Methods Research Report

Developing and Testing a Tool for the Classification of Study Designs in Systematic Reviews of Interventions and Exposures

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

Contract No. 290-02-0023

Prepared by:
University of Alberta Evidence-based Practice Center
Edmonton, Alberta, Canada

Investigators
Lisa Hartling, M.Sc.
Kenneth Bond, B.Ed., M.A.
Krystal Harvey, B.Sc.
P. Lina Santaguida, Ph.D.
Meera Viswanathan, Ph.D.
Donna M. Dryden, Ph.D.

AHRQ Publication No. 11-EHC007-EF
December 2010

This report is based on research conducted by the University of Alberta Evidence-based Practice Center (EPC) under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No. 290-02-0023). The findings and conclusions in this document are those of the author(s), 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 clinicians, employers, policymakers, and others make informed decisions about the provision of health care services. This report is intended as a reference and not as a substitute for clinical judgment.

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.

This document was written with support from the Effective Health Care Program at AHRQ. None of the authors has a financial interest in any of the products discussed in this document.

Suggested citation: Hartling L, Bond K, Harvey K, Santaguida PL, Viswananthan M, Dryden DM. Developing and Testing a Tool for the Classification of Study Designs in Systematic Reviews of Interventions and Exposures. Agency for Healthcare Research and Quality; December 2010. Methods Research Report. AHRQ Publication No. 11-EHC-007. Available at http://effectivehealthcare.ahrq.gov/.

The investigators have no relevant financial interests in the report. The investigators have no employment, consultancies, honoraria, or stock ownership or options, or royalties from any organization or entity with a financial interest or financial conflict with the subject matter discussed in the report.

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. 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.

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.gov.

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

Jean Slutsky, P.A., M.S.P.H.
Director, Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Stephanie Chang, M.D., M.P.H.
Director, EPC Program
Agency for Healthcare Research and Quality

Capt. Karen Lohmann Siegel, P.T., M.A.
Task Order Officer
Agency for Healthcare Research and Quality

Acknowledgments

We are very grateful to the following individuals from the University of Alberta Evidence-based Practice Center who were involved in testing the taxonomies and providing feedback: Ahmed Abou-Setta, Liza Bialy, Michele Hamm, Nicola Hooton, David Jones, Andrea Milne, Kelly Russell, Jennifer Seida, and Kai-On Wong. We thank Ben Vandermeer for conducting the statistical analyses. We thank Karen Siegel from AHRQ for her input and advice during the course of the project. We are also appreciative of the individuals who provided sample taxonomies and of the members of other Evidence-based Practice Centers who provided the studies that were used for testing.

Structured Abstract

Background. Classification of study design can help provide a common language for researchers. Within a systematic review, definition of specific study designs can help guide inclusion, assess the risk of bias, pool studies, interpret results, and grade the body of evidence. However, recent research demonstrated poor reliability for an existing classification scheme.

Objectives. To review tools used to classify study designs; to select a tool for evaluation; to develop instructions for application of the tool to intervention/exposure studies; and to test the tool for accuracy and interrater reliability.

Methods. We contacted representatives from all AHRQ Evidence-based Practice Centers (EPCs), other relevant organizations, and experts in the field to identify tools used to classify study designs. Twenty-three tools were identified; 10 were relevant to our objectives. The Steering Committee ranked the 10 tools using predefined criteria. The highest-ranked tool was a design algorithm for studies of health care interventions developed, but no longer advocated , by the Cochrane Non-Randomised Studies Methods Group. This tool was used as the basis for our classification tool and was revised to encompass more study designs and to incorporate elements of other tools. A sample of 30 studies was used to test the tool. Three members of the Steering Committee developed a reference standard (i.e., the "true" classification for each study); 6 testers applied the revised tool to the studies. Interrater reliability was measured using Fleiss' kappa (κ) and accuracy of the testers' classification was assessed against the reference standard. Based on feedback from the testers and the reference standard committee, the tool was further revised and tested by another 6 testers using 15 studies randomly selected from the original sample.

Results. In the first round of testing the inter-rater reliability was fair among the testers (κ = 0.26) and the reference standard committee (κ = 0.33). Disagreements occurred at all decision points in the algorithm; revisions were made based on the feedback. The second round of testing showed improved interrater reliability (κ = 0.45, moderate agreement) with improved, but still low, accuracy. The most common disagreements were whether the study was "experimental" (5/15 studies) and whether there was a comparison (4/15 studies). In both rounds of testing, the level of agreement for testers who had completed graduate-level training was higher than for testers who had not completed training.

Conclusion. Potential reasons for the observed low reliability and accuracy include the lack of clarity and comprehensiveness of the tool, inadequate reporting of the studies, and variability in user characteristics. Application of a tool to classify study designs in the context of a systematic review should be accompanied by adequate training, pilot testing, and documented decision rules.

Contents

Figure

Tables

Appendixes

Developing and Testing a Tool for the Classification of Study Designs in Systematic Reviews of Interventions and Exposures

Executive Summary

Introduction

Emphasis on understanding the comparative effectiveness of health care interventions has increased attention on the applicability of research. As a result, systematic reviews aiming to inform clinical practice have expanded beyond randomized controlled trials, which have limited generalizability, to include nonrandomized studies. However, different nonrandomized study designs differ in their relative strengths and weaknesses. A standard nomenclature and taxonomy for categorizing these nonrandomized studies may help promote a common language and understanding among Evidence-based Practice Centers (EPCs) and other systematic reviewers regarding the inherent strengths and weaknesses of particular study designs.

The development of a standard classification tool may help to streamline and facilitate scoping of evidence and making decisions as to what kind of evidence should be considered for any particular review. Accurate classifications by study design are also important for the efficient and accurate communication of the results of a systematic review.

The primary objectives of this Methods Research Paper were:

1. To identify classification tools that are currently used by systematic reviewers and other researchers to identify studies according to design.
2. To select a classification tool for modification and evaluation.
3. To develop instructions, including an algorithm and decision rules, for application of the modified tool to studies of interventions and exposures.
4. To test the tool and accompanying instructions for concurrent validity and interrater reliability.

Methodology

A Steering Committee of seven members from three EPCs (University of Alberta, McMaster University, RTI International–University of North Carolina [UNC] at Chapel Hill) and the Agency for Healthcare Research and Quality was formed to guide the methods and direction of the project.

Objective 1

A sample of classification tools was compiled by contacting representatives from all EPCs and other relevant organizations, as well as individuals with expertise in this area identified by the Steering Committee. Individuals were contacted by e-mail and asked to identify any taxonomies, guidelines, or other systems used to classify study designs. All EPCs were asked to provide examples of intervention or exposure studies for which the assignment of study design had been problematic.

Objective 2

The Steering Committee collaboratively developed the following criteria to rank the tools collected in Objective 1:

- Ease of use (e.g., contains a logic that users can readily follow).
- Unique classification for each study design (no overlap).
- Unambiguous nomenclature and decision rules/definitions (if applicable).
- Comprehensiveness (complete in terms of range of study designs).
- Potential to allow for identification of threats to validity and provide a guide to strength of inference.
- Development by a well-established organization.

Five members of the Steering Committee independently rated the tools collected in Objective 1 and ranked them according to their ability to satisfy the six criteria. The rankings were individually presented in a teleconference and an overall ranking was discussed until consensus was selected on the top ranked tool.

Objective 3

Three members of the Steering Committee used an iterative process to test and modify the tool selected in Objective 2. Decisions to modify the tool were based on the collective experience of the Steering Committee members. After several repetitions of this process, the final version of the modified tool was used to produce a glossary of study design definitions and related concepts. The tool and accompanying glossary were sent to all members of the Steering Committee for review.

Two members of the Steering Committee who were not involved in producing the reference standard selected 30 sample studies from the pool of studies collected from the EPCs. Studies were selected to cover most of the key decision nodes within the algorithm in order to ensure adequate testing of the tool.

Objective 4

Six individuals from the University of Alberta EPC (UAEPC) used the tool to assess the designs of the 30 sample studies with minimal additional instruction or direction. Testers were told that it would take approximately 5 to 10 hours to categorize the 30 studies and were asked to complete the assignment over a 2-week period.

Concurrently, three members of the Steering Committee independently applied the tool to the same 30 studies to develop the reference standard (i.e., the "true" classification for each study). Disagreements were resolved by discussion and consensus.

Overall interrater reliability was calculated using Fleiss' kappa (κ). Interrater reliability was calculated separately for the reference standard raters and the testers, as well as based on formal training of the testers (completed relevant graduate training vs. currently enrolled in graduate training). Accuracy of the testers was measured against the reference standard. The mean time taken to classify the sample of studies and the mean time taken per study were also calculated.

After the first round of testing, the tool was modified further based on the results of semistructured interviews with the testers to ensure the tool's usefulness and usability in the

context of a systematic review. Six testers from the UAEPC participated in a second round of testing, using a random sample of 15 studies from the 30 studies used for the first round of testing. Three of the testers had been involved in the first round of testing and three had not. The same analyses were conducted for the second round of testing.

Results

Objective 1

We contacted 31 organizations or individuals to identify taxonomies/study design classification tools. The Steering Committee reviewed the 23 tools that were received; 10 were considered relevant to the context of study design classification in systematic reviews.

Objective 2

The three top-ranked tools were:

- A design algorithm for studies of health care interventions (DASHCI) developed by the Cochrane Non-Randomised Studies Methods Group (NRSMG). Note that this tool is no longer advocated by the NRSMG,
- A tool developed by the American Dietetic Association (ADA).
- A tool developed by the RTI-UNC EPC.

The three tools were all algorithms (i.e., they provided a logical sequence of "yes or no" decisions to make when classifying studies). None of the algorithms covered the range of study designs that systematic reviewers might encounter when developing EPC reports. Further, the study nomenclature was inconsistent among the algorithms. The DASHCI algorithm was considered the most preferred tool and was used as the basis for further development.

Objective 4

Reference Standard

Three members of the Steering Committee developed the reference standard. Each member independently applied the flow diagram to assign design labels to the 30 studies. The initial agreement was fair (κ=0.33). Disagreements were resolved through discussion and consensus.

Disagreements occurred at most decision points in the algorithm. The area that created the greatest uncertainty and disagreements for the reference standard raters was the decision node "Was there a single cohort?" The initial decision node ("Was there a comparison?") was also a source of disagreement.

Test 1

Tester characteristics. Six staff members at the UAEPC with varying levels of training and experience in systematic reviews tested the modified taxonomy.

Agreement. There were no studies for which all six testers agreed on the classification (Table A). Five of six testers agreed on the classification of seven studies, four agreed on five studies,

three agreed on nine studies, two agreed on eight studies. The overall level of agreement was fair (κ=0.26). The levels of agreement for testers who had completed vs. those who were undertaking graduate-level training were fair (κ=0.38) and slight (κ=0.17), respectively.

Disagreements occurred at all decision points in the taxonomy; however, testers identified the determination of whether there was a single cohort as particularly problematic. The testers also said that certain terminology in the flow diagram was unclear (e.g., "group" vs. "cohort") and that disagreements arose due to poor study reporting. There was some variation in the manner in which testers used the flow diagram (e.g., whether or not they used the glossary, working forward vs. backward through the algorithm).

Accuracy of testers compared to reference standard. There were no studies for which all six testers agreed with the reference standard, and there was wide variation in the testers' accuracy of classification.

Test 2

Tester characteristics. Six staff members at the UAEPC with varying levels of training and experience in systematic reviews were involved in the second round of testing. Three of the testers had been involved in the first round of testing, and three of the testers had no previous involvement with the project or knowledge of the taxonomy being tested.

Agreement. There were three studies for which all six testers agreed on the classification (Table A). Five of six testers agreed on two studies, four agreed on six studies, three agreed on two studies, and two agreed on two studies. The overall level of agreement was considered moderate (κ=0.45). The levels of agreement for testers who had completed vs. those undertaking graduate-level training were moderate (κ=0.45) and fair (κ=0.39), respectively.

Accuracy of testers compared to reference standard. There were three studies for which all six testers agreed with the reference standard, but there was wide variation in the testers' accuracy of classification.

The least common agreement occurred at four key decision nodes: whether the study was "experimental" (5/15 studies), whether there was a comparison (4/15 studies), whether the assessment of exposure and outcome was prospective or retrospective, and whether the intervention or exposure and outcome data were gathered concurrently (2/15 studies).

Table A. Results of taxonomy testing

	Test 1—30 studies; 6 testers	Test 2— 15 studies; 6 testers
Overall agreement	κ=0.26	κ=0.45
Item agreement (number of studies):		
6/6 testers agreed	0	3 (20%)
5/6 testers agreed	7 (23%)	2 (13%)
4/6 testers agreed	5 (17%)	6 (40%)
3/6 testers agreed	9 (30%)	2 (13%)
2/6 testers agreed	8 (27%)	2 (13%)
No agreement	1 (3%)	0
Number of testers with same design classification as reference standard:		
6	0	3 (20%)
5	6 (20%)	2 (13%)
4	4 (13%)	3 (20%)
3	7 (23%)	1 (4%)
2	3 (10%)	2 (13%)
1	7 (23%)	2 (13%)
0	3 (10%)	2 (13%)

Discussion

We identified over 20 tools and selected 1 for modification and testing. The final testing of the modified tool showed moderate agreement among six testers and low accuracy against the predetermined reference standard. The moderate level of agreement is consistent with that observed in a previous study.

There are a variety of reasons for the moderate and low levels of agreement and accuracy observed in our study. The results likely reflect issues with the taxonomy itself, as well as attributes of the studies that were selected for testing. The studies used during testing were identified and selected because they had posed challenges for previous reviewers with respect to their design classification. Agreement might be better with a sample of studies that is more representative of the studies that would be included in a systematic review. Further, the sample of studies that we tested covered a wide range of topics. If the studies had been on the same topic, which would be the case in a systematic review, there might have been greater reliability. One of the main reasons that the selected studies were difficult to classify was poor reporting within the studies, which resulted in the need for testers to make assumptions (e.g., whether the timing of a study was prospective or retrospective). We also found classification challenging when there were discrepancies between the intent of the investigator and the conduct of the study, between the design and how data were analyzed, and between the investigators' initial plan and study implementation.

Shortcomings of the taxonomy itself also resulted in moderate agreement. Many of the decision points were challenging. For example, in one-third of the studies, there were discrepancies as to whether or not the study was truly "experimental." Identifying "quasi-experimental" studies is challenging, as the investigator has some control over certain aspects of design and study execution, and the study may not be considered either purely experimental (a "trial") or purely observational. This area of study design needs to be more clearly reflected in the taxonomy, and clear guidelines are needed for interpreting the extent of control an investigator has. The practical repercussion of this uncertainty in classification is that some

"quasi-experimental" studies (e.g., before-after or controlled before-after studies) may incorrectly be classified as trials; hence, their validity may be exaggerated and the results given too much weight in the context of a systematic review. One design that is particularly problematic has been variously referred to as an "uncontrolled trial" or "single-arm trial." It is our opinion that this design should not be considered a "trial" because of the serious risk of bias associated with the lack of a control or comparison group. Consequently, studies with such a design should be considered "before-after" studies, and our taxonomy was designed to channel them toward this classification.

Other decision nodes that yielded inconsistent results concerned whether there was a comparison, whether the study was experimental, and whether the data collection was prospective or retrospective. Several factors may have contributed to this inconsistency, including a lack of clarity in the questions posed in the algorithm, the testers' relevant background knowledge, the testers' experience or training, and the inconsistent use of design terminology among the studies. While we provided a glossary in an attempt to offer standard definitions and clarity in terminology, there may have been shortcomings with the glossary that created confusion, including ambiguity and inconsistency in terms and definitions. Testers who had completed relevant graduate-level training had greater agreement than those who were undertaking graduate-level training.

We observed a fair level of agreement among the reference standard raters as well. The three reference standard raters had substantial expertise in research methods and systematic reviews. The low level of agreement among these raters may reflect the more general complexities of study designs and the challenge of including all design considerations in a single flow diagram.

Variability in classification of studies may also reflect differences in how individuals applied or worked through the taxonomy. For example, some testers worked backward or backtracked in order to classify the studies according to what they felt was the most appropriate description. The testers also used the glossary accompanying the tool to varying degrees.

The difficulties in interpreting study design labels and the consequent difficulties in reaching agreement in assigning these labels to individual studies are consistent with those of other researchers. These issues have led some authors to direct systematic reviewers to focus on features of designs rather than on design labels when assessing studies for inclusion and evaluating potential risk of bias. The use of a taxonomy may provide greater transparency and consistency to the process.

Implications for Practice

The appropriate classification of studies by design or by design features is important in a systematic review in order to guide the selection of studies, the assessment of the risk of bias, the analysis of study results, the interpretation of results, and the grading of the body of evidence. There is a clear need for consistent use of terminology and study design labels, as well as a clear understanding of the terminology used in a particular field by those undertaking a systematic review in that field. We believe that a tool such as the one developed and tested in this study would be useful to guide this process, although the application of the tool requires several considerations in order to optimize agreement and reliability among reviewers. First, training in research methods, as well as in the use of the tool, is essential. Pilot testing the tool in the context of each review is highly recommended. Second, decision rules are needed for different fields of

research or review topics. Specifically, there need to be clear decisions about how to handle a lack of clarity in study reporting. We recommend that when the response to a question in the taxonomy is unclear, the reviewer assume that the condition was not met. Documentation of the decision rules will allow for consistency and transparency. Users of the algorithm need to use standardized definitions of study designs and design features.

Future Research

The tool developed and tested in this study serves as a basis for further research. Future research is needed to evaluate the tool within the context of a real systematic review, as well as to conduct more indepth testing for specific study designs that are difficult to classify. Additional critical review and refinement of the accompanying glossary are needed. We provide some preliminary data on factors that might create differences in reliability across individuals, including varied experience, training, and education. Further research is needed to provide more definitive results about these various factors and how they impact the performance of the tool. Our experiences also provide direction for methods to be employed in subsequent work, such as development of the reference standard (e.g., involving senior researchers with epidemiological training) and contacting authors for clarification when methods within the individual studies are unclear.

