of comments and suggestions on these tools and the potential approaches that systematic reviewers might use (Box 1). Our suggestions are intended to stimulate and provide a basis for future discussions that might ultimately lead to the development of explicit guidance for systematic reviewers.

ORBIT

Challenges Using the ORBIT Classification

We encountered a number of problems when we used ORBIT to assess included trials for SOR and SAR. First, we had difficulty making distinctions among ORBIT categories. Nine categories are a large number to have assessors consider with an adequate degree of reliability. Our team's assessors had difficulty making the distinction between categories E and F, for example, which depends upon an assessment of why an outcome might have been measured but not reported. The distinction between G and H was also difficult because it hinges on clinical judgment.

Second, in some studies it was difficult to determine if an exact p-value could be determined from the data presented in the trial publication, particularly if the assessor did not have a statistical background. Kirkham and colleagues[2] indicate that determination of the ORBIT class can use data calculated indirectly from the results (e.g., an exact p-value calculated from the standard error of an estimate). With adjusted and between-group analyses it was sometimes unclear to the assessor whether an exact p-value could be calculated without access to the underlying dataset.

Third, we had difficulty implementing ORBIT in the situations where there were multiple companion papers for a trial. Kirkham and colleagues[2] note that they determined the ORBIT class for a trial based on all identified publications for that trial. However, authors can legitimately parse publications based on followup intervals which are prespecified in a publication methods section or registry, but where the outcomes are not (yet) available in publications.

Limitations of the ORBIT Classification System

The most important issues that we encountered with ORBIT were not with its implementation, but rather with the limited nature of its intended use and scope. ORBIT was designed for the assessment of SOR/SAR using information within the publication(s) plus clinical judgment.[2] We assert, however, that the most important indicators of SOR and SAR are obtained from sources outside of the published report, such as trial registries or databases of research protocols. ORBIT does not incorporate this additional information. Our efforts to apply ORBIT by comparing information in the trial registry, including results (if any), to the publication (judgments #2 and #3) were met with limited success. Registry information only contributed to assessments of ORBIT classes G and H - when it was unclear in the publication if the outcome was measured. We did not encounter G and H assessments, however, thus had no

data for comparing publication and registry ORBIT classifications and our attempts to validate ORBIT proved infeasible.

In addition, ORBIT only addresses missing or incomplete outcome reporting as it considers outcomes that the reader is led to believe will be in the results section - either because they are mentioned in the methods section, or clinical judgment or information from other studies suggests that an outcome should be reported. The far more frequent scenario in our experience, however, was the reporting of outcomes in publications that had not been mentioned either in the methods section of the publication or in the trial registry. ORBIT also does not include the frequent changes that appear to occur in the primary or secondary outcomes in a study: either when or how an outcome was measured, or the evolution of an outcome initially poorly specified (e.g., "glycemic control") to one with specificity after subject recruitment had been completed. Such changes can only be identified by a careful review of the History of Changes in ClinicalTrials.gov. ORBIT also does not address the issue of the validity of study conclusions based on data reported in a given publication. Studies have been documented to present biased conclusions that correlate with the funders' interests.[26, 28, 29, 33]

Perhaps most importantly, because ORBIT focuses exclusively on the study publication(s), poor writing can lead directly to the appearance of SOR and SAR, where the "selectively" reported outcome is missing simply because of an error or omission or because of constraints on the number of words in the publication. In this situation, apparent SOR and SAR might not lead to actual bias in reporting, if the selective reporting does not relate to the direction and statistical significance of the results. On the other hand, SOR and SAR (and the presence of outcome reporting bias) could be obscured by careful writing.

These limitations in the scope of ORBIT as a tool for assessing SOR/SAR are not criticisms of ORBIT *per se*, as it was not intended to address these additional types of selective reporting of outcomes and analyses. These limitations, however, point to the need for additional tools to identify and assess SOR/SAR in primary studies.

