Methods Research Report

Development of Quality Criteria To Evaluate Nontherapeutic Studies of Incidence, Prevalence, or Risk Factors of Chronic Diseases: Pilot Study of New Checklists

AHRQ
Agency for Healthcare Research and Quality
Advancing Excellence in Health Care • **www.ahrq.gov**

This report is based on research conducted by the Minnesota Evidence-based Practice Center under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No. 290-2002-0009). The findings and conclusions in this document are those of the author(s), who are responsible for its content, and do not necessarily represent the views of AHRQ. 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 the 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 or actions may not be stated or implied.

Methods Research Report

Development of Quality Criteria To Evaluate Nontherapeutic Studies of Incidence, Prevalence, or Risk Factors of Chronic Diseases: Pilot Study of New Checklists

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. 290-02-0009

Prepared by:
Minnesota Evidence-based Practice Center
Minneapolis, Minnesota

Investigators:
Tatyana A. Shamliyan, M.D.
Robert L. Kane, M.D.
Mohammed T. Ansari, M.B.B.S.
Gowri Raman, M.D.
Nancy D. Berkman, Ph.D.
Mark Grant, M.D., M.P.H.
Gail Janes, Ph.D., M.S.
Margaret Maglione, M.P.P.
David Moher, Ph.D.
Mona Nasser, D.D.S.
Karen A. Robinson, M.D.
Jodi B. Segal, M.D.
Sophia Tsouros

AHRQ Publication No. 11-EHC008-EF
January 2011

The information in this report is intended to help clinicians, employers, policymakers, and others ake 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.

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

Suggested citation: Shamliyan T, Kane RL, Ansari MT, Raman G, Berkman ND, Grant M, Janes G, Maglione M, Moher D, Nasser M, Robinson KA, Segal JB, Tsouros S. Development of the Quality Criteria To Evaluate Nontherapeutic Studies of Incidence, Prevalence, or Risk Factors of Chronic Diseases: Pilot Study of New Checklists. Agency for Healthcare Research and Quality; January 2011. Methods Research Report. AHRQ Publication No. 11-EHC008-EF. Available at http://effectivehealthcare.ahrq.gov/.

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 they 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 evidence report. 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

Carmen Kelly, Pharm.D., M.P.H., R.Ph.
Task Order Officer
Agency for Healthcare Research and Quality

Investigator Affiliations

Tatyana A. Shamliyan, M.D.[a]
Robert L. Kane, M.D.[a]
Mohammed T. Ansari, M.B.B.S.[b]
Gowri Raman, M.D.[c]
Nancy D. Berkman, Ph.D.[d]
Mark Grant, M.D., M.P.H.[e]
Gail Janes, Ph.D., M.S.[f]
Margaret Maglione, M.P.P.[g]
David Moher, Ph.D.[b]
Mona Nasser, D.D.S.[h]
Karen A. Robinson, M.D.[i]
Jodi B. Segal, M.D.[i]
Sophia Tsouros[b]

[a]Division of Health Policy and Management, University of Minnesota School of Public Health, Minneapolis, MN, USA
[b]Ottawa Methods Centre, Clinical Epidemiology Program, Ottawa Health Research Institute, Ottawa, Ontario, Canada
[c]Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Boston, MA, USA
[d]RTI International, Research Triangle Park, NC, USA
[e]Blue Cross and Blue Shield Association, Chicago, IL, USA
[f]Centers for Disease Control and Prevention, Atlanta, GA, USA
[g]Southern California EPC, RAND Corporation, Santa Monica, CA, USA
[h]German Institute for Quality and Efficiency in Health Care, Cologne, Germany
[i]Johns Hopkins University School of Medicine, Baltimore, MD, USA