Conclusions

We developed and tested a taxonomy for the classification of study designs. The level of agreement among six testers was moderate and the accuracy against a reference standard was low. There are a number of explanations for the observed reliability and accuracy, including shortcomings of the taxonomy and accompanying glossary, inadequate reporting of the studies, and differences in tester characteristics. Application of such a tool in the context of a systematic review should be accompanied by adequate training, pilot testing, and documented decision rules. This study demonstrates that systematic testing and refinement enhance the reliability of the tool. At the study level, clear reporting, adherence to published reporting guidelines, and appropriate and consistent use of design terminology should be enforced.

Table B is a summary of study findings, implications for practice, and directions for further research.

Table B. Summary of findings, implications for practice, and directions for future research

Challenges in classifying study designs

- Poor reporting resulting in lack of clarity.
- Mixed methods utilized in the same study, or more than one question or hypothesis that employ different methods being investigated in the same study.
- Discrepancies between intent of investigator and study conduct.
- Discrepancies between study design and how data were analyzed.
- Discrepancies between investigators' original plan and study implementation.
- Inconsistent, inaccurate, or imprecise use of design terminology.

Implications for practice

- Clear directions and decision rules specific to topic area or systematic review questions.
- Adequate training and relevant education for those applying the tool.
- Pilot testing of tool and instructions in the context of specific topic area.
- At the study level, critical need for clear reporting, adherence to published reporting guidelines, and appropriate use of design terminology.

Future research

- Testing the final revised tool in different topic areas, within the context of a real systematic review, and

for specific study designs that are difficult to classify.
- Further investigation of the effects of training, education, and experience on reliability.
- Refinement of the methods employed in this study, including development of the reference standard and contacting authors of primary studies for clarification.

Methods Research Paper

Chapter 1. Introduction

Background

Evidence-based Practice Center (EPC) evidence reports and technology assessments aim to review the relevant scientific literature and to evaluate clinical and behavioral interventions, prognostic and diagnostic tools, health care utilization, and other health care organization and delivery issues.[1] These reports are used for informing and developing coverage decisions, quality measures, educational materials and tools, guidelines, and research agendas. The reports are based on rigorous, comprehensive syntheses (systematic reviews) and analyses (meta-analyses) of the scientific literature[2] on topics relevant to clinical, social science/behavioral, economic, and other health care organization and delivery issues. In most circumstances, controlled trials, and especially randomized controlled trials (RCTs), are the study design least likely to produce biased estimates of the effect of an intervention;[3] however, alternative study designs are often needed to capture information important to clinicians and other end-users of the reports, particularly when controlled trials are lacking, not appropriate for the outcomes or questions of interest (e.g., long-term outcomes, rare outcomes, adverse effects), or not generalizable to a broader population.[4] In the context of such reviews, the appropriate classification of studies according to their design or design features is important in order to guide (1) decisions around inclusion, (2) the assessment of methodological quality or risk of bias, (3) the combining of study results in a narrative synthesis or by statistical pooling and (4) grading the body of evidence.

The Use of Nonrandomized Study Designs

RCTs are considered the gold standard for judging therapeutic efficacy and effectiveness;[5,6] however, RCTs are often unnecessary, inappropriate, impossible or inadequate to address particular research questions.[6-8] Given the wide range of topics addressed by EPC reports, researchers frequently need to include nonrandomized studies in order to provide "a more detailed picture of our current knowledge and its limitations for clinicians and policymakers."[9] This is especially true for outcomes that RCTs may not be adequately designed to address (e.g., adverse effects)[10] and in areas in which few RCTs have been conducted, such as devices and surgical procedures (where RCTs account for less than 10 percent of the evidence base)[11] or educational interventions (e.g., medical education[12]).

In the past, many EPC reports have restricted evidence to RCTs when addressing effectiveness questions in order to protect against sources of bias in other study designs even though the extent of bias associated with different nonrandomized designs varies and the direction and magnitude of bias can be unpredictable.[5,13-15] When there are few well-conducted RCTs, however, end-users are often dissatisfied if a report concludes that the evidence is insufficient without considering other study designs, especially since these other designs may be the very studies that are influencing current practice and policy debates.[16] As a result, a past criticism of EPC reports is that they are too restrictive in their consideration of evidence.[16]

The fact that a high proportion of published intervention studies use nonrandomized designs is taken as evidence that this research is valued by clinicians.[17] Moreover, some researchers believe that including nonrandomized designs lacking control groups may increase

the evidence base concerning a health technology and strengthen the credibility of a review. Even if evidence to evaluate the effectiveness or harm is lacking, nonrandomized studies may provide evidence to guide the development of future research questions. There is also value in reviewing nonrandomized studies to clearly describe their limitations and to recommend the types of studies that would provide better evidence.[12,16]

There is evidence that EPC reports that address the efficacy or effectiveness of a clinical intervention are becoming more inclusive in terms of designs other than RCTs.[4] In addition, there is increasing attention and funding directed toward comparative effectiveness reviews (CERs) and an accompanying broadening of perspective;[18] the topics identified by the AHRQ for CERs are broad and require practice-based evidence, which, if available, is often collected through nonrandomized studies. A particularly important and related principle is the move from efficacy to effectiveness.[19] These changes in the focus of EPC work heighten the urgency for evaluating current practices in evidence synthesis for nonrandomized studies. The starting point in such a synthesis is the classification of study design.

Classification of Study Designs

The diversity of study designs and the similarities among them present challenges to reviewers who wish to clearly classify designs or design features for the purpose of assessing the strength of the evidence regarding a particular intervention. Textbooks in systematic reviews,[5] epidemiology,[20] and social science research[21] as well as health technology assessment (HTA) reports[7] provide detailed descriptions of study designs commonly used in those areas; however, none of these resources on its own provides a comprehensive treatment of all the design types encountered when evaluating studies within EPC reports, nor do any provide convenient summaries or decision rules for distinguishing among designs. Perhaps most troubling for systematic reviewers is the fact that the terminology and classification systems used to describe different nonrandomized studies are inconsistent.[9,14] This inconsistency is problematic because researchers often make decisions about research approaches and the interpretation of results on the basis of study design classification and labels. The use of a variety of similar and ambiguous study classifications (e.g., a "prospective study") may lead to low sensitivity in study identification and inaccurate quality assessment of the conduct of studies.[22] This results in uncertainty about the study designs that were used to address the research questions and how the evidence provided by these studies should be weighed.

An important goal of the EPC program is to advance the methods for conducting and reporting systematic reviews.[16] Though much research has focused on systems for grading evidence,[23,24] and a variety of approaches have been developed for evaluating the quality of nonrandomized studies, less attention has been paid to the importance of developing a standard classification scheme or nomenclature system or an algorithm to correctly classify study designs. Research has demonstrated poor agreement among reviewers using a "traditional" study design classification scheme in epidemiology (called a "taxonomy" by the researchers) to classify a set of studies in the field of low back pain.[14] The assignment of study design labels was also found to be unreliable, even when specific instructions and definitions were provided.[14]

The appropriate classification and assessment of a study's design relies heavily, though not exclusively, on the adequacy of reporting by a study's authors. Without an indication of the elements considered crucial for labelling studies, assessing the appropriateness of the study design, and assessing a study's strengths and potential weaknesses, study authors are likely to

omit these important features. For example, the Strengthening of Reporting of Observational Studies (STROBE) statement,[6] a guideline that promotes transparency in reporting analytic observational studies, encourages authors to "indicate the study design with a commonly used term" and "to present key elements of study design".[6] However, the document provides limited guidance on the elements needed to adequately conduct the above processes. A standard classification tool may help authors to identify these crucial study design features, thus improving the transparency of reporting.

Finally, the development of a standard classification tool may help to streamline and facilitate the process of scoping the available scientific literature and of deciding what methodologies should be considered for any particular review. Accurate study design classifications are also important for the efficient and accurate communication of the results of a systematic review. A classification tool may also complement other approaches to defining study designs or design features that are important to consider when evaluating the strength of evidence.

Objectives of the Project

The University of Alberta Evidence-based Practice Center (UAEPC) undertook this project to identify a valid and reliable tool for the classification of randomized and nonrandomized studies of interventions and exposures that could be used in the conduct of systematic reviews.

The specific objectives were

1. To identify classification tools that are currently used by systematic reviewers and others to identify studies according to design.
2. To select a classification tool for modification and evaluation.
3. To develop instructions, including an algorithm and decision rules, for application of the modified tool to studies of interventions and exposures.
4. To test the tool and accompanying instructions for concurrent validity and inter-rater reliability.

Chapter 2. Methods

The Steering Committee for this project was composed of Kenneth Bond (UAEPC), Donna Dryden (UAEPC), Lisa Hartling (UAEPC), Krystal Harvey (UAEPC), P. Lina Santaguida (McMaster EPC), Meera Viswanathan (RTI-UNC EPC), and Karen Siegel (AHRQ). The roles and credentials of individual members are listed in Appendix A.

A revision of the second objective made by the Steering Committee meant that the methods implemented during the course of this project differed slightly from those originally specified in the project work plan. Specifically, we had planned to select two taxonomies for further testing; however, none of the tools we identified incorporated the full range of nonrandomized study designs that commonly appear in systematic reviews of interventions and exposures. Consequently, instead of selecting two tools for further testing, the Steering Committee decided to select one tool and to modify it to better meet the needs of systematic reviewers.

Identification of Taxonomies (Objective 1)

In order to compile a sample of classification tools, representatives from all EPCs and other research organizations, as well as individuals with expertise in this area, were identified by the Steering Committee. Individuals were contacted by email and asked to send any taxonomies, guidelines, or other systems used to classify study designs. The organizations contacted are listed in Appendix B and the sample letter of request appears in Appendix C. The 15 EPCs were also asked to provide examples of intervention or exposure studies where the assignment of study design had been problematic. The collection of these studies continued until the testing of the modified tool began.

Selection of Taxonomies (Objective 2)

Five members of the of Steering Committee (DD, PS, MV, KB, LH) collaboratively developed six criteria for assessing the classification tools based on the experience of the EPCs and on individual knowledge regarding the characteristics considered desirable in a classification tool. The six criteria were:

- ease of use (e.g., contains a logic that users can readily follow);
- unique classification for each study design (no overlap);
- unambiguous nomenclature and decision rules/definitions (if applicable);
- comprehensiveness (complete in terms of the range of study designs);
- potentially allows for identification of threats to validity and provides a guide to strength of inference; and,
- developed by a well-established organization.

The five raters independently reviewed the tools collected in Objective 1 and ranked them according to their ability to satisfy the six criteria. The rankings were individually presented in a teleconference and an overall ranking was discussed until consensus was achieved between the five members.

Development of the Taxonomy Tool (Objective 3)

The top-ranked tool was modified to incorporate relevant elements of the other tools and to ensure comprehensiveness in terms of study designs. Three members of the Steering Committee (KB, DD, LH) modified the tool using an iterative process. Ten studies[25-34] identified in Objective 1 were purposefully selected to highlight some challenges in study design classification and to represent a range of designs. These studies and challenges were considered in modifying the tool. Decisions to modify the tool were based on the collective experience of the Steering Committee members. After several iterations of this process, the final version of the modified tool was used to produce a glossary of study design definitions and related concepts. The tool and accompanying glossary were sent to all members of the Steering Committee for review.

In preparation for testing the tool, 30 studies were selected from the pool of 71 studies that were received from three EPCs (47 studies from Minnesota EPC, 16 from UAEPC, and 8 from McMaster University EPC). Two members of the Steering Committee (LH, KB) who were not involved in developing the reference standard (see below Objective 4) selected the sample studies that would be used to test the new tool. Studies were selected to cover all of the key decision nodes within the algorithm in order to ensure adequate testing of the tool. Further, at least one study of each design was included in the sample pool. For the majority of designs (all but two—case-control and cross-sectional) at least two sample studies were included. Additional studies were included for the designs that we felt to be more commonly encountered in systematic reviews of therapeutic interventions and exposures (e.g., before-after studies, controlled before-after studies, and cohort studies). The selection of studies was based on the design determined by LH and KB, which was not consistent in all cases with the final reference standard classification. According to the reference standard classification, all decision nodes and all but two designs (prospective cohort, nested case control) in the taxonomy were represented.

Taxonomy Testing and Development of Reference Standard (Objective 4)

Six individuals from the UAEPC tested the tool by applying it to the sample of 30 studies. The number of testers was based on previous work in this area[14] and published guidelines for reliability studies.[34-37] We selected testers with a range of experience and training in identifying study designs to reflect a cross-section of the resources typically available when producing an evidence report. Three testers had master's level training in epidemiology, one had a master's degree in a health science field, and two had undergraduate degrees in a health science or related field. The testers were given the 30 studies (Appendix D), the tool, and the accompanying glossary (Appendix E) with minimal additional instruction or direction. They were told that it would take approximately 5 to 10 hours to categorize the 30 studies and were asked to complete the assignment over a 2-week period. The study design names in the tool and glossary were masked and letter codes were used in their place in an attempt to have testers work through the flow diagram in a systematic fashion rather than relying on study design labels. This is supported by recommendations from the Cochrane Collaboration to focus on study design features rather than study design labels when assessing studies for inclusion in a systematic review.[3]

Concurrently, the sample of 30 studies, the tool, and the accompanying glossary were given to three members of the Steering Committee (DD, PS, VM) to develop the reference standard (i.e., the "true" classification for each study). After independently classifying the studies, the three reviewers met by teleconference to discuss disagreements and reach consensus.

We recorded the number of studies for each level of agreement (i.e., all testers agreed, 5 of 6 testers agreed, etc.). Inter-rater reliability was calculated using Fleiss' kappa (κ). Fleiss' κ was used as there were more than two raters; however, Fleiss' κ tends to be lower when there are a large number of possible categories. In this case, there were 13 possible classifications which is considered large. Inter-rater reliability was calculated separately for (1) the testers and (2) the reference standard raters before consensus. We also calculated inter-rater reliability based on the formal training of the testers (completed relevant graduate training vs. currently undertaking graduate training). We then calculated the accuracy of the testers against the reference standard classifications of study designs. The mean time taken to classify the sample of studies and the mean time per study for the testers were also calculated.

Due to the low level of agreement for both the testers and the reference standard raters, we believed that it was important to further modify the tool to enhance its usefulness and usability in the context of a systematic review. We conducted semi-structured interviews independently with each tester to gather information on their general impressions of the tool; how they applied the tool; what items they found confusing; and, what aspects of the process they found challenging (e.g., an unclear or confusing tool vs. complex studies). The feedback from these interviews and from the reference standard raters was used to modify the tool. The changes that were made to the algorithm and glossary between the two rounds of testing are detailed in Appendix F.

A second round of testing was conducted to evaluate the modified tool (Appendix G). We selected six testers from the UAEPC. Three of the testers were involved in the first round of testing and three had not been previously involved. The three testers that were involved in the first round received no feedback after the first round of testing and were not aware of the reference standard design classifications of the sample of studies used for testing. Of these three testers, two had graduate level training in epidemiology and one had undergraduate training in a health-related field. Our strategy for selecting testers was similar to Round One. Specifically, we selected testers with a range of training (completed relevant graduate training vs. currently undertaking graduate training). Four of the testers received the flow diagram with the study design labels. This was based on the feedback we received from the testers in Round 1, indicating a preference for unmasked labels, identifying the conventional names of the study types. In order to allow for a comparison between the changes in the algorithm between Round 1 and Round 2, exclusive of unmasking the study design labels, two testers received the flow diagram with letter codes masking the study design labels. A random sample was selected of 15 of the 30 studies used for the first round of testing. As per the first round of testing, the testers were given the 15 studies, the modified tool, and the accompanying glossary. They were told that it would take up to 5 hours to classify the studies and were asked to complete the assignment within a week.

The same analyses were conducted for the second round of testing. These included calculations of agreement (number of studies with different levels of agreement between reviewers), inter-rater agreement (Fleiss' κ), and accuracy against the reference standard. The agreement was calculated for the testers overall, by level of training, by participation in the first

round of testing, and whether the study designs were labeled or coded. We calculated the mean time taken to classify the sample of studies and the mean time per study.

Chapter 3. Results

Identification of Taxonomies (Objective 1)

We contacted 31 organizations or individuals to identify taxonomies/study design classification instruments for further evaluation. Figure 1 shows the number of contacts made and the responses received.

The Steering Committee reviewed the 23 tools that were received; 10 were considered relevant for the purposes of this project. Tables 1 and 2 describe the tools received.