Recommendations for Future Research on the ORBIT Classification System

A new tool is needed for the assessment of SOR/SAR based on all available information, both within and beyond the study publication(s). Such a tool needs to have a limited number of categories and should be broadly applicable to a variety of study interventions and outcomes. This new instrument needs to incorporate both the selection of outcomes from those prespecified, and the addition of new outcomes that differ from those prespecified in the methods section or trial registry. The types of SOR/SAR incorporated in our judgments #2 and #3 (Table 3) may be a useful starting point for development of such a tool. Based on our experiences with the ORBIT classification system, we suggest that the number of categories be relatively small, and judgments based on undocumented information be removed (e.g., the distinction between ORBIT classes E and F, and G and H). Such a tool needs early evaluation, including both reliability and validity assessments.

At a later date, classification systems for nonrandomized studies need to be developed also. We recommend, however, that systems for RCTs be the initial focus for research, as trials are more uniform in design, more information is available in registries and other databases, and nonrandomized studies are often exploratory – purposefully without prespecification of all outcomes and analyses.

If ORBIT continues to be used, research is needed to determine the intra- and inter-rater reliability of the ORBIT classification system. To our knowledge, there are no reliability data on

this instrument. A cohort of individuals with training and experience in the critical appraisal of trials should examine studies from a variety of subject fields and with a variety of index outcomes. Efficacy, effectiveness, and harms outcomes should be examined for SOR/SAR, as the frequency of SOR and SAR and the reliability of the ORBIT tool likely vary across types of outcomes.

In addition, the ORBIT classification system needs to be validated. Kirkham and coauthors[2] performed a limited assessment of the accuracy of ORBIT classes G and H (unclear whether the outcome was measured or not) by comparing their designated class to information provided by the trialists. The sensitivity for predicting that the outcome had been measured was 92 percent and the specificity for predicting that the outcome had not been measured was 77 percent.

Trial Registries

Problems Encountered Using Trial Registries

Trial registries are an important recent advancement in biomedical research, improving public knowledge about ongoing and completed trials, promoting access to research results, and delineating prespecified study methods. Study registries were not as useful as we had anticipated for identifying and characterizing SOR and SAR, however. The following comments focus on ClinicalTrials.gov, as that was the predominant registry that we encountered, other registries that we examined contained less information than did ClinicalTrials.gov, and only this registry contains results.

The list of primary and secondary study outcomes on the summary page for each study in ClinicalTrials.gov often did not meet our needs or it was unclear how best to use the information that was provided. ClinicalTrials.gov indicates the "original" and "current" outcomes, both primary and secondary. This information can be used to quickly determine if there was a change in the primary and secondary outcomes and/or a change in an outcome's designation as primary or secondary. This information was most useful when outcomes were fully specified in terms of how and when they were measured. Frequently, however, outcomes listed in the registry were inadequately specified, particularly the original outcome, thus we could not determine if there was a significant change between the original prespecification and the published outcomes. For example, "A1c" might be specified as the original outcome in the registry, and the current outcome listed as "change in A1c (percent) from baseline to 26 weeks". Although we classified this as a change in the index outcome and considered it potential SOR, the trialists' original outcomes and their motives for the change were unclear.

In addition, the summary page with the "original" and "current" outcomes was only part of the story on changes in specified outcomes over the course of a trial. Additional information could be found by clicking on the "History of Changes" link, where various aspects of study outcomes could change numerous times during the course of the trial, but only the "original" and "current" outcomes are captured on the registry summary page. It is very time-consuming to review the "History of Changes", and difficult to determine if changes represented potential SOR. Although one might suspect SOR/SAR based on the timing of changes, it was not possible to determine if the trialists were purposefully manipulating reported outcomes after they had performed analyses and identified outcomes that they considered favorable.

Safety outcomes appeared to be less precisely specified in the registry, although our experiences were confined to the outcome of mortality for the CER on lipid agents.[17] In

reviewing the other two CERs, we rarely encountered any specification of safety outcomes, although we did not quantify those findings.

The study information sections of trial registries provided no information informing SAR for any of our included studies. No trial registration provided any information on analyses set, subgroups, proposed analyses, or covariates. The registry results frequently provided such information, but since that information was entered after analyses were completed, it was not prespecified and thus did not inform our assessment of SAR.