Acknowledgments

We would like to thank our reviewers, David Atkins, M.D., John Hoey, M.D., and Christine Laine, M.D., for reviewing and commenting on the draft; our collaborating experts, Ethan Balk, M.D., Chantelle Garritty, and Thomas Trikalinos, Ph.D., for their scientific input throughout this project; and Carmen Kelly, Pharm.D., our Task Order Officer from AHRQ, and Stephanie Chang, M.D., M.P.H., Medical Officer at AHRQ, for their guidance throughout the project. We also want to thank the librarian, Judith Stanke, for her contributions to the literature search; research assistants Stacy Dickinson, M.P.H., Emily Zabor, M.S. candidate in biostatistics, and Akweley Ablorh, M.S. candidate in biostatistics, for data abstraction, quality control, and synthesis of evidence; Zhihua Bian M.S. candidate in biostatistics, for her statistical help; Zhiyuan Xu, M.S. candidate in applied economics, for his work creating the ACCESS database; Dean McWilliams for his assistance in database development; Qi Wang, research fellow, for her statistical expertise in reliability testing; Susan Duval, Ph.D., for her help estimating sample size; Marilyn Eells for editing and formatting the report; and Nancy Russell and Rebecca Schultz for their assistance gathering data from the experts and formatting the tables.

Development of Quality Criteria To Evaluate Nontherapeutic Studies of Incidence, Prevalence, or Risk Factors of Chronic Diseases: Pilot Study of New Checklists

Structured Abstract

Objective. To develop two checklists for the quality of observational studies of incidence or risk factors of diseases.

Study design and setting. Initial development of the checklists was based on a systematic literature review. The checklists were refined after pilot trials of validity and reliability were conducted by seven experts, who tested the checklists on 10 articles.

Results. The checklist for studies of incidence or prevalence of chronic disease had six criteria for external validity and five for internal validity. The checklist for risk factor studies had 6 criteria for external validity, 13 criteria for internal validity, and 2 aspects of causality. A Microsoft Access database produced automated standardized reports about external and internal validities. Pilot testing demonstrated face and content validities and discrimination of reporting vs. methodological qualities. Interrater agreement was poor. The experts suggested future reliability testing of the checklists in systematic reviews with preplanned protocols, a priori consensus about research-specific quality criteria, and training of the reviewers.

Conclusions. We propose transparent and standardized quality assessment criteria of observational studies using the developed checklists. Future testing of the checklists in systematic reviews is necessary to develop reliable tools that can be used with confidence.

Contents

Figures

Tables

Appendixes

Introduction

The prevalence and incidence of chronic conditions have implications for policy and healthcare utilization. Valid information about risk factors is important in reducing the burden of chronic diseases.[1-2] Although systems to rank the strength of the recommendations about effective interventions consider all evidence from observational studies as low,[3-4] prevalence and risk factors for chronic diseases can be evaluated only in observational studies.[5] Public policy decisions should be based on applicable and unbiased results from high quality studies.[6-8] Assessing the quality of observational studies is an important part of evidence-based reports made for the Agency for Healthcare Research and Quality (AHRQ).[9]

An extensive review of all available systems for rating the strength of scientific evidence and concluded that future efforts need to identify valid and reliable quality ratings for observational studies.[9-10] Different methodological aspects, including selective treatment assignment, access to health care, or provider characteristics may have different importance for studies that examine treatment effects and prevalence of chronic conditions or the association of disease risk factors with patient mortality and morbidity.[9-10] Therefore, quality evaluation that is part of grading of a body of evidence must be tailored to the methodological aspects and quality standards of nontherapeutic observational studies.