Figure 1. Identification of study design classification tools

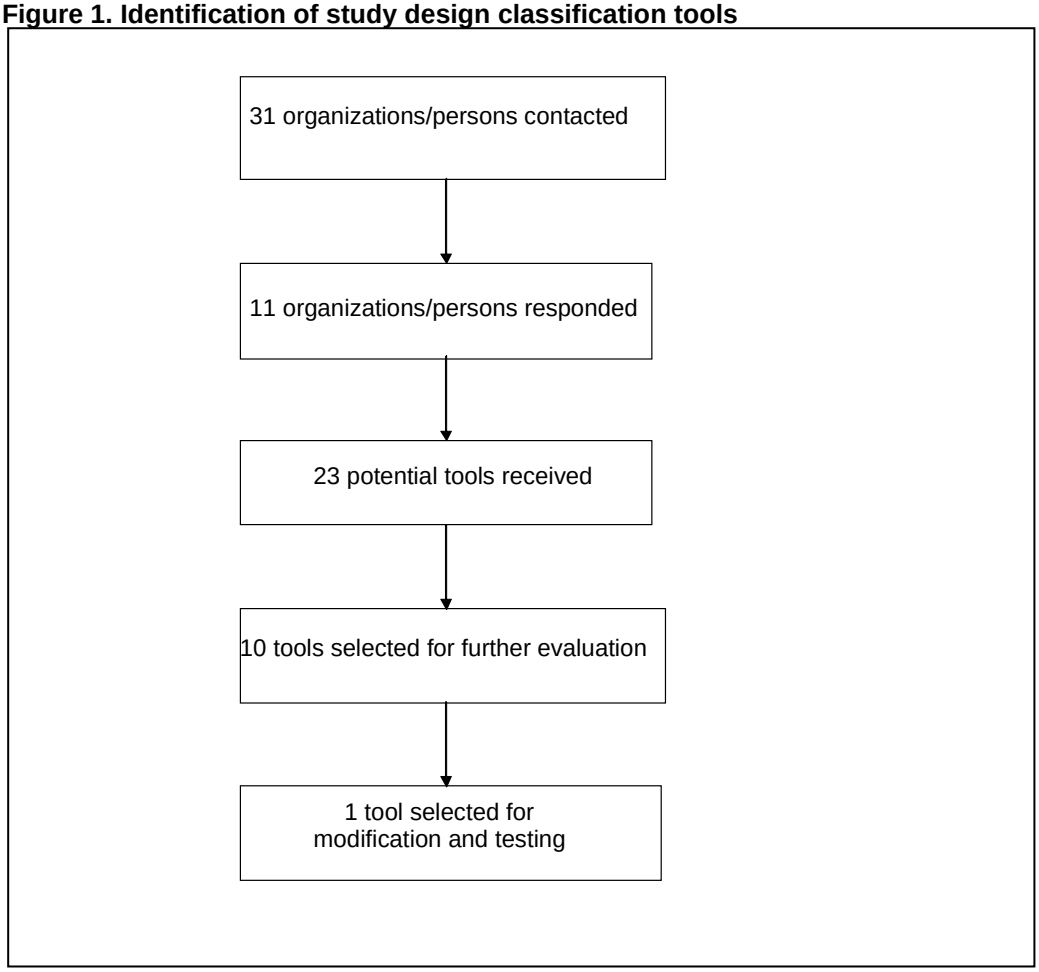


Table 1. Study design classification tools selected for further evaluation

Tool	Reference or source	Tool name used in this report
Research paper	Brown et al. 2008[43]	Brown
Design algorithm for studies of health care interventions	Cochrane Non-Randomised Studies Methods Group (NRSMG)*[44]	DASHCI
Definitions (based on Aschengrau et al. 2003; National Library of Medicine and the National Institute of Health); levels of evidence (based on Hamer and Collinson 1999)	Compiled by Minnesota EPC[45]	Minnesota
Systematic literature review specification manual: Study design algorithm (Appendix J)	World Cancer Research Fund[46]	SLR
Taxonomy of quasi-experimental studies	Campbell and Stanley 1966[47]	Campbell and Stanley
Traditional taxonomy of study design	Furlan 2006[14]; Furlan et al. 2008[48]	Traditional
Algorithm for classifying the research design of primary studies	American Dietetic Association (ADA) Evidence Analysis Manual[49]	ADA
List of study design features (Table 13.2.a) and some types of nonrandomized study (NRS) designs (Box 13.1.a)	Chapter 13, written by the Cochrane NRSMG, in Cochrane Handbook for Systematic Reviews of Interventions (Higgins & Green, 2008)*[5]	Cochrane Handbook
Algorithm of designs for treatment studies	RTI International-University of North Carolina at Chapel Hill (UNC) EPC[50]	RTI-UNC
Quality assessment tool for quantitative studies dictionary	McMaster University, School of Nursing, Effective Public Health Practice Project (EPHPP)[51]	EPHPP

*These documents were produced by the same group but at different times; the most recent approach to study design classification advocated by the Cochrane NRSMG is the second tool listed which appears in the current version of the Cochrane Handbook for Systematic Reviews of Interventions.

Table 2. Description of study design classification tools rejected for further evaluation

Tool	Reference or source	Comments
Algorithm of epidemiologic studies	Figueiras et al. 1997[52]	Algorithm for selecting designs for questions related to drug surveillance, not for study design classification
Research paper	Reeves 2004[53]	Framework presented was not comprehensive; provides background and design elements to consider when classifying studies
Systematic review	Estabrooks et al. 2001[54]	Details a systematic review and describes their approach to methodological evaluation of RCTs and observational studies in the context of a systematic review
Systematic review	Wong and Cummings, 2007[55]	Provides a quality assessment and validity tool for correlational studies
Systematic review	Lee and Cummings, 2008[56]	Provides a quality assessment and validity tool for correlational studies
Research paper	Zaza et al. 2008[57]	Same algorithm as ADA Evidence Analysis Manual (above)
Research paper	Heinsman and Shadish, 1996[58]	No algorithm or definitions
Research paper	Shadish et al. 2000[59]	No algorithm or definitions
Systematic review	Cummings and Estabrooks, 2003[60]	No algorithm or definitions
Systematic review	Cummings et al. 2008[61]	Provides a quality assessment and validity tool for correlational studies and pre/post intervention designs
Research paper	Brown et al. 2008a[62]	No algorithm or definitions
Research paper	Brown et al. 2008b[63]	No algorithm or definitions
Research paper	Cook et al. 2008[64]	No algorithm or definitions

Selection of Taxonomies (Objective 2)

Five members of the Steering Committee (DD, LS, MV, KB, LH) independently reviewed and ranked the 10 tools based on criteria presented in Chapter 2. Table 3 provides the results of the ranking and observations of the tools made during the process.

The three top-ranked tools were a "design algorithm for studies of health care interventions" (DASHCI) developed by the Cochrane Non-Randomised Studies Methods Group (NRSMG; note that this tool is no longer advocated by the NRSMG) and tools developed by the American Dietetic Association (ADA), and the RTI-UNC (Appendix H). All three were algorithms, i.e., they provided a logical sequence of "yes or no" decisions to make when classifying studies. The tools featured two starting points for design classification: (1) the assignment of the intervention/exposure and (2) the number of comparison groups. None of the taxonomies covered the range of study designs that systematic reviewers might encounter when conducting EPC evidence reports. Further, the nomenclature was inconsistent among the algorithms.

Due to the perceived failure of any single tool to meet the needs of systematic reviewers in terms of the comprehensiveness of study designs, the Steering Committee decided to select a single tool and to incorporate the desirable features from the other tools. The DASHCI tool emerged as the most preferred design algorithm and was used as the basis for further development (Table 1).

Table 3. Steering committee rankings of tools from most (1) to least (10) preferred

Tool	Median (modal) ranking	Comments
ADA	2 (2)	Not as easy to use as other flowcharts, not as comprehensive
Brown	10 (10)	Useful for controlled trials, does not seem as useful for cohort studies
Campbell and Stanley	7.5 (n/a)	Interesting additions to the design, but uses nomenclature that is unfamiliar which may reflect the age and/or context of the original document
Cochrane Handbook	5.5 (7)	Not as easy to read as a flowchart, but more comprehensive list of assignments for interventions; cannot be used to assign design, but could be used to check the response
DASHCI	1 (1)	Able to assign all studies from test sample into boxes
EPHPP	8 (9)	Not comprehensive enough, not able to deal with complex designs
Minnesota	7 (8)	Not clear that categories do not overlap, but some interesting additions in design (e.g., ambi-directional cohort study); no flow diagram
RTI-UNC	3 (2)	Not comprehensive enough, not able to deal with complex designs, but clean visual lines
SLR	4 (3)	Not able to deal with complex designs, but clear nomenclature and clean visual lines
Traditional	6 (4)	Not comprehensive enough, not able to deal with complex designs

n/a: not applicable because raters all gave different rankings

Taxonomy Testing and Development of Reference Standard (Objective 4)

Reference Standard

Three members of the Steering Committee (DD, LS, MV) developed the reference standard by independently applying the flow diagram to the 30 studies. Disagreements were resolved through discussion and consensus. All three reviewers had doctoral level training in epidemiology or research design and 4 to 8 years experience in systematic reviews and/or EPC work.

The three reviewers agreed on the classification of seven studies (23 percent), two of the three agreed on the classification of 14 (47 percent), and there was no agreement on the classification of nine (30 percent). The overall agreement was fair (κ=0.33).

Disagreements occurred at most decision points in the algorithm except for the following three: (1) "Were at least three measurements made before and after intervention/exposure?" (2) "Was intervention/exposure data registered prior to disease?" and (3) "Were both exposure/intervention and outcome assessed prospectively?" Each of these decision nodes was the last in their respective branches of the flow diagram.

The area that created the greatest confusion and disagreements for the reference standard raters was the decision node "Was there a single cohort?" Specifically, it was often difficult to determine whether the two groups under study were derived from the same cohort and the tool did not provide any criteria to make this decision. A second decision node where disagreements occurred was the first in the flow diagram: "Was there a comparison?" Specifically, it was unclear whether or not to classify the study as having a comparison when subgroup analyses were performed within a single group. A third point of disagreement was determining when a study was an interrupted time series (i.e., measurements taken at a minimum of three timepoints before and three timepoints after the intervention). While there is a precedent for this definition (http://www.epoc.cochrane.org/Files/Website/Reviewer%20Resources/inttime.pdf), the number of required timepoints may not be universally accepted.

Test 1

Tester characteristics. Six staff members at the UAEPC tested the modified classification tool. These individuals had varying levels of relevant training, experience with systematic reviews in general, and experience with EPC work specifically. The length of time they had worked with the UAEPC ranged from 9 months to 9 years. Three of the testers had obtained a master's degree in public health or epidemiology and three testers were undertaking graduate level training in epidemiology or library and information sciences.

The time taken to classify the 30 studies ranged from 7 to 9 hours with a mean of 8 hours overall and 16 minutes per study. Since the tool was new to the testers, this time reflects, in part, the process of familiarizing themselves with the flow diagram and the accompanying definitions.

Agreement. There were no studies for which all six testers agreed on the classification (Table 4). Five of six testers agreed on seven studies, four agreed on five studies, three agreed on nine studies, two agreed on eight studies. The overall level of agreement was considered fair (κ=0.26) (see Table 5 for interpretation of Fleiss' kappa statistic). The degree of agreement for testers who

had completed graduate level training was fair (κ=0.38), while for testers undertaking graduate training it was slight (κ=0.17).

We examined the disagreements in design classification and no clear patterns emerged (Table 6). Disagreements occurred at all decision points in the taxonomy. One decision point, "Was there a single cohort?", emerged as particularly problematic. The following terminology and contrasts used in the flow diagram were described as unclear or confusing: "group", "group" vs. "cohort", "concurrently", "comparison", and "exposure" vs. "intervention."

The testers were asked whether they thought the source of disagreement was due to the tool, the studies, or both. In one case the tester thought the tool was good and two felt it was easy to use; however, they generally remarked that the disagreement arose due to poor reporting at the study level. For example, it was often unclear whether a study was prospective or retrospective; in fact, one study was described as retrospective in the Abstract and prospective in the Methods section. One tester commented that the variety of topics covered by the 30 studies made classification challenging and the tool may be easier to apply in the context of a systematic review in which studies are more similar in terms of topic and design issues.

Four of the testers commented that they were often motivated by what design they thought the study to be. For instance, some testers said that they would read the study, determine their own sense of what the design was, and work backwards through the flow diagram to justify their design selection. Alternatively, they would work through the flow diagram to a design endpoint, check the definition to ensure that it was consistent with their own interpretation, and then work backwards to the decision node that would take them to the design they thought was more appropriate.

Finally, two testers indicated that they did not use the definitions that accompanied the flow diagram, while two testers said that the descriptions helped them make decisions and make sense of the letter answer. Several testers indicated that they preferred design labels on the flow diagram rather than the letter codes.

Accuracy of testers compared to reference standard. The accuracy of the testers was assessed against the reference standard (Table 7). There were no studies for which all six testers agreed with the reference standard and there was generally wide variation in level of accuracy across the studies.

Table 4. Results of testing

	Test 1 (30 studies)	Test 2 (15 studies)
Overall agreement	κ=0.26	κ=0.45
Item agreement (number of studies):		
6/6 testers agreed	0	3 (20%)
5/6 testers agreed	7 (23%)	2 (13%)
4/6 testers agreed	5 (17%)	6 (40%)
3/6 testers agreed	9 (30%)	2 (13%)
2/6 testers agreed	8 (27%)	2 (13%)
no agreement	1 (3%)	0
Time for assessment (mean)		
overall	8 hours	2 ¾ hours
by study	16 minutes	11 minutes

Table 5. Interpretation of Fleiss' kappa (κ)(from Landis and Koch 1977)[38]

κ	Interpretation
<0	Poor agreement
0.0-0.20	Slight agreement
0.21-0.40	Fair agreement
0.41-0.60	Moderate agreement
0.61-0.80	Substantial agreement
0.81-1.0	Almost perfect agreement

Test 2

Tester characteristics. Six staff members at the UAEPC were involved in the second round of testing. Three of the testers had been involved in the first round of testing, while three of the testers had no previous involvement with the project or the taxonomy being tested. One tester had a PhD in medicine, three testers had a master's degree in epidemiology, and two testers had undergraduate degrees in health sciences or related field and were undertaking graduate level training in epidemiology. The length of time the testers had worked with the UAEPC ranged from 2 months to 9 years. Four of the testers used a flow diagram that had the study design labels, while two of the testers used a flow diagram with letter codes.

Agreement. The time taken to classify the 15 studies ranged from 2.25 to 4 hours with means of 2.75 hours overall and 11 minutes per study. There were three studies for which all six testers agreed on the classification. Five of six testers agreed on two studies, four agreed on six studies, three agreed on two studies, and two agreed on two studies. The overall level of agreement was considered moderate (κ=0.45) (Table 4). The degree of agreement for testers who had completed graduate level training was moderate (κ=0.45), while for testers undertaking graduate training it was fair (κ=0.39). The level of agreement was moderate for both those who had the flow diagram with study design labels (κ=0.41) and for those with letter codes (κ=0.55).

Accuracy of testers compared to reference standard. The accuracy of the testers was assessed against the reference standard (Table 6). There were three studies for which all six testers agreed with the reference standard, but generally there was wide variation in the level of accuracy across the studies. Table 8 presents the accuracy of the testers against the reference standard by study design. There was improved accuracy for RCTs, nonrandomized trials, retrospective cohorts, interrupted time series (ITS) without comparison, and case-control studies. Accuracy decreased for controlled before-after studies, non-concurrent cohorts, and noncomparative studies. There was no difference for one before-after study. No comparisons could be made for ITS with comparison.

We examined the classification of studies to identify patterns of disagreements (Table 9). The most common disagreements occurred at four key decision nodes in the flow diagram: whether the study was "experimental" (5/15 studies), whether there was a comparison (4/15 studies), whether the assessment of exposure and outcome was prospective or retrospective (3/15 studies), and whether the intervention/exposure and outcome data were gathered concurrently (2/15 studies).

Table 6. Classification of studies: Round 1

Study (author, year)	Reference Standard	NonR trial	RCT	Prosp cohort	Retrosp cohort	ITS with comparison group	CBA	Non-concurrent cohort	Nested case-control	Case-control	ITS without comparison group	B-A	Cross-sectional	Non-comparative
Anderson 2006	B-A											√√√√		√
Bentas 2003	Non-comparative											√		√√√√
Blais 2003	ITS with comparison group					√√√	√√				√			
Boszotta 2004	B-A									√		√√√		√√
Cardo 1997	Case-control				√					√√√√				
Carey 1995	Cross-sectional												√√√	√√√
Chenot 2006	Non-comparative			√	√		√	√			√		√	
Cherkin 2002	Non-comparative							√					√√	√√√
Cranson 1991	CBA	√√					√√√√							
Darai 2002	Retrosp cohort	√		√	√		√√			√				
Davies 1996	B-A								√			√√√√	√	
DeVader 2007	Retrosp cohort			√	√			√√		√			√	
Happ 2008	NonR trial	√√	√√					√√						
Harris 1996	ITS with comparison group			√	√	√√√	√							
Herman 2001	B-A					√	√√	√√				√		
Hollabaugh 1998	CBA	√√			√√		√			√				
Kaplan 1996	Non-concurrent cohort				√√√√√			√						
Karlsson 2006	Non-concurrent cohort			√√	√			√	√	√				
Leclercq 2000	B-A	√					√√√√			√				
Lipscomb 2003	ITS without comparison group										√√√	√√		√
Minassian 2005	B-A			√√√						√		√		√

NonR trial=nonrandomized trial; RCT=randomized controlled trial; prosp=prospective; retrosp=retrospective; ITS=interrupted time series; CBA=controlled before-after; B-A=before-after

√ - indicates that one of the testers made the selection

Table 6. Classification of studies: Round 1 (continued)

Study (author, year)	Reference Standard	NonR trial	RCT	Prosp cohort	Retrosp cohort	ITS with comparison group	CBA	Non-concurrent cohort	Nested case-control	Case-control	ITS without comparison group	B-A	Cross-sectional	Non-comparative
Paulson 2004	Non-concurrent cohort				√		√	√√				√	√	
Qin 2002	CBA			√	√		√√√						√	
Scheurmier 1998	Non-concurrent cohort							√√√√		√		√		
Sit 2007	NonR trial	√√√√√					√							
Verrotti 1993	Non-concurrent cohort	√					√	√√	√			√		
Wells 2008	Cross-sectional			√			√						√√√√	
Wickizer 2004	Restrosp cohort			√		√√√	√				√			
Wilson 2008	RCT	√	√√√√√											
Zancanato 1990	RCT	√	√√√√√											

Table 7. Accuracy of testing compared to reference standard

Test 1 30 studies		Test 2 15 studies	
Number of testers with same design classification as reference standard	Occurrence	Number of testers with same design classification as reference standard	Occurrence
6	0	6	3 (20%)
5	6 (20%)	5	2 (13%)
4	4 (13%)	4	3 (20%)
3	7 (23%)	3	1 (4%)
2	3 (10%)	2	2 (13%)
1	7 (23%)	1	2 (13%)
0	3 (10%)	0	2 (13%)