Registry results, when available, did provide useful information that could be compared to the trial publication. In the majority of RCTs with available results, however, the registry results were less complete than the publication for our index outcome. In other words, additional analyses and results were published that were not presented in the registry, while we rarely encountered the situation where the registry results were more complete. We consider the addition of new outcomes and analyses an important type of SOR/SAR, because they may reflect results favorable to the authors and/or study funders.

Frequently results were not presented in the registry (53 percent of our studies overall). There are legitimate reasons why a trial might not have results posted in a registry.[34] ClinicalTrials.gov, however, does not provide a reason when trial results are not posted in the registry, and only the situation of an ongoing study can be deduced from the registry.

For each included RCT registered with ClinicalTrials.gov, we examined the "History of Changes" page of the registry for relevant changes and encountered a number of challenges. First, the vast majority of changes were of no relevance to our exploration of SOR and SAR, and it was difficult to efficiently identify potentially relevant changes among the plethora of information. The "History of Changes" contains spelling and punctuation corrections, the addition of abbreviations, the addition of citations for publication derived from the trial, the addition of study results, and changes in outcomes, among many other types of changes. Although there are categories of changes in ClinicalTrials.gov, including protocol, recruitment status, location/contact, administrative, and miscellaneous, protocol changes were often combined with administrative changes if they occurred on the same date. Thus the user often had to review irrelevant information when looking for significant protocol changes.

Second, in addition to the amount of information that had to be reviewed, the format for presentation of changes was suboptimal for our purpose. For example, dates for each change in outcome were specified, however, how that date related to important events such as initial or final subject recruitment was not transparent (although all relevant dates could be determined and compared with significant effort). In addition, all changes were formatted as fields for data entry (e.g., "<textblock>") and using abbreviations that were not defined. Such a format is not user-friendly and the meaning was unclear at times. When results were added to the registry, they were listed also on the "History of Changes" page, but without any formatting and with their data entry code: a format of little use to a reader. (The results were clearly presented in the "Study Results" tab, however).

Recommended Changes in ClinicalTrials.gov

From our experience trying to identify and characterize SOR and SAR from information in trial registries, and from using ClinicalTrials.gov in particular, we present a number of suggestions for improving ClinicalTrials.gov and registries in general. These suggestions are based solely on this exploratory work, and they need to be further vetted and validated before being implemented. Our suggestions are of two types: the first related to presentation and ease of

use of the information contained in the registry; and second to the content of the registry. We will address both types, focusing only on ClincialTrials.gov in view of its dominance in our study. Our suggestions are confined to the use of a trial registry for obtaining information on potential SOR and SAR, and are not intended to encompass other reasons for using a trial registry.

Changes in Format of ClinicalTrials.gov

A visual timeline of relevant dates would be useful to users of ClinicalTrials.gov. Currently, numerous dates are presented in narrative form in the introductory page and/or the "History of Changes" page. For many dates there is a logical sequence that should occur if bias in the trial is to be minimized. For example, trial registration should always occur prior to commencement of subject recruitment. The presentation of a timeline of critical points in trial design, implementation, analysis, and publication would allow the user to quickly assess whether the chronology of study design and registration was optimal.

The formatting of the "History of Changes" page needs extensive revision to be of optimal use. The user should be able to review types of changes without having to review categories that are not of interest. For example, if the user is only interested in changes in trial protocol methodology such as a change in the primary outcome or followup interval, the user should not have to review spelling corrections that were entered on the same date as the change in study protocol. The current categories of changes are reasonable; however the approach of organizing by date of change is not. In addition, the field codes should be changed to meaningful labels and headings, or eliminated completely. In addition, all listed changes should represent real changes: situations were encountered where a change was listed, but the text before and after appeared identical.

The "History of Changes" page as currently formatted is not appropriate for reporting results. Results are clearly and efficiently presented in the results section: the "History of Changes" could refer the reader there, with documentation of the date of addition of results to the registry.