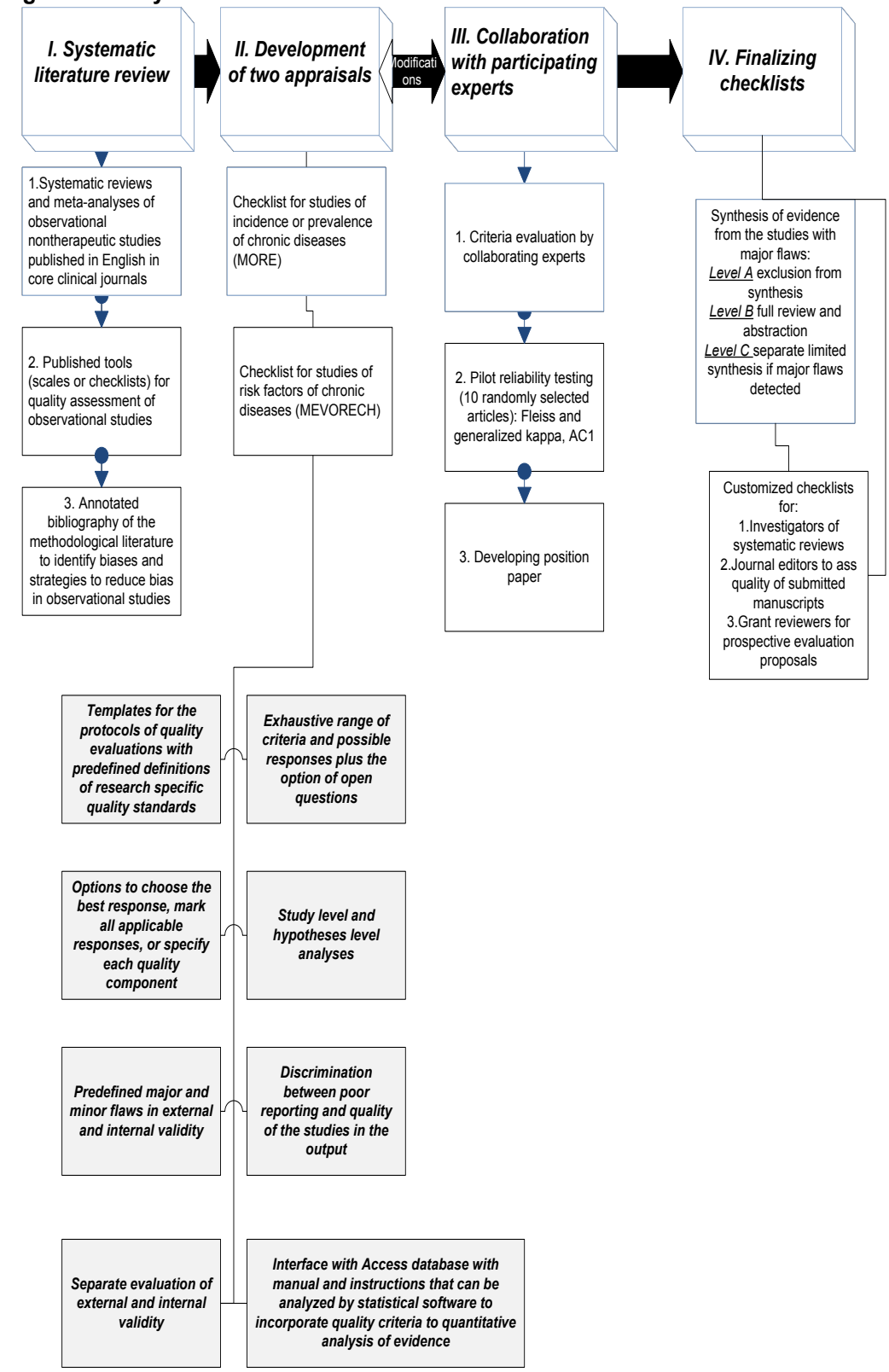
The present collaborative project sought to develop valid and reliable quality criteria of observational studies that examine the incidence or prevalence of chronic conditions and risk factors for diseases. We propose criteria for the design, reporting standards, and assessment of nontherapeutic observational studies in systematic reviews and evidence-based reports.

Methods

We developed two checklists, one for studies of incidence or prevalence and another for risk factors, based on our literature review and in collaboration with experts from other Evidence-based Practice Centers and the Centers for Disease Control and Prevention.

The protocol to construct the checklists was based on a conceptual model of the development of indexes, rating scales, or other appraisals to describe and measure symptoms, physical signs, and other clinical phenomena in clinical medicine (Figure 1).[11] We defined external validity as the extent to which the results of a study can be generalized to the target population.[5] Applicability may differ from external validity by the definition of the target population; well designed studies from different countries with good external validity can have low applicability to the U.S. population. We defined internal validity as the extent to which the results of a study are correct for the subjects and the associations detected are truly caused by exposure.[5] We defined biases the checklists should address, but avoided labeling biases in quality evaluation because of differences in definitions of biases and because of applicability of previously labeled selection, information, differential verification, context, treatment paradox, disease progression, and other biases to interventional studies.