Table 8. Accuracy of testing compared to reference standard by study design

	Number of testers with same design classification as reference standard	
	Test 1	Test 2
RCT (Wilson 2008)	5/6	6/6
RCT (Zancanato 1990)	5/6	6/6
Nonrandomized trial (Happ 2008)	2/6	4/6
Nonrandomized trial (Sit 2007)	5/6	5/6
Controlled before-after (Cranson 1991)	4/6	4/6
Controlled before-after (Hollabaugh 1998)	1/6	0/6
Controlled before-after (Qin 2002)	3/6	2/6
Before-after (Anderson 2006)	5/6	n/a
Before-after (Boszotta 2004)	3/6	n/a
Before-after (Davies 1996)	4/6	n/a
Before-after (Herman 2001)	1/6	n/a
Before-after (Leclerq 2000)	0/6	n/a
Before-after (Minassian 2005)	1/6	1/6
Nonconcurrent cohort (Kaplan 1996)	1/6	n/a
Nonconcurrent cohort (Karlsson 2006)	1/6	0/6
Nonconcurrent cohort (Paulson 2004)	2/6	n/a
Nonconcurrent cohort (Scheurmier 1998)	4/6	n/a
Nonconcurrent cohort (Verrotti 1993)	2/6	1/6
Retrospective cohort (Darai 2002)	1/6	n/a
Retrospective cohort (DeVader 2007)	1/6	3/6
Retrospective cohort (Wickizer 2004)	0/6	n/a
ITS with comparison (Blais 2003)	3/6	n/a
ITS with comparison (Harris 1996)	3/6	n/a
ITS without comparison (Lipscomb 2003)	3/6	4/6
Case control (Cardo 1997)	5/6	6/6
Cross-sectional (Carey 1995)	3/6	n/a
Cross-sectional (Wells 2008)	4/6	n/a
Noncomparative (Bentas 2003)	5/6	5/6
Noncomparative (Chenot 2006)	0/3	n/a
Noncomparative (Cherkin 2002)	3/3	2/3

n/a = study was not included in sample for second round of testing

Table 9. Classification of studies: Round 2

Study (author, year)	Reference Standard	NonR trial	RCT	Prosp cohort	Retrosp cohort	ITS with comparison group	CBA	Non-concurrent cohort	Nested case-control	Case-control	ITS without comparison group	B-A	Cross-sectional	Non-comparative
Bentas 2003	Non-comparative											√		√√√√
Cardo 1997	Case-control									√√√√√				
Cherkin 2002	Non-comparative									√			√√√	√√
Cranson 1991	CBA	√√					√√√√							
DeVader 2007	Retrosp cohort			√	√√√								√√	
Happ 2008	NonR trial	√√√√						√√						
Hollabaugh 1998	CBA	√√√√		√	√									
Karlsson 2006	Non-concurrent cohort	√√			√√√√									
Lipscomb 2003	ITS without comparison group										√√√√	√		√
Minassian 2005	B-A			√√√√								√		√
Qin 2002	CBA			√√	√		√√			√				
Sit 2007	NonR trial	√√√√√	√											
Verrotti 1993	Non-concurrent cohort	√			√		√√	√		√				
Wilson 2008	RCT		√√√√√√											
Zancanato 1990	RCT		√√√√√√											

NonR trial=nonrandomized trial; RCT=randomized controlled trial; prosp=prospective; retrosp=retrospective; ITS=interrupted time series; CBA=controlled before-after; B-A=before-after

√ - indicates that one of the testers made the selection

Chapter 4. Discussion

The goal of this project was to identify a tool that could be used within the context of systematic reviews to assist with the classification of study designs. This study builds on previous work to test a "traditional taxonomy" in the area of interventions for low back pain.[14] The previous study suggested a number of directions for further research including developing a more comprehensive taxonomy in terms of the scope of study designs, and testing the taxonomy in different fields of research. We identified over 20 tools and selected one for modification and testing. One of the critical criteria in the selection process was comprehensiveness of the tool in terms of study designs. The final testing of the modified tool showed moderate agreement among six testers and low accuracy against a predetermined reference standard. The moderate level of agreement is consistent with that observed in the previous study.[14] The level of agreement observed in these two studies raises questions and concerns around the reliability, validity, and ultimately the utility of available classification tools. There are numerous tools in existence and, to our knowledge, few (if any) have undergone testing either during or after development. However, our findings also demonstrate that it is possible to systematically test and modify a tool in order to yield more reliable results.

There are various reasons for the moderate and low level of agreement and accuracy observed in our study. In general, it is difficult to determine the extent to which the results reflect issues with the taxonomy itself versus attributes of the studies that were selected for testing. The studies used during testing were identified and selected because they had posed challenges with respect to design classification within various systematic reviews. It might be expected that the agreement would be better among a more representative sample of all studies that would be included in a systematic review. The complexity of studies is partially reflected in the time taken to classify each study (16 minutes in Round 1 and 11 minutes in Round 2 compared to 3 to 8 minutes in a similar study that was restricted to one topic area[14]). Further, in the context of a systematic review, the challenges of study design classification will vary by topic depending on the designs included and the general state of the literature in the area. The sample of studies that we tested covered a wide range of topics. There may have been greater reliability if the studies had been on the same topic.

One of the main reasons that the selected studies were difficult to classify was poor reporting within the studies. This resulted in the need for testers to make assumptions or judgments in many cases (e.g., whether the timing of a study was prospective or retrospective). There is an urgent need for clearer reporting and stricter adherence to reporting guidelines, and these should be enforced at the journal level through the editorial and peer review process. However, these changes will have limited impact on systematic reviewers who typically review studies done in the past. In certain cases, we found that the classification could vary depending on the intent of the authors and this was not regularly clear from the written report. In addition, there were cases of a discrepancy between how the study was designed and how the data were analyzed. Consequently, the classification could vary depending on the focus and interpretation of the individual assigning the design. Further, studies may use one particular design for some outcomes and a different design for other outcomes (e.g., RCT for primary or short-term outcomes, subgroup analysis using a prospective cohort approach for safety or longer-term outcomes).

Clearly the observed level of agreement and accuracy may be due in part to shortcomings of the taxonomy itself. Many of the decision points were challenging. This may in part reflect some lack of clarity within the field of epidemiology. For example, one study found substantial variation in the interpretation and understanding of blinding across 25 textbooks and a sample of physicians.[39] It may also partly be explained by the evolution of study designs over time; designs are becoming more complex and incorporating mixed methods. We aimed to address Furlan's observation that a useful taxonomy should be comprehensive in terms of study designs. One of the design categories that Furlan[14] found lacking in the "traditional taxonomy" was "quasi-experimental" studies; this led to one of the main problems she describes in terms of difficulty choosing between the only two options of "experimental" and "observational" designs. Furlan further described three categories of studies: experimental (RCTs, CCTs), observational (cohort, case-control, cross-sectional), and descriptive (case series, case reports). This classification completely overlooks the area of "quasi-experimental" designs where the investigator has some control over certain aspects of design and study execution but the study may not be considered either purely experimental (a "trial") or purely observational. While our intent was to capture these "quasi-experimental" designs using more precise methods (e.g., study design features) and terminology, this was not overtly apparent to the testers. This resulted in discrepancies in a third of the studies as to whether or not the study was truly "experimental." Of note is the fact that two of the testers labelled several studies as "quasi-experimental" even though this was not a design label in the flow diagram. These study designs need to be more clearly reflected in the taxonomy, and clear guidelines for interpreting the extent of control that an investigator has are needed. The practical repercussion is that some "quasi-experimental" studies (e.g., before-after or controlled before-after studies) may incorrectly be classified as trials; hence, their validity may be exaggerated and the results given too much weight in the context of a systematic review. One design that is particularly problematic has been variously referred to as an "uncontrolled trial" or "single-arm trial." It is our opinion that this design should not be considered a "trial" because of the high risk of bias associated with having no control or comparison group. These studies should be considered "before-after" studies, and our taxonomy was designed to channel them toward this classification.

Many of the questions or decision nodes that reviewers might consider relatively straightforward (e.g., was there a comparison, was the study experimental, was the data collection prospective or retrospective) did not yield consistent responses. There are several factors that may contribute to this inconsistency including a lack of clarity or definitions within the questions posed in the algorithm; variation in the level of background knowledge, experience or training of those classifying the studies; and, an inconsistent use of design terminology among studies. While we provided a glossary in an attempt to offer standard definitions and clarity in terminology, we acknowledge that there may have been shortcomings with the glossary that created confusion. In retrospect, we found there was some ambiguity and inconsistency in terms and definitions. For instance, the terms "group," "cluster," and "observation" required greater clarity and consistency across the definitions. Further, definitions for the different types of cohort studies could be revised for more consistency. These challenges with respect to terminology were discussed in the epidemiology literature as early as the 1950s and have yet to be resolved.[22,40]

Inconsistent, inaccurate, or imprecise use of terminology can be confusing. For example, a sample may be randomly selected but not necessarily randomly assigned to treatment groups;

therefore, the use of the word "random" does not necessarily mean that the study is an RCT. In addition, the term "cohort" was considered problematic, despite the fact that a definition had been provided in the accompanying glossary. This raises the issue that the same terms can refer to different things (e.g., "cohort" may describe a group of people or a study design) and clearly distinguishing between the common and technical meanings of terms may not always be sufficient to prevent potential ambiguity or confusion. The inappropriate use of terminology by the authors of the research studies also creates confusion for the reader. For example, one study was described by the authors as a "case-control"; however, the reference standard considered it a controlled before-after study. The terms "prospective" and "retrospective" are often used loosely or in an ill-defined manner in the literature[41] and may refer to different constructs, including "directionality" (i.e., "temporal relationship between the observation of study factor level and the observation of disease status"[42]) and "timing" (i.e., "chronological relationship between the onset of the study and the occurrence of the primary phenomena under study"[42]). Reporting guidelines for observational studies recognize these different dimensions and recommend that authors refrain from using "prospective" and "retrospective" in favor of an explicit description of these dimensions of the study.[41] In some cases, variation arises from terminology specific to different fields or similar terminology used variably across fields (e.g., social sciences, education, psychology, medicine). There is a clear need for consistent use of terminology and study design labels and/or an understanding of the terminology used in a particular field by those undertaking a systematic review in that field.

One practical reason for the low agreement is the relatively large number of testers we chose and the large number of potential response categories. Although the statistical tests accommodate for this to some extent, the more testers and response categories there are the greater the likelihood of disagreement. Restricting the testing to two or three individuals may have yielded better results and more closely replicated the systematic review process. Nevertheless, the fact that six testers with systematic review experience and relevant training showed moderate agreement is problematic. Moreover, the three investigators with doctoral training in epidemiology or research design who developed the reference standard showed fair agreement. These observations go beyond the studies and taxonomy used in this study and reflect the more general complexities of study designs. Perhaps it is unreasonable to distill the myriad design elements that populate textbooks into a single flow diagram. The fact that 15 people assigned eight different designs to the same study[14] further highlights the complexity and variability of study design classification.

Variability in design classification and moderate agreement may also reflect differences in how individuals applied or worked through the taxonomy, or as Furlan described "the creativity of humans."[14] Testers commented on the presentation of the tool in terms of using letter codes rather than study design labels. The use of letter codes was done intentionally in an effort to increase the probability that testers would work through the flow diagram and answer each question sequentially, rather than place the studies into the categories they deemed appropriate based on study design label. Based on the feedback we received from the testers, they often worked backwards or back-tracked in any event in order to classify the studies according to what they felt was the most appropriate description. The reliance on study design labels, and the inappropriate or inaccurate use of study design labels or terminology by authors, has implications for the application of these types of tool. The glossary that we developed to

accompany the tool should mitigate these inconsistencies to some extent; users of the algorithm need to use standardized definitions provided.

The difficulties in interpreting study design labels and the consequent difficulties in reaching agreement in assigning these labels to individual studies are consistent with those of other researchers. These issues have led some authors to direct systematic reviewers to focus on features of designs rather than on design labels when assessing studies for inclusion and evaluating potential risk of bias.[3] We endorse this approach and recommend that reviewers should be as explicit as possible about the design features that are being considered. However, we do not think that this obviates the need for or the usefulness of some design labels in describing studies being considered for inclusion in systematic reviews. AHRQ reports tend to have broad inclusion criteria and, as a pragmatic issue, often require the consideration of design features as part of the inclusion/exclusion process. The use of a taxonomy is intended to provide greater transparency and consistency to the process by closely examining the design features. In addition, many reviewers still find it useful to be able to describe and categorize studies according to broad design rather than referring only to specific features.[3] Moreover, groups that do not advocate the use of study design labels continue to recognize the importance of considering inherent weaknesses in design features. We believe that design labels allow reviewers to keep issues regarding inherent weaknesses in mind while retaining the ability to categorize studies both broadly and according to specific design features. Whether design labels and the differences between studies that they help to identify are also useful for assessing risk of bias has not been well investigated empirically. We consider the results of this study as a contribution to the developing evidence base that may help guide this larger discussion about the utility of design labels for assessing the quality of evidence.

Implications for Practice

The appropriate classification of studies by design is a critical step in a systematic review in order to guide inclusion, risk of bias assessments, pooling of studies for analysis, interpretation of results, and grading the body of evidence. We believe that a tool such as the one tested in this study would be useful to guide this process, although application of the tool requires several considerations in order to optimize agreement and reliability among reviewers. First, there was some indication that those with more training showed greater agreement. Therefore training in research methods, as well as use of the tool, is essential. As with all other aspects of the systematic review process, pilot testing the tool in the context of each review is highly recommended. Second, decision rules may need to be made in the context of different fields of study and/or review topics. Specifically, decisions around how to handle lack of clarity due to inadequate reporting need to be clear. We recommend that when the response to a question in the taxonomy is unclear, the reviewer assume that the condition was not met. Studies with mixed designs or that used different designs for different hypotheses within the same study report are difficult to classify. Decision rules need to be made in the context of the review. Further, in a systematic review, data from studies may be used outside of the context of the study's design (e.g., data extracted for a single group of interest where more than one group was actually studied). Questions then arise as to whether to classify the study according to the research question that the review is addressing (and for which the study data is being used) or the original intent of the researchers. Documentation of the decision rules will allow for consistency and transparency.

Finally, we identified the following user preferences for consideration in adapting or selecting a tool for this purpose. Users preferred the taxonomy to begin with a comparison (i.e., was there a comparison) rather than assignment (i.e., who assigned the intervention) as the latter does not easily allow for mixed methods. There was a preference for the assignment to be broadly inclusive of exposure and intervention. Users preferred a visual framework with an algorithm (step-by-step process) to arrive at a classification (i.e., a flow diagram that "makes the decision for you"). There was also a preference for the algorithm to be more-or-less self-contained (e.g., definitions or clear terminology used inasmuch as possible) so that users don't have to refer to other, often lengthy, documentation. Finally, the testers preferred to see study design labels on the algorithm rather than letter codes; however, when two versions of the taxonomy with letter codes vs. design labels were tested, those with the letter codes showed greater agreement (κ=0.55 vs. κ=41). This finding was based on four testers using the design labels and two testers using the letter codes. Further research is needed to confirm this finding before recommendations can be made for application of the tool in practice.

Future Directions

This study confirms the findings and validates many of the observations made in a recent, similar study.[14] Both studies underscore the complexity of study designs and the current inadequacies with reporting. We believe that the tool we have developed serves as a basis for use in systematic reviews and further research. We made minor revisions following the final round of testing that merit further testing. In addition, critical review of the glossary by researchers and methodologists would be beneficial. Future research is needed to evaluate the tool within the context of a real systematic review; this would offer more focus in terms of content area and design issues, while offering wider and more representative scope of studies in terms of ease and difficulty of classification. Further testing should be done targeting specific study designs, particularly those that are difficult to classify. We provide some preliminary data around factors that might lead to differences in reliability across individuals, including varied experience, training, and education. Future research is needed to provide more definitive results around these various factors and how they impact the performance of the tool. The methods we employed could be refined in future work. Specifically, we encountered unanticipated challenges around developing the reference standard classifications. We feel that it is important that this process be done by senior researchers with extensive and relevant experience and training. Researchers with similar backgrounds in terms of experience and training may result in greater agreement. Alternative methods to achieve this task would be to develop the reference standard as a group, rather than independently from each other. Further, where consensus is due to lack of clarity in reporting, information could be sought by contacting the authors of the studies. Finally, in order to inform this field more broadly, work is needed to quantify the bias associated with design labels and the differences between studies that they help to identify.

Strengths and Limitations

We built on previous work by using existing taxonomies and combining their different elements to yield a single taxonomy that we believed was the most comprehensive in terms of the number of study designs it could classify. We tested the taxonomy using studies covering a range of different topics and types of interventions (e.g., surgical, educational, legislative, etc.) which enhances the generalizability of our results. However, the studies included in a single

systematic review would likely be more homogeneous in terms of their designs and the corresponding design issues which may result in greater agreement between those classifying the studies. We had testers with a range of education, experience, and training which enhances the generalizability for individuals who may be involved in a systematic review. We made a number of comparisons regarding education, experience, and training. These results should be considered exploratory; however, they do provide a basis for hypotheses in future research. We did not train the testers in the use of the tool prior to it being tested. Training the testers in the use of the classification tool prior to its testing may more closely replicate the process used in a real systematic review and may have resulted in greater agreement. Finally, our reference standard was based on consensus among three individuals with substantial expertise in research methods and systematic reviews. However, there was a high level of disagreement among the reference standard raters and resolving some of these disagreements required lengthy and sustained discussion. Hence, our reference standard may not be the design that was actually implemented by the investigators, in which case the accuracy of the tool may be over or underestimated.

Summary

We developed and tested a tool for the classification of study designs. The level of agreement among six testers was moderate and the accuracy against a reference standard was low. There are a number of explanations for the observed reliability and accuracy including possible shortcomings of the taxonomy (e.g., lack of clarity and comprehensiveness); inadequate study reporting (e.g., poorly described methods, inaccurate and inconsistent use of terminology); and variation in user characteristics (e.g., training/education/experience, preferences, application of the tool). Application of such a tool in the context of a systematic review should be accompanied by adequate training, pilot testing, and documented decision rules. This study demonstrates that systematic testing and refinement enhances the reliability of the tool. At the study level, clear reporting, adherence to published reporting guidelines, and appropriate and consistent use of design terminology should be enforced.