Changes in Content of ClinicalTrials.gov

Primary and secondary outcomes need to be specified in detail in the registry: what the outcome is, how it will be measured (if appropriate), and when it will be measured. For example, the outcome A1c needs to be fully qualified, for example "Between-group change in A1c (percent) from baseline to 26 weeks". Listing of vague outcomes such as "glycemic control" should never be permitted in trial registries, particularly after subject recruitment has begun. Precise prespecification of outcomes needs to apply to both "original" and "current", primary and secondary, and benefits and harms outcomes.

Descriptions of changes in outcomes, both primary and secondary, should be readily accessible to the user. Ideally, in addition to the "original" and "current" outcomes now displayed on the trial's main page, a chronology of all changes in outcomes would be clearly presented. The trialist should also be required to indicate the reason for any change in primary or secondary outcomes that were made after subject recruitment commenced. Substantive changes after that point are rarely indicated, and thus when they occur the study authors need to provide a rationale. Population, intervention, and other important subgroups that are part of confirmatory (versus exploratory) analyses should also be specified in the registry prior to commencing subject recruitment.

For the purposes of identifying SAR, it would be very useful for trial registries to include prespecification of selected information such as the general analytic approach (superiority, non-inferiority, equivalence) and the analysis set (e.g., full analysis set and how that was defined). Other information, such as covariates for adjusted analyses would also be useful.

Box

Challenges using the ORBIT classification system

1. Difficulty making distinctions among the ORBIT categories
2. Problems determining if an exact p-value could be determined from the data presented
3. Difficulty applying ORBIT across multiple companion papers

Limitations of the ORBIT classification system

1. Scope and intended use: ORBIT was designed for use with the trial publication(s) only, and not for use with additional information such as that obtained from trial registries or protocols. We found that this additional information could rarely be applied to ORBIT categories.
2. ORBIT only addresses missing outcomes, i.e., those that the reader expects from reading the methods section or from clinical judgment. ORBIT does not include the addition of outcomes or analyses to results section that do not appear to have been specified *a priori*.

Problems encountered using ClinicalTrials.gov

1. Outcomes listed on the summary page were frequently inadequately specified (e.g., "glycemic control")
2. Only "original" and "current" primary and secondary outcomes are listed, although numerous changes in outcomes can have occurred between those two listings.
3. The "History of Changes" tab was difficult to use: labels were unclear and it was difficult to efficiently identify important changes in the outcomes.
4. Safety outcomes were rarely specified in the summary page, and when they were listed often lack specificity.
5. No information of use to detect SAR is provided, such as the analysis set or variables for adjustment.
6. Results were often not posted in the registry and when they were, they were often less complete than those in the publication.

Recommended Changes in ClinicalTrials.gov

1. Changes in format
    a. A visual timeline of important dates would assist the reader in evaluating when registration and changes in outcomes occurred with respect to subject recruitment and data analysis.
    b. The "History of Changes" pages should be formatted to facilitate efficient identification of important protocol changes and data labels should be self-explanatory.
2. Changes in content

a. Primary and secondary outcomes should be specified in detail: what, how, and when.
b. Any changes in outcomes, both primary and secondary, should be listed with the date and rationale on the registry summary page.
c. Information on planned analytic approach (e.g., superiority, non-inferiority), the analysis set (e.g., intention-to-treat, as-treated), and variables planned for adjusted analyses should be provided.

Suggestions for Identifying and Characterizing SOR and SAR

As a result of this exploratory work, we developed a number of specific suggestions for systematic reviewers to use when trying to determine the frequency and effect of SOR and SAR during a systematic review. These suggestions will inform further discussions as guidance is formulated in future. In view of the paucity of evidence on the prevalence and effect of SOR and SAR on effect estimates and conclusions in systematic reviews, our suggestions should also inform future research agendas.