Figure 1. Analytical framework



Selection of Candidate Variables

We reviewed all previously published checklists and scales for quality assessment of observational studies.[12-95] Then we generated a bank of criteria items by applicability to observational studies of incidence, prevalence, or risk factors, and by assessment of external or internal validity. Finally, we selected all components relevant to studies of incidence, prevalence, or risk factors of chronic conditions. We included possible responses for comprehensive objective quality evaluation with minimal evaluator judgments around quality criteria.

Composition of the Checklists

Based on the results of the literature review, we formulated the requirements for the proposed checklists to assess quality of observational nontherapeutic studies (Figure 1). We developed two checklists (Appendix 1); one for incidence or prevalence studies and one for risk factors following the requirements we formulated. The checklists were designed to reflect the best (gold standard) methodology that the CDC uses to conduct Public Health Surveillance for Chronic Conditions for incidence or prevalence studies. The quality evaluators have the flexibility to define biases that can be specific for research questions. The checklists are available in text format (Appendix A) or as a relational database (Microsoft Access), which can be downloaded from https://netfiles.umn.edu/xythoswfs/webui/_xy-17471658_1-t_aRG151Im.

The checklists address important biases to which cohort, cross-sectional, and case-control studies may be susceptible. The checklist for studies of risk factors included two aspects of causality-dose response and temporal association between risk factors and disease outcomes. Grading the level of evidence requires additional information about consistency of results across studies and should not be part of the standard output for individual studies. Basic knowledge of epidemiology is required to complete the checklists.

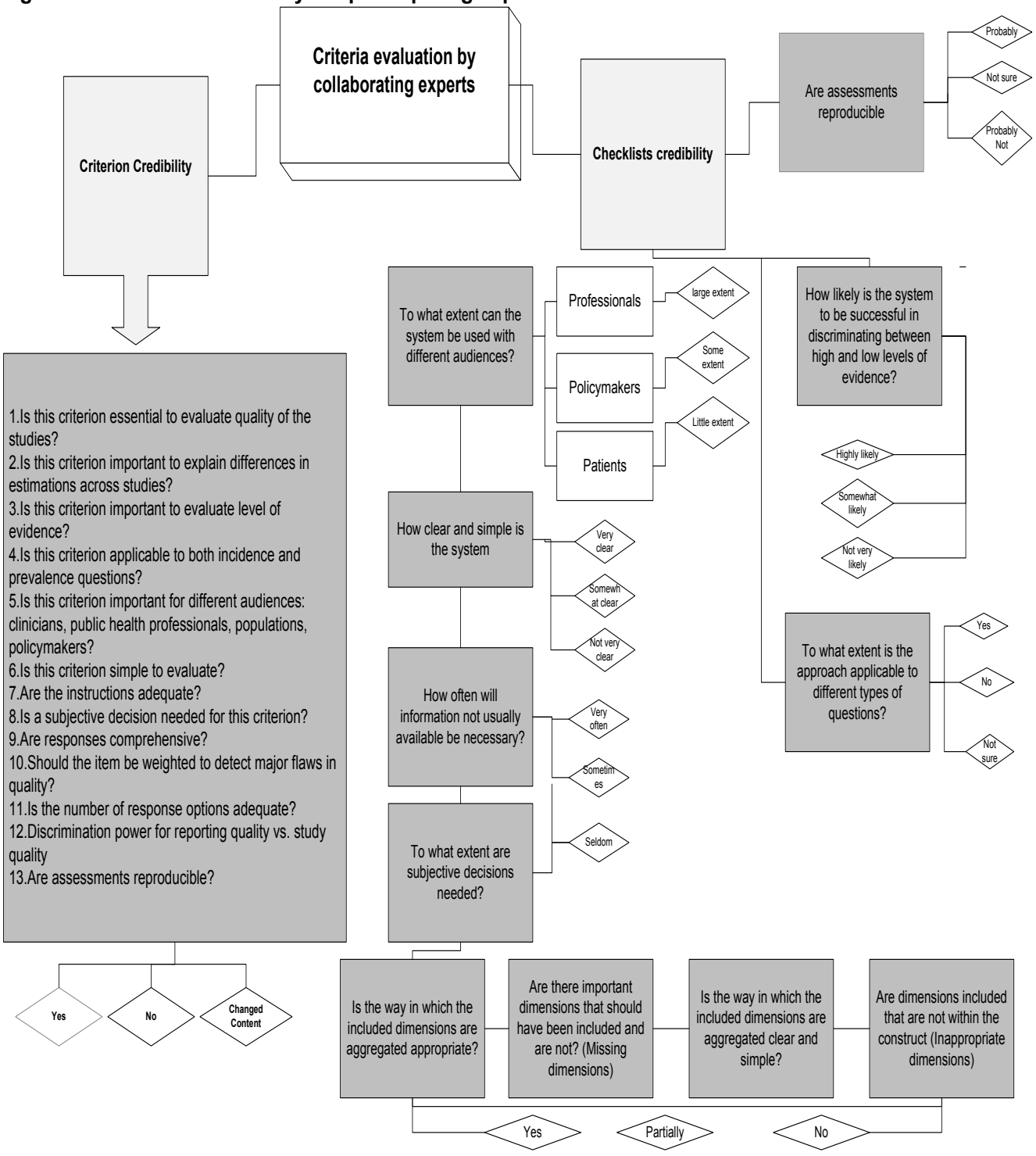
We discriminated reporting quality from methodological quality by having the option of "not reported" for all quality criteria. The proposed checklists discriminate methodological quality using Boolean operators; for example, when the response to the question about sampling of subjects is X (clinic based recruitment), then the study has a major flaw in external validity. If the evaluators chose responses predefined as poor reporting or methodological quality, the Access internal algorithm would list them as present major or minor flaws in methodological quality or not reported quality criteria.