Table 10 is a summary of study findings, implications for practice, and directions for future research.

Table 10. Summary of findings, implications for practice, and directions for future research

Challenges in classifying study designs

- Poor reporting resulting in lack of clarity
- Mixed methods utilized in the same study, or more than one question or hypothesis being investigated in the same study that employ different methods
- Discrepancies between intent of investigator and study conduct
- Discrepancies between design and how data were analyzed
- Discrepancies in investigators' original plan and study implementation
- Inconsistent, inaccurate, or imprecise use of design terminology

Implications for practice

- Clear directions, guidelines, and decision rules specific to topic area or systematic review questions
- Adequate training and relevant education for those applying the tool
- Pilot testing of tool and instructions in the context of specific topic area
- At the study level, clear reporting, adherence to published reporting guidelines, and appropriate use of design terminology are critical.
-

Future research

- Testing the final revised tool in different topic areas and in the context of a real systematic review, and for specific study designs that are difficult to classify
- Further investigation of the effects of training, education, and experience on reliability
- Refinement of the methods employed in this study, including development of the reference standard and contacting authors of primary studies for clarification

References

1. Agency for Healthcare Research and Quality: Evidence-based Practice Centers. Available at: http://www ahrq gov/clinic/epc/. Accessed: October 30, 2008.
2. Egger M, Davey Smith G. Principles of and procedures for systematic reviews. In: Egger M, Davey Smith G, Altman DG, eds. Systematic reviews in health care: meta-analysis in context. London: BMJ; 2001. p. 23-42.
3. Reeves BC, Deeks JJ, Higgins JPT, et al. Including non-randomized studies. In: Higgins JPT, Green S, eds. Cochrane handbook for systematic reviews of interventions 5.0.1 [September 2008]. Chichester, UK: John Wiley & Sons, Ltd; 2008. p. 391-432.
4. Norris SL, Atkins D. Challenges using nonrandomized studies in systematic reviews of treatment interventions. Ann Intern Med 2005;142:1112-1119.
5. Higgins JPT, Green S. Cochrane handbook for systematic reviews of interventions 5.0.1 [September 2008]. Chichester, UK: John Wiley & Sons, Ltd; 2008.
6. Elm von E, Altman DG, Egger M, et al. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. PLoS Med 2008;4(10):e296.
7. Deeks JJ, Dinnes J, D'Amico R, et al. Evaluating non-randomised intervention studies. Health Technol Assess 2003;7(27).
8. Black N. Why we need observational studies to evaluate the effectiveness of health care. BMJ 1996;312:1215-1218.
9. Egger M, Davey Smith G, Schneider M. Systematic reviews of observational studies. In: Egger M, Davey Smith G, Altman DG, eds. Systematic reviews in health care: meta-analysis in context. London: BMJ; 2001:211-227.
10. Chou R, Helfand M. Challenges in systematic reviews that assess treatment harms. Ann Intern Med 2005;142(12):1090-1099.
11. McCulloch P, Taylor I, Sasako M, et al. Randomised trials in surgery: problems and possible solutions. BMJ 2002;324:1448-1451.
12. Reed D, Price EG, Windish DM, et al. Challenges in systematic reviews of education intervention studies. Ann Intern Med 2005;142:1080-9.
13. Hartling L, McAlister FA, Rowe BH, et al. Challenges in systematic reviews of therapeutic devices and procedures. Ann Intern Med 2005;142:1100-1111.
14. Furlan AD. Non-randomized studies: an evaluation of search strategies, taxonomy and comparative effectiveness with randomized trials in the field of low-back pain University of Toronto; 2006.
15. MacLehose RR, Reeves BC, Harvey IM, et al. A systematic review of comparisons of effect sizes derived from randomised and non-randomised studies. Health Technol Assess 2000;4(34):1-153.
16. Atkins D, Fink K, Slutsky J. Better information for better health care: the evidence-based practice center program and the agency for healthcare research and quality. Ann Intern Med 2005;142(12):1035-1041.
17. Ray JG. Evidence in upheaval: incorporating observational data into clinical practice. Arch Int Med 2002;162:249-254.
18. IOM (Institute of Medicine). The criteria and process for setting priorities. In: Initial national priorities for comparative effectiveness research.Washington, DC: The National Academies Press; 2009. p. 77-96.
19. Gartlehner G, Hansen RA, Nissman D, et al. A simple and valid tool distinguishes efficacy from effectiveness studies. J Clin Epidemiol 2006;59:1040-1048.
20. Rothman KJ, Greenland S, Lash TL. Modern epidemiology. 3rd ed. Philadelphia: Lippincott Williams & Wilkins; 2008.
21. Shadish WR, Cook TD, Campbell DT. Experimental and quasi-experimental designs for generalized causal inference. New York, NY: Houghton Mifflin Company; 2002.
22. Vandenbroucke JP. Prospective or retrospective: what's in a name? BMJ 1991;302:249-250.

23. Owens DK, Lohr KN, Atkins D, et al. Grading the strength of a body of evidence when comparing medical interventions-agency for healthcare research and quality and the effective health care program. J Clin Epidemiol 2009;July 10 [ePub ahead of print].
24. Atkins D, Best D, Briss PA, et al. Grading quality of evidence and strength of recommendations. BMJ 2004;328:1490.
25. Anderson K, Boothby M, Aschenbrener D, et al. Outcome and structural integrity after arthroscopic rotator cuff repair using 2 rows of fixation. Am J Sports Med 2006;34(14):1899-1905.
26. Carey TS, Evans A, Hadler N, et al. Care-seeking among individuals with chronic low back pain. Spine 1995;20(3):312-7.
27. Creasey G, Grill J, Korsten M, et al. An implantable neuroprosthesis for restoring bladder and bowel control to patients with spinal cord injuries: A multicenter trial. Arch Phys Med Rehabil 2001;82:1512-9.
28. DeVader SR, Neeley HL, Myles TD, et al. Evaluation of gestational weight gain guidelines for women with normal prepregnancy body mass index. Obstet Gynecol 2007;110(4):745-51.
29. Davies HTO, Crombie IK, MacRae WA, et al. Audit in pain clinics, changing the management of low-back and nerve-damage pain. Anaesthesia 1996;51:641-646.
30. Golden MP, Russell BP, Ingersoll GM, et al. Management of diabetes mellitus in children younger than 5 years of age. Am J Dis Child 1985;139:448-452.
31. Leon A, Greenberg J, Kanuru N, et al. Cardiac resynchronization in patients with congestive heart failure and chronic atrial fibrillation. Effect of upgrading to biventricular pacing after chronic right ventricular pacing. J Am Coll Cardiol 2002;39(8):1258-1263.
32. Scheurmier N, Breen AC. A pilot study of the purchase of manipulation services for acute low back pain in the United Kingdom. J Manipulative Physiol Ther 2009;21(1):14-18.
33. Srinivasan S, Craig ME, Beeney L, et al. An ambulatory stabilisation program for children with newly diagnosed type 1 diabetes. Med J Aust 2004;180:277-280.
34. Wilson TE, Fraser-White M, Feldman J, et al. Hair salon stylists as breast cancer prevention lay health advisors for african american and afro-caribbean women. J Health Care Poor Underserved 2008;19:216-226.
35. Walter SD, Eliasziw M, Donner A. Sample size and optimal study desings for reliability studies. Stat Med 1998;17:101-110.
36. Donner A, Eliasziw M. Sample size requirements for reliability studies. Stat Med 1987;6:441-448.
37. Eliasziw M, Young SL Woodbury MG, et al. Statistical methodology for the concurrent assessment of interrater and intrarater reliability: using goniometric measurements as an example. Phys Ther 1994;74(777):788.
38. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-174.
39. Devereaux PJ, Manns BJ, Ghali WA, et al. Physician interpretations and textbook definitions of blinding terminology in randomized controlled trials. JAMA 2001;285:2000-2003.
40. White C, Bailar JI. Retrospective and prospective methods of studying association in medicine. Am J Pub Health 1956;46:35-44.
41. Vandenbroucke JP, von Elm P, Altman DG, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and Elaboration. PLoS Med 2007;4(10):e297.
42. Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic research: principles and quantitative methods. New York, NY: John Wiley & Sons, Inc.; 1982. p. 57-58.
43. Brown C, Hofer T, Johal A, et al. An epistemology of patient safety research: a framework for study design and interpretation. Part 2. Study design. Qual Saf Health Care 2008;17;163-169.
44. Design algorithm for studies of health care interventions. Available at: http://www.cochrane.dk/nrsmg/docs/chap2fig.gif Accessed January 15, 2009.
45. Minnesota EPC (personal communication).
46. World Cancer Research Fund. Systematic Literature Review Specification Manual—version 15. Available at; http://www.dietandcancerreport.org/?p=slr_specification_manual.

47. Campbell DT, Stanley JC. Experimental and Quasi-Experimental Designs for Research. Rand McNally, Chicago, Illinois; 1966.
48. Furlan AD, Tomlinson G, Jadad AR, et al. Methodological quality and homogeneity influenced agreement between randomized trials and nonrandomized studies of the same intervention for back pain. Journal of Clinical Epidemiology 2008 Mar;61(3):209-231.
49. American Dietetic Association. Evidence Analysis Manual: Steps in the ADA Evidence Analysis Process. 2009.
50. Gartlehner G, Lohr KN, Linda Lux, et al. "Study Taxonomy and Glossary," unpublished material from the RTI-University of North Carolina Evidence-based Practice Center, North Carolina, 2005.
51. Effective Public Health Practice Project (EPHPP) (2007) Quality assessment tool for quantitative studies dictionary. McMaster University, Faculty of Health Sciences, School of Nursing. Available at: http://www.myhamilton.ca/myhamilton/CityandGovernment/HealthandSocialServices/Research/EPHPP/. Accessed December 1, 2010.
52. Figueiras A, Tato F, Takkouche B, et al. An algorithm for the design of epidemiologic studies applied to drug surveillance. Eur J Clin Pharmacol 1997; 51(6): 445-448.
53. Reeves BC. A framework for classifying study designs to evaluate health care interventions. Forsch Komplementarmed Klass Naturheilkd 2004;11(Suppl 1):13-17.
54. Estabrooks C, Goel B, Thiel E, et al. Decision aids: are they worth it? A systematic review. Journal of Health Services Research and Policy 2001; 6(3): 170-182.
55. Wong CA, Cummings GG. The relationship between nursing leadership and patient outcomes: a systematic review. Journal of Nursing Management 2007;15:508-521.
56. Lee H, Cummings GG. Factors influencing job satisfaction of front line nurse managers: a systematic review. Journal of Nursing Management 2008;16:768-783.
57. Zaza S, Wright-De Aguero LK, Briss PA, Truman BI, Hopkins DP, Hennessy MH, Sosin DM, Anderson L, Carande-Kulis VG, Teutsch SM, Pappaioanou M. Data collection instrument and procedure for systematic reviews in the Guide to Community Preventive Services. Am J Prev Med 2000;18 (1S):44-74.
58. Heinsman DT, Shadish WR. Assignment methods in experimentation: when do nonrandomized experiments approximate answers from randomized experiments? Psychological Methods 1996; 1(2): 154-169.
59. Shadish WR, Matt GE, Navarro AM, Phillips G. The effects of psychological therapies under clinically representative conditions: a meta-analysis. Psychological Bulletin 2000; 126(4): 512-529.
60. Cummings G, Estabrooks CA. The effects of hospital restructuring that included layoffs on individual nurses who remained employed: A systematic review of impact. International Journal of Sociology and Social Policy 2003; 23(8/9): 8-53.
61. Cummings G, Lee H, MacGregor T, Davey M, Wong C, Paul L, Stafford E. Factors contributing to nursing leadership: a systematic review. Journal of Health Services Research and Policy 2008: 13(4): 240-248.
62. Brown C, Hofer T, Johal A, Thomson R, Nicholl J, Franklin BD, Lilford RJ. An epistemology of patient safety research: a framework for study design and interpretation. Part 3. End points and measurement. Qual Saf Health care 2008;17:170-177.
63. Brown C, Hofer T, Johal A, Thomson R, Nicholl J, Franklin BD, Lilford RJ. An epistemology of patient safety research: a framework for study design and interpretation. Part 4. One size does not fit all. Qual Saf Health Care 2008;17:178-181.
64. Cook TD, Shadish WR, Wong VC. Three conditions under which experiments and observational studies produce comparable causal estimates: new findings from within-study comparisons. Journal of Policy Analysis and Management 2008; 27(4): 724-750.

Developing and Testing a Tool for the Classification of Study Designs in Systematic Reviews of Interventions and Exposures

Appendixes

Appendix A. Steering Committee Members

Name	Affiliation	Training	Expertise	Role in project
Kenneth Bond	UAEPC	MA (Philosophy)	Project management, systematic review/health technology assessment methods	Technical expert
Donna Dryden	UAEPC	PhD (Epidemiology)	Epidemiology, systematic review methods	Co-task order leader
Lisa Hartling	UAEPC	MSc (Community Health and Epidemiology), PhD (Candidate, Pediatrics)	Epidemiology, systematic review methods	Co-task order leader
Krystal Harvey	UAEPC	BSc	Project coordination	Project coordinator
P. Lina Santaguida	McMaster EPC	PhD (Epidemiology)	Epidemiology, systematic review methods	Technical expert
Karen Siegel	AHRQ	PT, MA	Stakeholder perspective	AHRQ representative
Meera Viswanathan	RTI-UNC EPC	PhD	Research design, systematic review methods	Technical expert

UAEPC = University of Alberta Evidence-based Practice Center; AHRQ=Agency for Healthcare Research and Quality

Appendix B. Contacts for Identification of Study Design Classification Tools

Evidence-based Practice Centers:

- Blue Cross and Blue Shield Association, Technology Evaluation Center
- Duke University
- ECRI Institute
- Johns Hopkins University
- McMaster University
- University of Minnesota
- University of Oregon
- RTI International—University of North Carolina
- Southern California
- Tufts University—New England Medical Center
- University of Alberta
- University of Connecticut
- Minnesota EPC
- University of Ottawa
- Vanderbilt University

Additional organizations (and representatives where applicable):

- The Cochrane Collaboration Non-Randomized Studies Methods Group (Barney Reeves)
- The Campbell Collaboration (William Shadish)
- National Health Services (Richard Lilford)
- Faculty of Nursing, University of Alberta (Greta Cummings)
- Department of Educational Psychology, Faculty of Education, University of Alberta (Veronica Smith)
- Canadian Agency for Drugs and Technologies in Health
- American College of Physicians

Nine additional individual experts were identified and contacted. Names have been withheld.

Appendix C. Letter of Request To Identify Study Design Classification Tools

Hello [name of organization or person]:

The University of Alberta Evidence-based Practice Center (UAEPC) is undertaking a methods research project concerning classification of study designs for non-randomized studies. The UAEPC is one of 15 EPCs funded by the Agency for Healthcare Research and Quality (AHRQ) to review the available evidence and produce evidence reports on health care topics, and to conduct research on systematic review methodology. AHRQ is funding this methods project.

As part of this research project, the UAEPC in conjunction with two other EPCs—McMaster University and RTI-UNC—will 1) identify and test tools that have been developed to classify study designs in systematic reviews, and 2) recommend a classification tool that can be used by the EPCs and other systematic reviewers to bring consistency to the classification of non-randomized study designs.

In the first stage of this project, we are collecting different taxonomies, classification tools, guidelines or other systems that have been used when classifying study designs. To ensure that we have a broad spectrum of classification tools, we are wondering if you would send us any classification systems or tools that you have used or are aware of. If you are willing to send us these systems, please just reply to this email with the attachment or citation. If there are other individuals or organizations that you think may have useful information, please forward this e-mail, or send us their contact information and we would be happy to follow-up.

Thank you for any help you can provide.

Donna M. Dryden, Ph.D.
Associate Director, U of A/Capital Health Evidence-based Practice Center
University of Alberta
Aberhart Centre, Room 9417
11402 University Avenue
Edmonton, AB, Canada

Phone: (780) 492-1273
Fax: (780) 407-6435
Email: ddryden@ualberta.ca
Web site: http:\\www.epc.ualberta.ca

Appendix D. Bibliography and Summary of Studies for Classification (Objective 4)

Anderson K, Boothby M, Aschenbrener D, et al. Outcome and Structural Integrity After Arthroscopic Rotator Cuff Repair Using 2 Rows of Fixation. Am J Sport Med 2006;34(14):1899-1905.

Bentas W, Wolfram M, Jones J, et al. Robotic Technology and the Translation of Open Radical Prostatectomy to Laparoscopy: The early frankfurt experience with robotic radical prostatectomy and one year follow-up. Eur Urol 2003;44:175-181.

Blais L, Couture J, Rahme E, et al. Impact of a cost sharing drug insurance plan on drug utilization among individuals receiving social assistance. Health Policy 2003;64:163-172.

Boszotta H, Prünner K. Arthroscopically assisted rotator cuff repair. Arthroscopy 2004;20(6):620-626.

Cardo DM, Culver DH, Ciesielski CA, et al. A case-control study of HIV seroconversion in health care workers after percutaneous exposure. The N Engl J Med 1997;337(21):1485-1490.

Carey TS, Evans A, Hadler N, et al. Care-seeking among individuals with chronic low back pain. Spine 1995;20(3):312-317.

Chenot J-F, Becker A, Leonhardt C, et al. Determinants for receiving acupuncture for LBP and associated treatments: a prospective cohort study. BMC Health Serv Res 2006;6(149).

Cherkin DC, Deyo RA, Sherman KJ, et al. Characteristics of visits to licensed acupuncturists, chiropractors, massage therapists, and naturopathic physicians. J Am Board Fam Pract 2002;15(6):463-472.

Cranson RW, Orme-Johnson DW, Gackenbach J, et al. Transcendental meditation and improved performance on intelligence-related measures: Longitudinal study. Pers Individ Dif 1991;12(10):1105-1116.