Most importantly, the definition and scope of SOR and SAR need to extend beyond those usually referred to in the systematic review methodology literature: where only a subset of the original outcomes measured and analyzed in a trial are fully reported based on the magnitude of the treatment effect or the statistical significance of selected outcomes.[4] The identification and characterization of SOR and SAR for trials should also include the publication of outcomes and analyses that were not prespecified in the publication methods section or in the trial's registration information. The reporting of this expanded set of outcomes also represents the selective reporting of outcomes or analyses, likely based on the direction, magnitude, or statistical significance of the results.

Suggestions for Systematic Reviewers

The systematic review team needs to be strategic and parsimonious in their efforts to identify SOR and SAR. We suggest that the team consider the following steps in their assessment of potential SOR and SAR:

1. Review trial publication(s) for incompletely reported outcomes in the results section (i.e., ORBIT A through D). Although these classes may be infrequent and may also represent writing and journal styles and constraints on the number of words in a publication, they are important classes to identify. We do not suggest trying to identify and categorize the other ORBIT classes because they reflect for the most part poor writing and not necessarily SOR/SAR.
2. Seek information beyond the study publication only in limited situations when there is a suggestion in the publication or the systematic reviewer has a specific reason to seek out SOR or SAR. This step is potentially very resource intensive and the gain in understanding of SOR and SAR is likely to be minimal. In addition, we were not able to identify predictors of SOR/SAR in our small cohort of studies that had little variation in several of the potential predictors that we examined. Until further research is available to inform such predictors, we offer the following suggestions based on our experience.

a. If there is a concern about SOR in a trial publication, registry results should be searched for. In the absence of the posting of results in the registry, however, the additional information in the registry will likely add little information to what is in the publication.
b. Only search trial registries for RCTs. NRS are not registered frequently enough for searches to be worthwhile.
c. Only search for outcomes of benefit and not harms outcomes unless those are likely to have been prespecified as the trial's primary or secondary outcome. In our experience, useful prespecification of safety outcomes was even less common than for efficacy and effectiveness outcomes.
d. If the trial registry does not include results, we do not recommend further exploration of the summary information provided in the registry.

3. When searching for registered trials, use only the World Health Organization ICTRP portal. This search engine accesses a number of different trial registries at once and although the other registries that we examined were all less useful than ClinicalTrials.gov, the user can quickly identify all registered trials, including those in ClinicalTrials.gov.
4. Comparison of the methods and results sections of trial publications will rarely provide useful information for identifying and characterizing SOR or SAR. The methods section of publications is most likely written after data analyses were completed, and discrepancies between the methods and results sections likely represent poor writing on the part of the trial's authors and discrepancies cannot be attributed with certitude to SOR or SAR.

Conclusions

In this exploratory study of a small cohort of RCTs that were included in three CERs, we determined that trials published in or after 2005 and contributing to meta-analyses, were not consistently listed in trial registries. Publications of trials do not consistently report information on trial registration, and reporting of results in ClincialTrials.gov was inconsistent with no clear reasons for the inconsistencies.

We identified numerous challenges in searching for, and characterizing, SOR and SAR in our cohort of RCTs. We did not find the ORBIT classification tool,[2] designed for the assessment of SOR within trial publications, particularly useful. ORBIT has too many categories with too much ambiguity among those categories. ORBIT classes did not describe the types of SOR and SAR that we most frequently encountered: the addition of outcomes measures, subgroups, and other analyses to published results that were not prespecified in the publication methods section or listed in the registry. We consider this type of SOR and SAR as important as the nonreporting of prespecified outcomes.

Trial registries were of little use in identifying SOR unless trial results were listed in the registry. Registries were of no use in identifying SAR. The presentation and content of ClinicalTrials.gov, the predominant registry for the trials that we examined, could be improved to better assist the systematic reviewer in identifying potential SOR and SAR. Suggestions for improvements in trial registries, and ClinicalTrials.gov in particular, include: 1) a requirement that outcomes be precisely specified; 2) the reasons for any change in primary or secondary outcomes should be provided by the trialist; 3) improved formatting of the History of Changes section; 4) prespecification of the analysis set and general analytic approach be mandated; and 5) efforts to insure that all available results are posted.