The definitions of minor flaws can be research specific but must be pre-specified and justified in the protocols of quality evaluations to ensure that the evaluations are based on quality components rather than formal scaling of the criteria. Part one of the report lists poorly reported quality criteria and flaws in external validity; part two lists internal validity without formal scaling of criteria or summarizing them into global arithmetic scores or obscured nontransparent quality rank.

Pilot Testing of Validity and Reliability by Experts

We invited members of the Evidence-based Practice Centers and the Centers for Disease Control and Prevention to participate in the project. Responders from the U.S., Canada, and Germany evaluated face and content validity of the checklists (Figure 2).[3,11] Two centers had more than one volunteering expert; but each center represented one opinion about quality criteria and submitted one completed checklist per article.

Figure 2. Criteria evaluation by the participating experts



We conducted a pilot test of the checklists. The experts each evaluated ten articles to test reliability and discriminant validity.[96-105] We randomly selected the articles from the reference lists of several completed EPC reports about incidence or prevalence or risk factors of chronic diseases.[10,106-109] We aimed to develop quality assessments of individual studies that represent

different topics of biomedical research. We expected the same level of reliability for different topics.

The outcomes for each step of the collaboration included**:**

(1) Credibility test of each quality criterion, domains of external and internal validity, and overall conclusions about quality (Figure 2).[11]
(2) Reliability evaluation: agreement, generalized[110] and Fleiss kappa[111-112] and AC1 statistics[113-115] overall, by topic, by article, by domains of external and internal validity, and by quality component.
(3) Revision of the appraisals according to comments from the experts.

We modified both checklists for different audiences, depending on the goals of quality appraisals: retrospective quality assessment of individual studies for systematic reviews or quality assessment of submitted manuscripts.

Analysis of Credibility, Validity, and Reliability of the Checklists

In the absence of a gold standard, we created an analytical plan based on published validations of previously peer reviewed checklists[25,62] and scales for quality assessment[49-50] such as the Overview Quality Assessment Questionnaire.[25] The index assessed quality of evaluated systematic reviews compared with the best published systematic reviews or meta-analyses with six of the seven hypotheses used to test construct validity held true. Sensibility of the index was evaluated with 13 questions and a seven point scale, the mean rating was 5 or greater, indicating good sensibility of the tool.[25] The authors validated the methodological index for nonrandomized studies (MINORS)[74] and reported that scores estimating nonrandomized therapeutic studies were significantly lower when compared with well designed randomized controlled clinical trials. Credibility criteria of the index were assessed on a 7 point scale. Good credibility was defined with a total score of 13 or more.