Darai E, Jeffry L, Deval B, et al. Results of tension-free vaginal tape in patients with or without vaginal hysterectomy. Eur J Obstet Gynecol Reprod Biol 2002;103:163-167.

Davies HTO, Crombie IK, MacRae WA, et al. Audit in pain clinics, Changing the management of low-back and nerve-damage pain. Anaesthesia 1996;51:641-646.

DeVader SR, Neeley HL, Myles TD, et al. Evaluation of Gestational Weight Gain Guidelines for Women With Normal Prepregnancy Body Mass Index. Obstet Gynecol 2007;110(4):745-751.

Happ MB, Sereika S, Garrett K, et al. Use of the quasi-experimental sequential cohort design in the Study of Patient-Nurse Effectiveness with Assisted Communication Strategies (SPEACS). Contemp Clin Trials 2008;29:801-808.

Harris CM, Scrivener G. Fundholders' prescribing costs: the first five years. BMJ 1996;313:1531-1534.

Herman RM, Richter P, Walega P, et al. Anorectal sphincter function and rectal barostat study in patients following transanal endoscopic microsurgery. Int J Colorectal Dis 2001;16:370-376.

Hollabaugh RS, Dmochowski RR, Kneib TG, et al. Preservation of putative continence nerves during radical retropubic prostatectomy leads to more rapid return of urinary continence. Urology 1998;51:960-967.

Kaplan SA, Santarosa RP, Te AE. Comparison of fascial and vaginal wall slings in the management of intrinsic sphincter deficiency. Urology 1996;47:885-889.

Karlsson I, Bondemark L. Intraoral Maxillary Molar Distalization, Movement before and after Eruption of Second Molars. Angle Orthod 2006;76(6):923-929.

Leclercq C, Victor F, Alonso C, et al. Comparative Effects of Permanent Biventricular Pacing for Refractory Heart Failure in Patients with Stable Sinus Rhythm or Chronic Atrial Fibrillation. Am J Cardiol 2000;85:1154-1156.

Lipscomb HJ, Li L, Dement J. Work-Related Falls Among Union Carpenters in Washington State Before and After the Vertical Fall Arrest Standard. Am J Ind Med 2003;44:157-165.

Minassian VA, Al-Badr A, Pascali DU, et al. Tension-Free Vaginal Tape: Do Patients Who Fail to Follow-up Have the Same Results as those Who Do? Neurourol Urodyn 2005;24:35-38.

Paulson DL, Wash L. A Comparison of Wait Times and Patients Leaving Without Being Seen When Licensed Nurses Versus Unlicensed Assistive Personnel Perform Triage. J Emerg Nurs 2004;30(4):307-311.

Qin L, Au S, Choy W, et al. Regular Tai Chi Chuan Exercise May Retard Bone Loss in Postmenopausal Women: A Case-Control Study. Arch Phys Med Rehabil 2002;83:1355-1359.

Scheurmier N, Breen AC. A Pilot Study of the Purchase of Manipulation Services for Acute Low Back Pain in the United Kingdom. J Manipulative Physiol Ther 2009;21(1):14-18.

Sit JWH, Yip VYB, Ko SKK, et al. A quasi-experimental study on a community-based stroke prevention programme for clients with minor stroke. J Clin Nurs 2007;16:272-281.

Verrotti A, Chiarelli F, Sabatino G, et al. Education, knowledge and metabolic control in children with type 1 diabetes. Eur Rev Med Pharmacol Sci 1993;15:5-10.

Wells NM, Yang Y. Neighborhood Design and Walking. Am J Prev Med 2008;34(4):313-319.

Wickizer TM, Kopjar B, Franklin G, et al. Do Drug-Free Workplace Programs Prevent Occupational Injuries? Evidence from Washington State. Health Serv Res 2004;39(1):91-110.

Wilson TE, Fraser-White M, Feldman J, et al. Hair Salon Stylists as Breast Cancer Prevention Lay Health Advisors for African American and Afro-Caribbean Women. J Health Care Poor Underserved 2008;19:216-226.

Zanconato S, Baraldi E, Santuz P, et al. Effect of inhaled disodium cromoglycate and albuterol on energy cost of running in asthmatic children. Pediatr Pulmonol 1990;8:240-244.

Table D1. Characteristics of studies selected for classification

Study	Study design label*	Study objective	Population	Intervention and comparator	Outcome(s)
Anderson 2006	Before-after	To investigate the outcomes and structural integrity of the arthroscopic repair of rotator cuff tears using 2 rows of fixation	Patients with full-thickness rotator cuff tears that could be restore to anatomical position on the greater tuberosity	Arthroscopic rotator cuff repair using 2 rows of suture anchors	Functional scores and evidence of retear or defect
Bentas 2003	Non-comparative	To evaluate the da Vinci surgical system for laparoscopic radical prostatectomy for prostate cancer	Adult patients with prostate cancer and eligible for radical prostatectomy	da Vinci surgical system	Operative morbidity (e.g., surgical time, blood loss, hospitalization), post-operative complications
Blais 2003	ITS with comparison group	To evaluate impact of a cost-sharing drug insurance plan for people receiving social assistance	Quebec residents receiving social assistance	Residents receiving social assistance vs. privately insured residents	Monthly consumption of medications
Boszotta 2004	Before-after	To evaluate arthroscopically assisted rotator cuff repair	Patients who received arthroscopically assisted transosseous rotator cuff repair	Arthroscopically assisted transosseous rotator cuff repair	Function scores
Cardo 1997	Case-control	To identify risk factors for the transmission of HIV to health care workers after percutaneous exposure to HIV-infected blood	Healthcare workers with documented exposure to HIV-infected blood	Workers who subsequently became seropositive vs. those who did not	Personal information, information on source patient and injury
Carey 1995	Cross-sectional	To determine the prevalence of chronic low back pain and the extent to which treatment is sought	Adults who had experienced back pain within the last 2 years	NA	Demographics and clinical characteristics, types of care sought, types of treatment received
Chenot 2006	Non-comparative	To explore factors associated with acupuncture treatment for low back pain and the association of acupuncture with use of other health care services	Adult patients with low back pain being seen by general practitioners	NA	Use of pharmacologic and nonpharmacologic treatments for low back pain

CAM=complementary and alternative medicine; CBA=controlled before-after; ITS=interrupted time series; NonR trial=nonrandomized trial; RCT=randomized controlled trial; prosp=prospective; retrosp=retrospective; ITS=interrupted time series; RCT=randomized controlled trial
*Study design label is that assigned by the reference standard.

Table D1. Characteristics of studies selected for classification (continued)

Study	Study design label*	Study objective	Population	Intervention and comparator	Outcome(s)
Cherkin 2002	Non-comparative	To describe patients and problems seen by CAM practitioners	Licensed acupuncturists, chiropractors, massage therapists, and naturopathic physicians	NA	Patient and practice characteristics (e.g., role of practitioner in path of care, reason for and duration of visit)
Cranson 1991	CBA	To investigate the effect of Transcendental Meditation™ and TM-Sidhi™ programs on IQ scores and reaction time	First-year university students enrolled in a psychology course	Participation in daily TM and TM-Sidhi programs vs. nonparticipation in TM and TM-Sidhi programs	Cattel's Culture Fair Intelligence Test and Hick's reaction time
Darai 2002	Retrosp cohort	To assess complications and cure rates of tension-free vaginal tape (TVT) procedure with or without hysterectomy	Women treated with stress urinary incontinence who had undergone TVT procedure alone or TVT with vaginal hysterectomy	TVT procedure alone vs. TVT procedure combined with vaginal hysterectomy	Post-operative urinary flow and subjective cure rate
Davies 1996	Before-after	To identify and promote appropriate changes in management of low back pain and nerve-damage pain	Patients suffering from low back or nerve-damage pain attending outpatient pain clinics	Active feedback (mailed report and personal visit) of variations in practice within and between clinics	Changes in practice
DeVader 2007	Retrosp cohort	To investigate the relationship between gestational weight (GW) gain and adverse pregnancy outcomes among women with normal prepregnancy body mass index (BMI)	Women with normal prepregnancy BMI who delivered full-term singletons	Lower than recommended GW vs. recommended GW vs. more than recommended GW	Adverse pregnancy outcomes
Happ 2008	NonR trial	To investigate the impact of a systematically implemented assistive communication intervention with nonspeaking ICU patients and their nurse caregivers using a control group comparison	ICU nurses and nonspeaking ICU patients	Basic communication skills training (BCST) vs. BCST and assistive communication strategies vs. usual care	Observational measures of nurse-patient communication performance

Table D1. Characteristics of studies selected for classification (continued)

Study	Study design label*	Study objective	Population	Intervention and comparator	Outcome(s)
Harris 1996	ITS with comparison group	To determine whether the first 5 waves of English fundholding practices have reduced their prescribing costs relative to non-fundholding practices	All general practices in England	Fundholding practices vs. nonfundholding practices	Changes and rates of change in cost per prescribing unit and changes in number of items per prescribing unit
Herman 2001	Before-after study	To evaluate the effect of transanal endoscopic microsurgery on anorectal motility and continence control and to define potential risk factors of post-surgical anorectal dysfunction	Adults with mobile rectal tumors qualifying for surgery	NA	Anorectal vector volume manometry, sphincter reflex evaluation, and barostat study
Hollabaugh 1998	CBA	To evaluate incontinence rates after radical retropubic prostatectomy	Men with clinically localized prostate cancer	Modified radical retropubic prostatectomy vs. standard anatomic retropubic prostatectomy	Assessment of continence (dryness)
Kaplan 1996	Non-concurrent cohort	To compare safety and efficacy of fascial vs. vaginal wall slings in the management of women with intrinsic sphincter deficiency (ISD)	Women who underwent surgical repair of ISD	Fascial sling vs. vaginal wall sling	Incontinence and patient satisfaction
Karlsson 2006	Non-concurrent cohort	To evaluate maxillary molar distalization and anchorage loss in 2 groups before and after eruption of second maxillary molars	Children receiving orthodontic treatment	Distal movement performed before eruption of second molar vs. simultaneous movement of first and second molars	Cephalometric measures (e.g., maxillary base, molar position and inclination)
Leclercq 2000	Before-after	To compare long-term clinical effects of permanent biventricular pacing in patients with stable sinus rhythm or chronic atrial fibrillation	Adults with with dilated cardiomyopathy and intraventricular conduction delay	Implanted permanent cardiac biventricular pacemaker	QRS duration and axis, NYHA class, LV ejection fraction, and peak VO_2

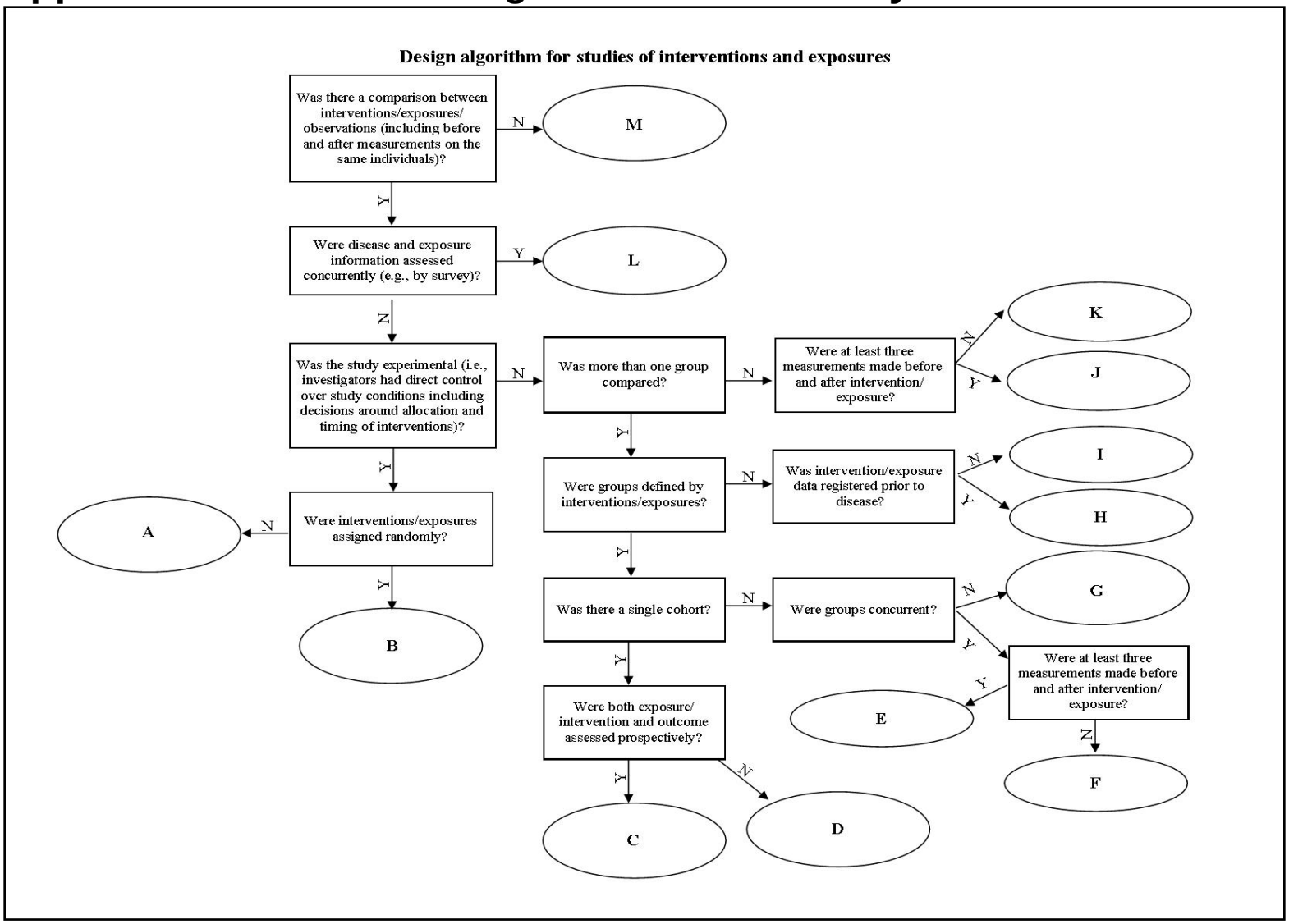
Table D1. Characteristics of studies selected for classification (continued)

Study	Study design label*	Study objective	Population	Intervention and comparator	Outcome(s)
Lipscomb 2003	ITS without comparison group	To evaluate changes in rates and severity of falls from elevations among union carpenters after fall prevention standard change	Union carpenters working in Washington state between 1989-1998 who worked at least 3 mo.	Rates before vs. rates after enactment of standard change	Injuries, paid lost time due to falls from elevation, mean direct payments for falls from heights
Minassian 2005	Before-after	To compare success and complication rates of the tension-free vaginal tape between patients with good vs. poor followup	Women with stress urinary incontinence	Tension-free vaginal tape procedure	Absence of stress urinary incontinence
Paulson 2004	Non-concurrent cohort	To compare wait time and number of patients who leave without being seen between triage systems that use nurses vs. unlicensed assistive personnel	Patients presenting to the emergency department	Patient triage by nurses vs. by unlicensed assistive personnel	Wait time and number patients leaving without being seen
Qin 2002	CBA	To evaluate potential benefits of regular Tai Chi Chuan exercise on the weight-bearing bones of postmenopausal women	Postmenopausal women who practice Tai Chi Chuan regularly and nonexercising controls	Tai Chi Chuan vs. no exercise	Bone mineral density
Scheurmier 1998	Non-concurrent cohort	To test the cost implications of recommendations for purchasing arrangements for low back pain in primary care settings	General practices purchasing manipulation services under National Health Service arrangements	Recommended purchasing arrangements vs. previous purchasing arrangement	Waiting time for first attendance, number of consultations, drug use and cost, recovery time, x-ray utilization and cost of care
Sit 2007	NonR trial	To determine the effectiveness of a community based stroke prevention program in improving knowledge about stroke, improving self-monitoring, and maintaining behavioral changes for stroke prevention	Adults with diagnosed minor stroke who live independently and are cognitively intact	Conventional medical treatment and community-based stroke prevention program vs. conventional medical treatment only	Lifestyle habits, medication compliance, and stroke knowledge

Table D1. Characteristics of studies selected for classification (continued)

Study	Study design label*	Study objective	Population	Intervention and comparator	Outcome(s)
Verrotti 1993	Non-concurrent cohort	To investigate whether quality of metabolic control is related to knowledge of disease	Children with type 1 diabetes mellitus	Diabetes education started at diagnosis vs. education started after diagnosis	Blood glucose control
Wells 2008	Cross-sectional	To examine associations between neighborhood design and walking	Women partnered with Habitat for Humanity neighborhood developments	Neo-traditional neighborhoods vs. conventional suburban neighborhoods	Environmental characteristics (e.g., street networks, land use), walking
Wickizer 2004	Retrosp cohort	To evaluate the effect of a publicly sponsored drug-free workplace program on reducing the risk of occupational injuries	Private companies insured through the Department of Labor and Industries	Companies participating and not participating in the drug-free workplace program	Injury rates, work time lost due to serious injury
Wilson 2008	RCT	To assess the effectiveness of breast cancer health promoting messages administered by salon stylists to clients in the salon setting	Hair salons in urban minority area	Salons promoting health messages vs. salons not promoting health messages	Conduct of monthly breast self-exam, mammogram, receipt of information on breast health
Zancanato 1990	RCT	To evaluate the effect of disodium cromoglycate and albuterol on energy cost of running in children with exercise-induced asthma	Children with mild to moderate exercise-induced asthma	Exercise test with premedication vs. exercise test without premedication	$FEV_{1,}$ peak VO_2, energy cost of running, ventilation

Appendix E. Round One Algorithm and Glossary



DESIGN ALGORITHM FOR STUDIES OF INTERVENTIONS AND EXPOSURES

When using the algorithm, it is recommended that you do not rely on the design labels assigned by the authors of the report, but rather work through the questions in the algorithm based on the methods presented in the report and the definitions provided below.