Much further research is needed to develop efficient, tailored approaches to identifying and characterizing SOR and SAR in trials starting with an expanded and simplified classification system and changes to trials registries. Without such improvements, the increased time needed by systematic reviewers to try and identify and characterize SOR and SAR are not likely to be worthwhile. Research ultimately needs to guide systematic reviewers in assessing the direction and magnitude of the effects of missing outcomes and analyses on effect estimates and conclusions in systematic reviews.

References

1. Song F, Parekh S, Hooper L, et al. Dissemination and publication of research findings: an updated review of related biases. *Health Technol Assess* 2010;14(8):1-220.
2. Kirkham JJ, Dwan KM, Altman DG, et al. The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. *BMJ*. 2010;340:c365.
3. Sterne J, Egger M, Moher D. Addressing reporting biases. *Cochrane Handbook for Systematic Reviews of Interventions*. 2008;Chichester: John Wiley & Sons Ltd:297-333.
4. Hutton JL, Williamson PR. Bias in meta-analysis due to outcome variable selection within studies. *Applied Statistics 2000*. 2002;49(3):359-370.
5. Higgins J, Altman D. Assessing risk of bias in included studies. *Cochrane Handbook for Systematic Reviews of Interventions*. 2008;Chichester: John Wiley & Sons Ltd:187-242.
6. Dickersin K. The existence of publication bias and risk factors for its occurrence. *JAMA*. Mar 9 1990;263(10):1385-1389.
7. Dickersin K, Chan S, Chalmers TC, Sacks HS, Smith H, Jr. Publication bias and clinical trials. *Control Clin Trials*. Dec 1987;8(4):343-353.
8. Hopewell S, Loudon K, Clarke M. Publication bias in clinical trials due to statistical significance or direction of trial results. *Cochrane Database Syst Rev*. 2009;1:MR000006.
9. Stern JM, Simes RJ. Publication bias: evidence of delayed publication in a cohort study of clinical research projects. *BMJ*. Sep 13 1997;315(7109):640-645.
10. Dickersin K, Olson CM, Rennie D, et al. Association between time interval to publication and statistical significance. *JAMA*. Jun 5 2002;287(21):2829-2831.
11. Dickersin K, Chalmers I. Recognising, investigating and dealing with incomplete and biased reporting of clinical research: from Francis Bacon to the World Health Organisation. *James Lind Library*. 2010.
12. Gotzsche PC. Reference bias in reports of drug trials. *Br Med J (Clin Res Ed)*. Sep 12 1987;295(6599):654-656.
13. Dwan K, Altman DG, Arnaiz JA, et al. Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias. *PLoS ONE*. 2008;3(8):e3081.
14. Copas J, Jackson D. A Bound for Publication Bias Based on the Fraction of Unpublished Studies. *Biometrics*. 2004;60(1):146-153.
15. Bennett W, Wilson L, Bolen S, et al. Oral Diabetes Medications for Adults With Type 2 Diabetes: An Update. Comparative Effectiveness Review No. 27. (Prepared by Johns Hopkins University Evidence-based Practice Center under Contract No. 290-02-0018.) AHRQ Publication No. 11-EHC038-EF. Rockville, MD: Agency for Healthcare Research and Quality. March 2011. Available at: www.effectivehealthcare.ahrq.gov/reports/final.cfm
16. MacLean C, Alexander A, Carter J, et al. Comparative Effectiveness of Treatments To Prevent Fractures in Men and Women With Low Bone Density or Osteoporosis. Comparative Effectiveness Review No. 12. (Prepared by Southern California/RAND Evidence-based Practice Center under Contract No. 290-02-0003). Rockville, MD: Agency for Healthcare Research and Quality. December 2007. Available at: www.effectivehealthcare.ahrq.gov/reports/final.cfm.
17. Sharma M, Ansari M, Soares-Weiser K, et al. Comparative Effectiveness of Lipid-Modifying Agents. Comparative Effectiveness Review No. 16. (Prepared by the University of Ottawa Evidence-based Practice Center under contract No. 290-02-0021.)

Rockville, MD: Agency for Healthcare Research and Quality. September 2009. Available at: www.effectivehealthcare.ahrq.gov/reports/final.cfm.
18. World Health Organization. International Clinical Trials Registry Platform (ICTRP). 2011; http://www.who.int/ictrp/en/. Accessed August 9, 2011.
19. Food and Drug Administration Modernization Act of 1997. 1997;Pub L No. 105-115.
20. Food and Drug Administraion Revitalization act, Title III, Subtitle C, S 1082. 2007.
21. DeAngelis CD, Drazen JM, Frizelle FA, et al. Clinical trial registration: a statement from the International Committee of Medical Journal Editors. *JAMA*. Sep 15 2004;292(11):1363-1364.
22. Zarin DA, Ide NC, Tse T, Harlan WR, West JC, Lindberg DAB. Issues in the registration of clinical trials. *JAMA*. May 16 2007;297(19):2112-2120.
23. International Committee of Medical Journal Editors. 2009; http://www.icmje.org/faq_clinical.html. Accessed August, 9, 2011.
24. United States Code. US Public Law 110–85: Food and Drug Administration Amendments Act of 2007. 2008; http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=110_cong_public_laws&docid=f:publ085.110. Accessed August 9, 2011.
25. FDA Failed to Enforce Law Requiring Drugmakers to Disclose Test Data- WashPost. 2004. http://www.ahrp.org/infomail/04/07/06.php. Accessed January 3, 2012.
26. Als-Nielsen B, Chen W, Gluud C, Kjaergard LL. Association of funding and conclusions in randomized drug trials: a reflection of treatment effect or adverse events? *JAMA*. Aug 20 2003;290(7):921-928.
27. Bekelman JE, Li Y, Gross CP. Scope and Impact of Financial Conflicts of Interest in Biomedical Research. *JAMA*. January 22, 2003 2003;289(4):454-465.
28. Lexchin J, Bero LA, Djulbegovic B, Clark O. Pharmaceutical industry sponsorship and research outcome and quality: systematic review. *BMJ*. May 31 2003;326(7400):1167-1170.
29. Sismondo S, Sismondo S. How pharmaceutical industry funding affects trial outcomes: causal structures and responses. *Soc Sci Med*. May 2008;66(9):1909-1914.
30. Raz I, Chen Y, Wu M, et al. Efficacy and safety of sitagliptin added to ongoing metformin therapy in patients with type 2 diabetes. *Curr Med Res Opin*. Feb 2008;24(2):537-550.
31. Scott R, Loeys T, Davies MJ, Engel SS, Sitagliptin Study G. Efficacy and safety of sitagliptin when added to ongoing metformin therapy in patients with type 2 diabetes. *Diabetes Obes Metab*. Sep 2008;10(10):959-969.
32. Rosenstock J, Rood J, Cobitz A, Biswas N, Chou H, Garber A. Initial treatment with rosiglitazone/metformin fixed-dose combination therapy compared with monotherapy with either rosiglitazone or metformin in patients with uncontrolled type 2 diabetes. *Diabetes Obes Metab*. Nov 2006;8(6):650-660.
33. Golder S, Loke YK. Is there evidence for biased reporting of published adverse effects data in pharmaceutical industry-funded studies? *Br J Clin Pharmacol*. Dec 2008;66(6):767-773.
34. About the ClinicalTrials.gov Results Database. 2011; http://clinicaltrials.gov/ct2/info/results. Accessed January 5, 2012.

Abbreviations and Acronyms

ARB, analysis reporting bias
AHRQ, Agency for Healthcare Research and Quality
CERs, reviews of comparative effectiveness
EPC, Evidence-based Practice Centers
ICMJE, International Committee of Medical Journal Editors
ICTRP, International Clinical Trials Registry Platform
NA, not applicable
NR, not reported
ORB, outcome reporting bias
ORBIT, Outcome Reporting Bias in Trials
RCT, randomized controlled trial
SAR, selective outcome reporting
SOR, selective analysis reporting
WHO, World Health Organization