Study Design Key

Below is a list of definitions that correlate with study designs assigned by the accompanying taxonomy. At the end of this list are some additional concepts that may be useful during study design classification.

A
A study in which individuals or groups of individuals (e.g., community, classroom) are assigned to the intervention or control by a method that is not random (e.g., date of birth, date of admission, judgement of the investigator). Individuals or groups are followed prospectively to assess differences in the outcome(s) of interest. The unit of analysis is the individual or the group, as appropriate.

B
A study designed to test the efficacy of an intervention on an individual, a group of individuals, or clusters (e.g., classrooms, communities). Individuals or clusters are randomly allocated to receive an intervention or control/comparison (e.g., placebo or another intervention) and are followed prospectively to assess differences in outcomes. The unit of analysis is the individual, group of individuals, or the cluster, as appropriate. Variations in treatment assignment and measurement produce different types of studies including factorial, cross-over, parallel, stepped wedge and Solomon four-group.

C
A study in which individuals in the group without the outcome(s) of interest (e.g., disease) are classified according to exposure status at baseline (exposed or unexposed) and then are followed over time to determine if the development of the outcome of interest is different in the exposed and unexposed groups.

D
A study in which a group of individuals is identified on the basis of common features that were determined in the past. The group is usually assembled using available data sources (e.g., administrative data). Individuals are classified according to exposure status (exposed or unexposed) at the time the group existed and are followed up to a prespecified endpoint to determine if the development of the outcome of interest is different in the exposed or unexposed groups.

E
A study in which multiple observations over time are “interrupted” by an intervention or exposure and in which two series are examined (one is a comparison group). There must be at

least 3 observations before and at least 3 observations after the intervention or exposure for each group. The investigator(s) does not assign or have control over the intervention/exposure, which may be an environmental variable (e.g., airborne toxin) or administrative assignment (e.g., seatbelt legislation, educational program, service delivery model) but does have control over the timing of the measurement and the variables being measured.

F
A study in which the outcome(s) of interest is measured both before and after the intervention or exposure in two or more groups of individuals. In this study design the study group receives the intervention or exposure and the comparison group(s) does not. This type of study includes interventions that may be in the control of the investigator (e.g., a surgical procedure) as well as interventions that may be an environmental variable (e.g., airborne toxin) or administrative assignment (e.g., seatbelt legislation). In all cases the investigator(s) has control over the timing of the measurement and the variables being measured.

G
A study in which 2 or more groups of individuals are identified on the basis of common features at different time points. Individuals in each group are classified according to exposure status (exposed or unexposed) at the time the groups existed or were created. They are followed to determine if the development of the outcome of interest is different in the exposed or unexposed groups.

H
A study where exposed and control subjects are drawn from the population of a prospective cohort study. Baseline data are obtained at the time the population is identified; the population is then followed over a period of time. The study is then carried out using persons in whom the disease or outcome has developed and a sample of those who have not developed the outcome of interest (controls).

I
A study in which participants are selected based on the known outcome(s) of interest (e.g., disease, injury). Exposure status is then collected based on the participants' past experiences. Exposure status is compared between the two (or more) groups: those who have the outcome of interest and those who do not have the outcome of interest (controls). This is a retrospective study that collects data on events that have already occurred.

J
A study in which multiple observations over time are "interrupted" by an intervention or exposure. There must be at least 3 observations before the intervention and at least 3 observations after the intervention; otherwise, the study is considered a before-after study. The investigator(s) does not assign or have control over the intervention/exposure, which may be an environmental variable (e.g., airborne toxin) or administrative assignment (e.g., seatbelt legislation, educational program, service delivery model) but does have control over the timing of the measurement and the variables being measured.

K
A study of an intervention or exposure in which the investigator(s) compares the outcome(s) of interest both before and after the intervention in the same group of individuals. This includes interventions that may be in the control of the investigator (e.g., a surgical procedure) as well as interventions that may be an environmental variable (e.g., airborne toxin) or administrative assignment (e.g., seatbelt legislation). In all cases the investigator(s) has control over the timing of the measurement and the variables being measured.

L
A study in which both the exposure and the outcome status in a target population are assessed concurrently, that is, at the same point in time or during a brief period of time. The temporal sequence of cause and effect cannot necessarily be determined. They are most commonly used to assess prevalence. A common method for data collection is a survey.

M
Examples of this design include:

- A study that presents a description of a single patient or participant. Studies are usually retrospective and typically describe the manifestations, clinical course, and prognosis of the individual.
- A study that describes the experience of a group of patients with a similar diagnosis and/or treatment. Studies are usually retrospective and typically describe the manifestations, clinical course, and prognosis of a condition.
- A study in which data are collected at a series of points in time on the same population to observe trends in the outcome(s) of interest.

Additional Concepts

Cluster
The term ‘cluster’ refers to a unit of allocation or analysis in a clinical trial. Examples of clusters include hospitals, schools, neighborhoods, or entire communities.

Cluster randomized controlled trial
Synonym: *community trial; group randomized trial*
A randomized controlled trial in which the units of randomization and analysis are groups of people or communities (e.g., classroom, hospital, town). Typically, several communities receive the intervention and several different communities serve as controls.

Cohort
The term ‘cohort’ refers to a group of individuals (or other organizational units) who have a common feature when they are assembled (e.g., birth year, place of employment, medical condition, place or time period of medical treatment) and are followed over time. They can be followed prospectively or examined retrospectively.

Experimental study
A type of study in which investigators have direct control over the timing, course, and assignment of the intervention. Experimental studies investigate an intervention to determine its effect on the outcome(s) of interest. In an experimental study a population is selected to receive a specific intervention the effects of which are measured by comparing the outcomes in the experimental group with the outcomes of a control group that has received another intervention or placebo. Examples include randomized controlled trial, cluster randomized controlled trial, nonrandomized trial, n-of-one trial. See also **observational study**.

Observational study
A study in which the investigator(s) does not control the exposure/ intervention status of study participants (i.e., the assignment of the intervention or exposure of interest is not under the control of the investigator(s)). The simplest form of observational study is the case report or case series, which describes the clinical course of individuals with a particular condition or diagnosis. Observational studies include descriptive and analytic studies. See also **experimental study**.

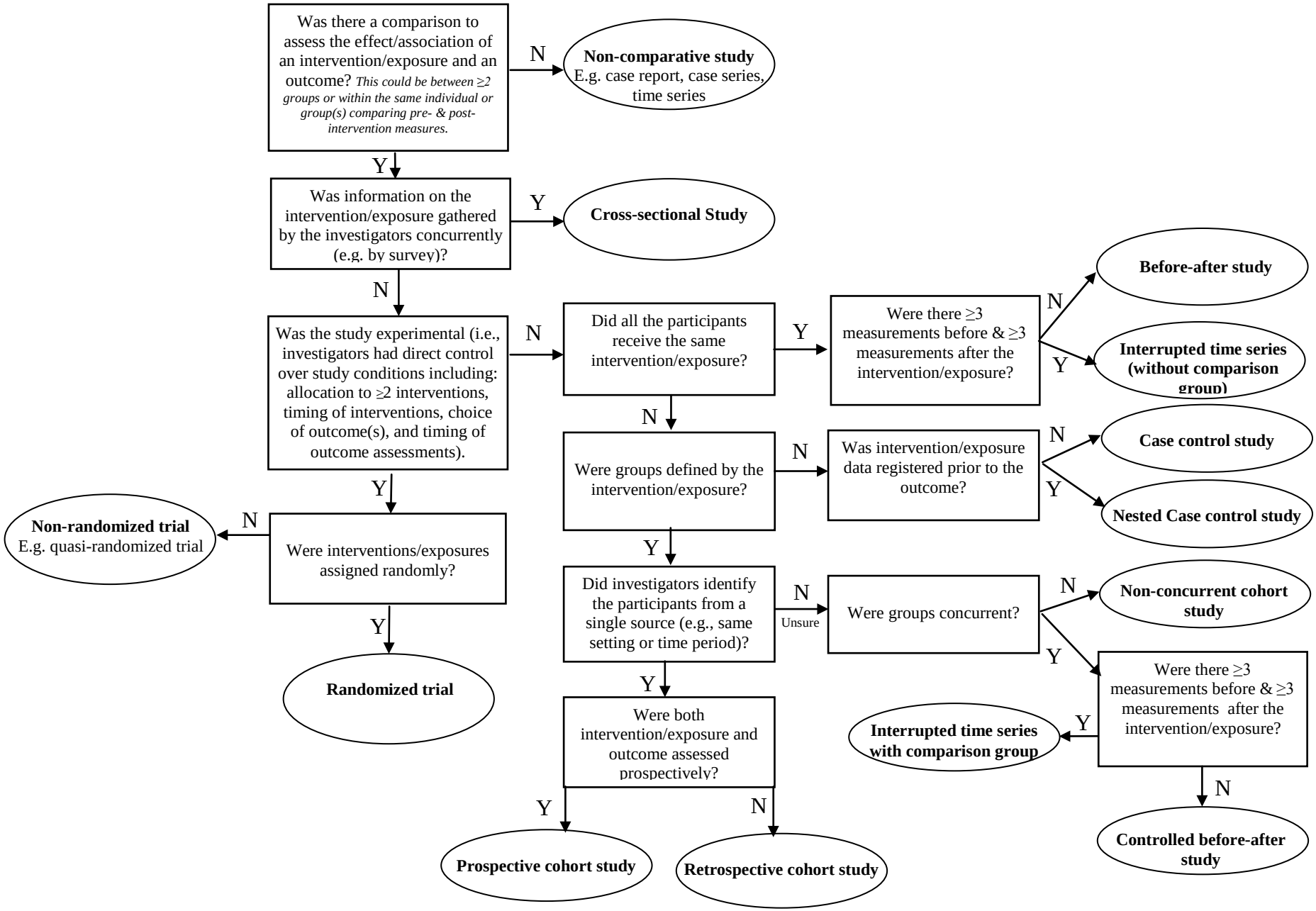
Quasi-experimental study
A type of study in which the investigator(s) evaluates the effect of an intervention but does not have full control over the timing, course, or allocation of the intervention. They are often used when it is not possible to conduct a true experimental study.

Appendix F. Changes Made Between Round One and Round Two Algorithm

Round One	Round Two
Was there a comparison between interventions/exposures/observations (including before and after measurements on the same individuals)?	Was there a comparison to assess the effect/association of an intervention/exposure and an outcome? This could be between ≥2 groups or within the same individual or group(s) comparing pre- and post-intervention measures.
Were disease and exposure information assessed concurrently (e.g., by survey)?	Was information on the intervention/exposure gathered by the investigators concurrently (e.g., by survey)?
Was the study experimental (i.e., investigators had direct control over study conditions including decisions around allocation and timing of interventions)?	Was the study experimental (i.e., investigators had direct control over study conditions including: allocation to ≥2 interventions, timing of interventions, choice of outcome(s), and timing of outcome assessments?
Was more than one group compared?	Did all the participants receive the same intervention/exposure?
Was there a single cohort?	Did investigators identify the participants from a single source (e.g., same setting or time period)?
Were at least three measurements made before and after intervention/exposure?	Were there ≥3 measurements before and ≥3 measurements after the intervention/exposure?
Were at least three measurements made before and after intervention/exposure?	Were threre ≥3 measurements before and ≥3 measurements after the intervention/exposure?

Appendix G. Round Two Algorithm and Glossary

*Before beginning identify the **P**opulation, **I**ntervention/exposure, and key **O**utcomes of the study.*



DESIGN ALGORITHM FOR STUDIES OF INTERVENTIONS AND EXPOSURES

When using the algorithm, it is recommended that you do not rely on the design labels assigned by the authors of the report, but rather work through the questions in the algorithm based on the methods presented in the report and the definitions provided below.

Study Design Key

Below is a list of definitions that correlate with study designs assigned by the accompanying taxonomy. At the end of this list are some additional concepts that may be useful during study design classification.

Non-randomized trial
A study in which individuals or groups of individuals (e.g., community, classroom) are assigned to the intervention or control by a method that is not random (e.g., date of birth, date of admission, judgement of the investigator). Individuals or groups are followed prospectively to assess differences in the outcome(s) of interest. The unit of analysis is the individual or the group, as appropriate.

Randomized trial
A study designed to test the efficacy of an intervention on an individual, a group of individuals, or clusters (e.g., classrooms, communities). Individuals or clusters are randomly allocated to receive an intervention or control/comparison (e.g., placebo or another intervention) and are followed prospectively to assess differences in outcomes. The unit of analysis is the individual, group of individuals, or the cluster, as appropriate. Variations in treatment assignment and measurement produce different types of studies including factorial, cross-over, parallel, stepped wedge and Solomon four-group.

Prospective cohort study
A study in which individuals in the group without the outcome(s) of interest (e.g., disease) are classified according to exposure status at baseline (exposed or unexposed) and then are followed over time to determine if the development of the outcome of interest is different in the exposed and unexposed groups.

Retrospective cohort study
A study in which a group of individuals is identified on the basis of common features that were determined in the past. The group is usually assembled using available data sources (e.g., administrative data). Individuals are classified according to exposure status (exposed or unexposed) at the time the group existed and are followed up to a prespecified endpoint to determine if the development of the outcome of interest is different in the exposed or unexposed groups.

Interrupted time series with comparison group
A study in which multiple observations over time are “interrupted” by an intervention or exposure and in which two series are examined (one is a comparison group). There must be at least 3 observations before and at least 3 observations after the intervention or exposure for each group. The investigator(s) does not assign or have control over the intervention/exposure, which may be an environmental variable (e.g., airborne toxin) or administrative assignment (e.g., seatbelt legislation, educational program, service delivery model) but does have control over the timing of the measurement and the variables being measured.

Controlled before-after study
A study in which the outcome(s) of interest is measured both before and after the intervention or exposure in two or more groups of individuals. In this study design the study group receives the intervention or exposure and the comparison group(s) does not. This type of study includes interventions that may be in the control of the investigator (e.g., a surgical procedure) as well as interventions that may be an environmental variable (e.g., airborne toxin) or administrative assignment (e.g., seatbelt legislation). In all cases the investigator(s) has control over the timing of the measurement and the variables being measured.

Non-concurrent cohort study
A study in which 2 or more groups of individuals are identified on the basis of common features at different time points. Individuals in each group are classified according to exposure status (exposed or unexposed) at the time the groups existed or were created. They are followed to determine if the development of the outcome of interest is different in the exposed or unexposed groups.

Nested case control study
A study where exposed and control subjects are drawn from the population of a prospective cohort study. Baseline data are obtained at the time the population is identified; the population is then followed over a period of time. The study is then carried out using persons in whom the disease or outcome has developed and a sample of those who have not developed the outcome of interest (controls).

Case control study
A study in which participants are selected based on the known outcome(s) of interest (e.g., disease, injury). Exposure status is then collected based on the participants' past experiences. Exposure status is compared between the two (or more) groups: those who have the outcome of interest and those who do not have the outcome of interest (controls). This is a retrospective study that collects data on events that have already occurred.

Interrupted time series (without a comparison group)
A study in which multiple observations over time are "interrupted" by an intervention or exposure. There must be at least 3 observations before the intervention and at least 3 observations after the intervention; otherwise, the study is considered a before-after study. The investigator(s) does not assign or have control over the intervention/exposure, which may be an environmental variable (e.g., airborne toxin) or administrative assignment (e.g., seatbelt legislation, educational program, service delivery model) but does have control over the timing of the measurement and the variables being measured.

Before-after study
A study of an intervention or exposure in which the investigator(s) compares the outcome(s) of interest both before and after the intervention in the same group of individuals. This includes interventions that may be in the control of the investigator (e.g., a surgical procedure) as well as interventions that may be an environmental variable (e.g., airborne toxin) or administrative assignment (e.g., seatbelt legislation). In all cases the investigator(s) has control over the timing of the measurement and the variables being measured.

Cross-sectional study
A study in which both the exposure and the outcome status in a target population are assessed concurrently, that is, at the same point in time or during a brief period of time. The temporal sequence of cause and effect cannot necessarily be determined. They are most commonly used to assess prevalence. A common method for data collection is a survey.

Non-comparative study
Examples of this design include:

- A study that presents a description of a single patient or participant. Studies are usually retrospective and typically describe the manifestations, clinical course, and prognosis of the individual.
- A study that describes the experience of a group of patients with a similar diagnosis and/or treatment. Studies are usually retrospective and typically describe the manifestations, clinical course, and prognosis of a condition.
- A study in which data are collected at a series of points in time on the same population to observe trends in the outcome(s) of interest.

Additional Concepts

Cluster
The term 'cluster' refers to a unit of allocation or analysis in a clinical trial. Examples of clusters include hospitals, schools, neighborhoods, or entire communities.

Cluster randomized controlled trial
Synonym: *community trial; group randomized trial*
A randomized controlled trial in which the units of randomization and analysis are groups of people or communities (e.g., classroom, hospital, town). Typically, several communities receive the intervention and several different communities serve as controls.

Cohort
The term 'cohort' refers to a group of individuals (or other organizational units) who have a common feature when they are assembled (e.g., birth year, place of employment, medical condition, place or time period of medical treatment) and are followed over time. They can be followed prospectively or examined retrospectively.

Experimental study
A type of study in which investigators have direct control over the timing, course, and assignment of the intervention. Experimental studies investigate an intervention to determine its effect on the outcome(s) of interest. In an experimental study a population is selected to receive a specific intervention the effects of which are measured by comparing the outcomes in the experimental group with the outcomes of a control group that has received another intervention or placebo. Examples include randomized controlled trial, cluster randomized controlled trial, nonrandomized trial, n-of-one trial. See also **observational study**.

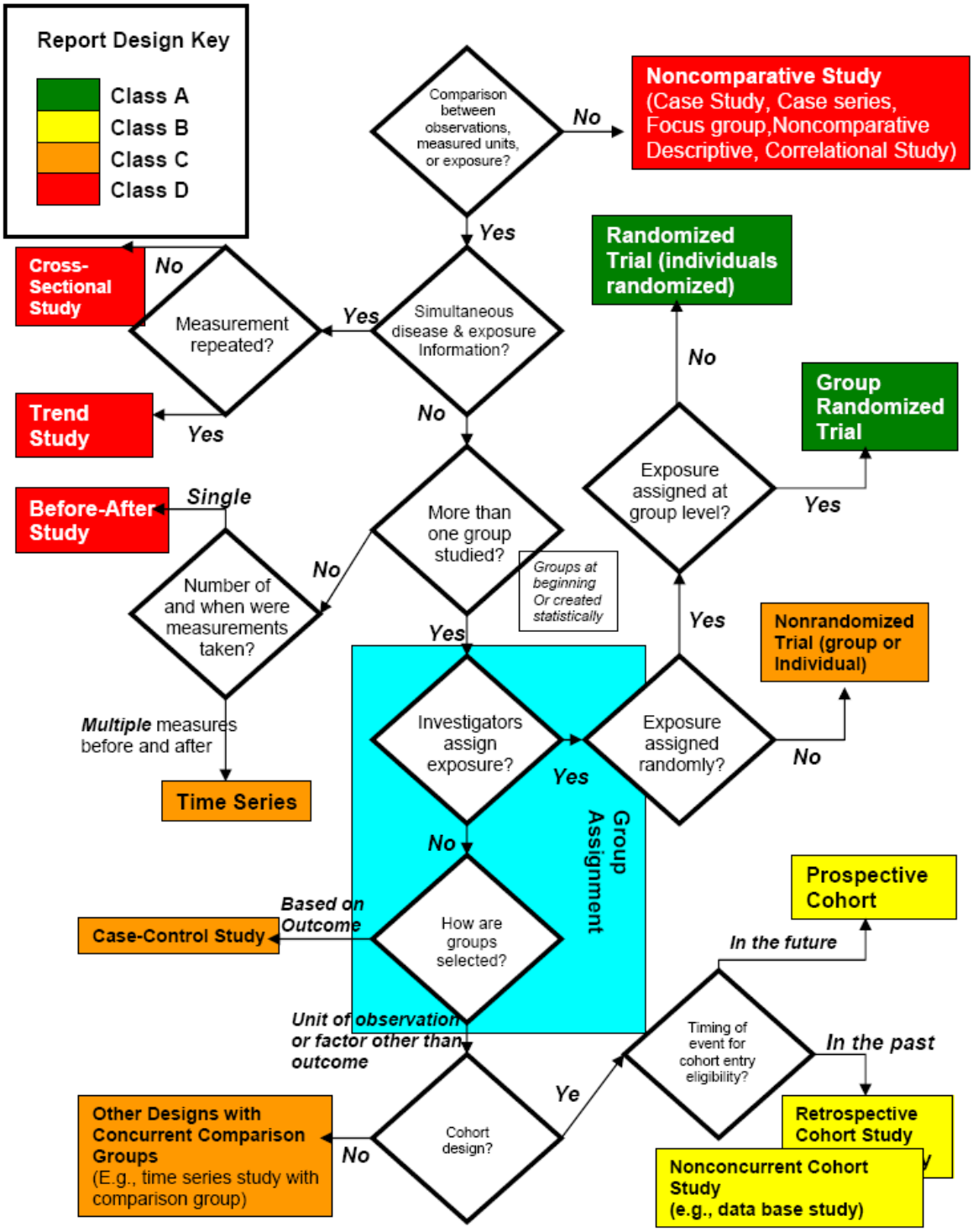
Observational study
A study in which the investigator(s) does not control the exposure/ intervention status of study participants (i.e., the assignment of the intervention or exposure of interest is not under the control of the investigator(s)). The simplest form of observational study is the case report or case series, which describes the clinical course of individuals with a particular condition or diagnosis.
Observational studies include descriptive and analytic studies. See also **experimental study**.

Quasi-experimental study
A type of study in which the investigator(s) evaluates the effect of an intervention but does not have full control over the timing, course, or allocation of the intervention. They are often used when it is not possible to conduct a true experimental study.

Appendix H. Top Ranked Classification Tools Design Algorithm for Studies of Health Care Interventions

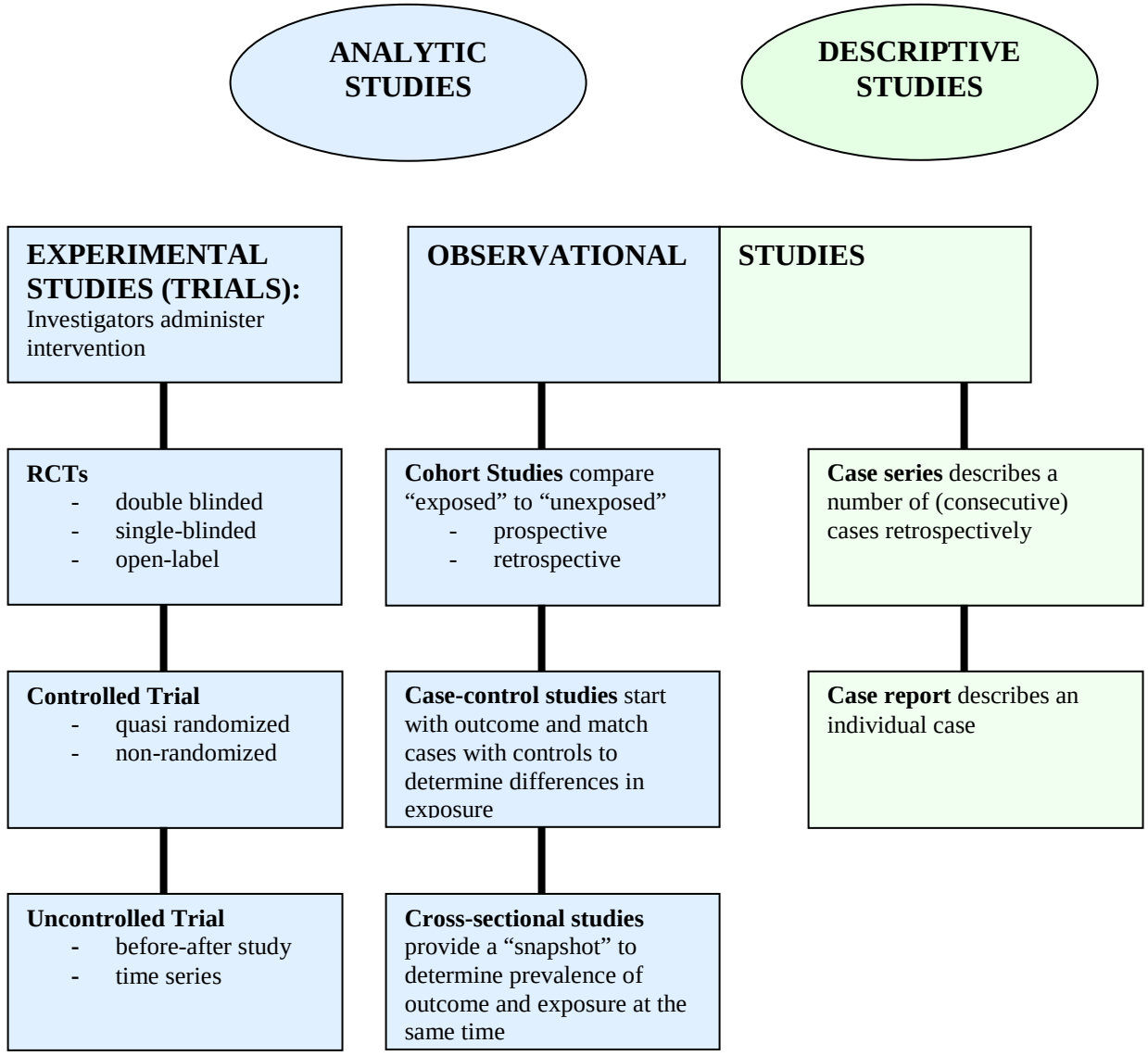
from http://www.cochrane.dk/nrsmg/docs/chap2fig.gif. Accessed January 15, 2009.

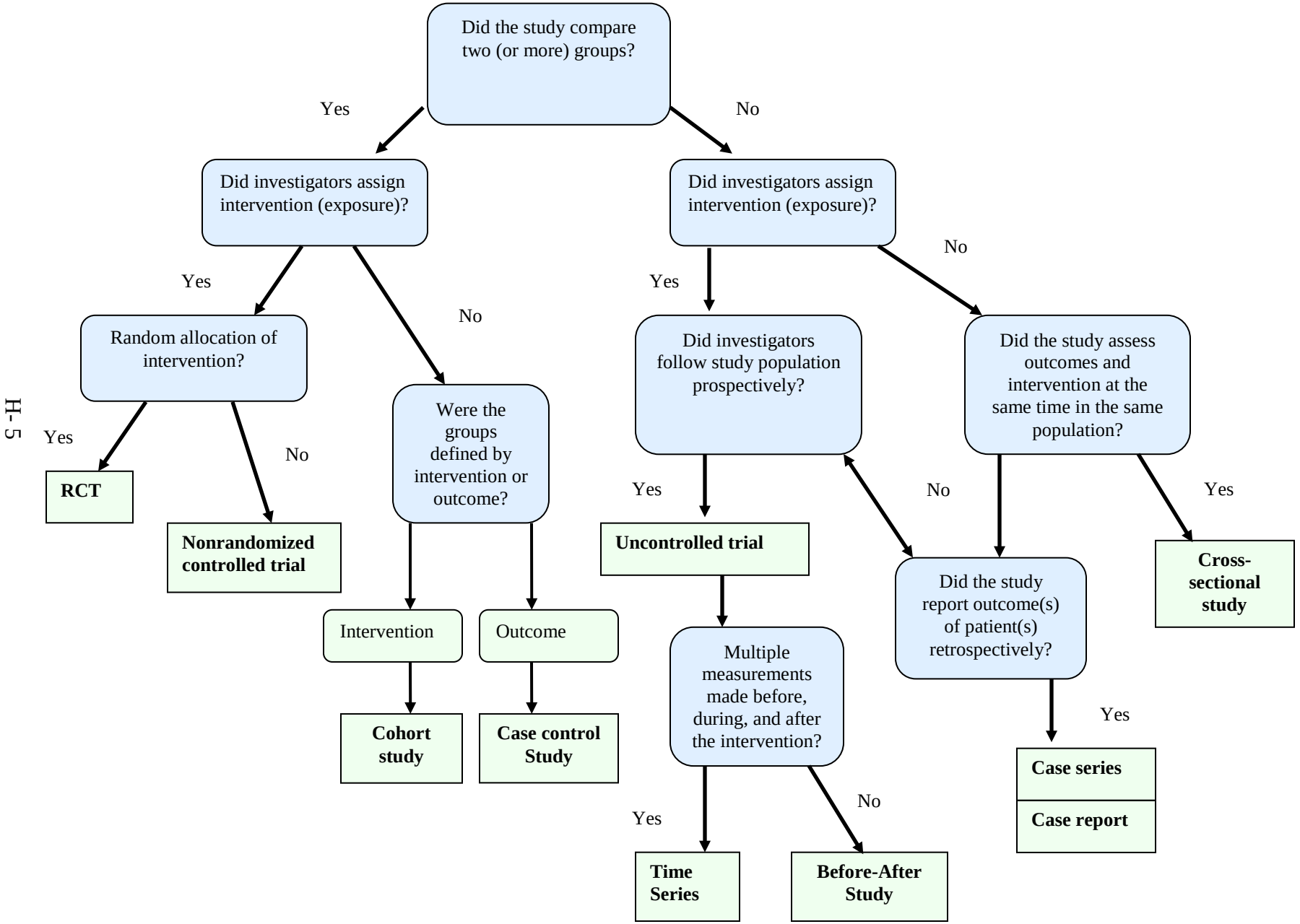
American Dietetic Association* (©American Dietetic Association. Reprinted with permission.)



*This algorithm was updated in the Fall of 2010. The updated version is available at: http://www.adaevidencelibrary.com/topic.cfm?cat=1315

RTI–UNC


ANALYTIC STUDIES
DESCRIPTIVE STUDIES
EXPERIMENTAL STUDIES (TRIALS): Investigators administer intervention
OBSERVATIONAL
STUDIES
RCTs
- double blinded
- single-blinded
- open-label
Cohort Studies compare "exposed" to "unexposed"
- prospective
- retrospective
Case series describes a number of (consecutive) cases retrospectively
Controlled Trial
- quasi randomized
- non-randomized
Case-control studies start with outcome and match cases with controls to determine differences in exposure
Case report describes an individual case
Uncontrolled Trial
- before-after study
- time series
Cross-sectional studies provide a "snapshot" to determine prevalence of outcome and exposure at the same time


Did the study compare two (or more) groups?
Yes
No
Did investigators assign intervention (exposure)?
Did investigators assign intervention (exposure)?
Yes
No
Yes
No
Random allocation of intervention?
Were the groups defined by intervention or outcome?
Did investigators follow study population prospectively?
Did the study assess outcomes and intervention at the same time in the same population?
Yes
No
Yes
No
RCT
Nonrandomized controlled trial
Intervention
Outcome
Uncontrolled trial
Did the study report outcome(s) of patient(s) retrospectively?
Cross-sectional study
Cohort study
Case control Study
Multiple measurements made before, during, and after the intervention?
Yes
Case series
Case report
Yes
No
Time Series
Before-After Study

H-5

Accompanying Glossary for RTI-UNC Algorithm

Analytic Studies
Analytic studies are designed to examine causal associations (e.g. treatment efficacy). Investigators assemble groups of individuals for the specific purpose of determining a cause- effect relationship. Analytic studies employ control groups to compare exposed (i.e. treated) to unexposed groups (e.g. placebo) to determine differences in outcomes. In head-to-head studies, outcomes of two or more different exposures (treatments) are assessed. Analytic studies can be divided into two broad design strategies: experimental studies and observational studies.

Before-After Study (Pre-post Study)
Investigators administer intervention and follow participants prospectively. Patients serve as their own controls, that is, the patients' pre-data is compared with the same patients' post-data. If multiple outcome measurements are conducted during the follow-up period the study becomes a time series. Regression to the mean and non-specific effects of procedures threaten the validity of before and after studies and can cause misleading inferences.

Case Control Study
Case control studies identify patients with and without a given outcome (cases and controls, respectively). Patient characteristics should be similar and controls must be representative of those individuals who would have been selected as cases, had they developed the disease. Investigators look back in time and determine the exposure status for cases and controls. Case-control studies are always retrospective.

Case Report
Case reports present outcomes of individual cases (disease at an individual level) Often they report rare or new diseases or rare outcomes. Case reports can be valuable for assessing rare but severe adverse events.

Case Series
Case series aggregate individual cases in one report. Case series are defined by a group of people with similar diagnoses or undergoing the same procedure over time. Case series are aggregated retrospectively using registries or clinical records. Ideally, case series report on consecutive cases of a well-described study population with a well-defined intervention using validated outcome measures. Prospective case series resemble before-after studies or time series. Case series are useful for hypothesis generation, improved case definition, or detection of rare adverse events. Case series do not have comparison groups and therefore cannot test treatment efficacy. Historical controls often differ in co-interventions and other baseline characteristics and can be misleading.

Cohort Study
A cohort study follows two or more groups from exposure to outcome. Groups are defined by being exposed or unexposed. Ideally, groups are similar with respect to all other patient characteristics except exposure. Cohort studies can be prospective or retrospective. In prospective cohort studies, participants

are tracked forward in time (investigators age with participants). In retrospective cohort studies, investigators look backward to choose cohorts.

Cross-sectional Study
Cross-sectional studies examine the relationship between disease and other variables of interest at one particular point in time ("snapshot"). Cross-sectional studies are most commonly used to assess prevalence data. A common format is surveys.

Descriptive Studies
Descriptive studies provide a better understanding of the general characteristics of a disease or treatment. They are important for hypothesis generation and are often the first emerging evidence. Descriptive studies use information from various sources such as census data, registries or clinical records. They do not have comparison groups and therefore do not allow assessments of association (i.e. relative risks [benefits], odds ratios, absolute and relative risk reductions). Furthermore, descriptive studies cannot provide any evidence about causation or treatment efficacy. Two main types of descriptive treatment studies exist: Case reports or case series; and cross sectional studies of individuals.

Experimental Studies (Trials)
A study in which investigators assign the intervention and follow study participants prospectively to assess differences in outcomes. Experimental studies are RCTs, nonrandomized controlled trials, before-after studies, and time series. Experimental studies are conducted prospectively.

Nonrandomized Controlled Trials
Investigators assign the intervention and have control over the allocation of participants to groups (no underlying random process). Selection bias is frequently an issue in nonrandomized studies.

Observational Studies (non-experimental studies)
Investigators do not assign an intervention but rather observe groups with and without an intervention. Observational studies can be conducted prospectively (investigator ages with the participants) or retrospectively.

Randomized Controlled Trial (RCT)
Participants are randomly allocated to intervention or control and followed over time to assess differences in outcomes. Randomization and allocation concealment (i.e. allocation of the randomization sequence) ensures that known and unknown prognostic factors (confounders) are distributed equally between groups. RCTs are considered the gold standard to assess treatment effects. RCTs can be blinded (masked) or unblinded. Except for outcomes that do not involve any subjective judgment (e.g. overall mortality: death–no death) blinding of outcomes assessors, patients, and health care providers is important to minimize measurement bias. A disadvantage of RCTs is that study populations are often highly selected and results therefore sometimes lack generalizability (external validity).

Time Series
Investigators administer intervention and follow participants prospectively. Patients serve as their own controls, that is, the patients' pre-data are compared with the same patients during and after the intervention. Multiple outcome measurements are conducted during the follow-up period of the study.

Regression to the mean and non-specific effects of procedures threaten the validity of time series and can cause misleading inferences.

Uncontrolled Trial
Investigators assign intervention without a control group. Depending on the timing of the outcomes assessments uncontrolled trials can be before-after studies or time series.