We used the same approach to evaluate the validity of the checklists with an agreement around criteria. We analyzed a proportion of positive responses among all binary responses about face and content validity. We collected open-ended comments and suggestions to modify the checklists. We scored all responses as 0, 0.5, or 1 and calculated the means and standard deviations of total scores. We did not score missing responses and did not include them in the analyses. For binary responses, we added criteria to the checklists when more than 75 percent of the experts agreed on validity.[25] We concluded that a criterion was credible when the mean score was ≥6 on a scale of 9.[74]

We examined discriminant validity by testing the hypothesis that the checklists can detect differences in quality across studies, and discriminate reporting vs. methodological quality.[74] We assigned scores for responses detecting poor reporting or the presence of major or minor flaws in the methodological quality of the articles. Then we calculated means and standard deviations for the total score of poor reporting and major or minor flaws in each article. The significant mean differences in total scores at 95-percent confidence level indicated the discriminating power of the checklists to detect poor reporting vs. poor methodological quality.

To test reliability of the checklists, we estimated that we needed to ask at least five experts to review ten publications for kappa=0.8 and null hypothesis of kappa=0.5 based on alpha=0.05 and beta=0.2.[116]

We measured agreement among raters separately for all ten of the papers selected, using Landis & Koch's measure of inter-rater agreement for multiple raters (with studies in place of subjects).[111] Because of the small sample size and the lack of an analytical standard error, we computed 95-percent confidence intervals for Landis & Koch's measure using a bootstrap procedure[112] with 1,000 bootstrap samples, each sample consisting of the appropriate number of papers sampled with replacement. We report the 2.5th and 97.5th percentiles of these bootstrap samples as a 95-percent confidence interval. All computing was done using a custom code in the R system (v. 2.3.1).[117-118]

We also calculated generalized kappa[110] and AC1 statistics for each quality component and each article[113-115] (Appendix B Exhibit 1) using Excel[110] (Appendix B Exhibit 2) and SAS[114,119] software (codes are available by request from the authors). We calculated a generalized kappa statistic for each article using the number of raters (seven) who marked the same response category.[111] The checklist for incidence or prevalence studies had 18 subjects and 11 response categories. The checklist for risk factor studies had 26 subjects and nine response categories. We calculated the total number of ratings, generalized kappa for each response category, and the total generalized kappa for the study (Appendix B Exhibit 2).

Kappa statistics are sensitive to differences in rate marginal probabilities and can be paradoxically low in cases when the propensity of positive ratings is very small or very large (high agreement with a high concentration of observations in one cell). In contrast, the AC1 statistic is a more robust measure of agreement among multiple raters because it estimates the likelihood of agreement by chance as the probability that a randomly chosen rater classifies randomly chosen subjects into the same response category. Since none of the statistical tests for reliability of nominal multi-rater responses using checklists is ideal,[11] we compared percentage agreement, Fleiss and generalized kappa, and AC1 statistics to detect areas of disagreement. We interpreted kappa values of 0.0-0.19 as poor, 0.20-0.39 as fair, 0.40-0.59 as moderate, 0.60-0.79 as substantial, and 0.80-1.00 as almost perfect agreement.

Results

We identified 84 publications that described 96 tools[12-95] to assess the quality of observational studies; 47 of 96 tools were created for therapeutic studies; 47 percent were modified from previously published peer reviewed appraisals, 18 percent were developed based on methodological standards; 35 percent did not provide any information about development of the tools; 22 percent reported reliability; and 10 percent reported validation procedures. None of the tools ranked internal validity or applicability; 35 percent categorized quality by the presence of predefined major flaws in design, or by total score. The level of evidence was proposed in 22 percent of the tools, by criteria of causality or internal validity. None of the tools discriminated poor reporting quality from overall quality. None of the tools gave separate conclusions about external and internal validity. Evaluation required different degrees of subjectivity. Numerical estimates of quality did not provide transparent conclusions about the degree of bias.

Evaluation of the Tools by Experts Participating in the Project

Face and Content Validity. Participating experts from nine organizations (of 15 invited) evaluated content and face validity of the checklists (Table 1). The experts found the majority of the criteria proposed for incidence or prevalence or risk factor studies to be valid. Seventy-three percent thought the dimensions in the checklists were appropriate, and 90 percent thought they were complete. They also agreed that suggested responses were comprehensive and reproducible but were difficult to evaluate because of reporting quality and the observational nature of the research. Comments included examples of data mining for purposes different from the aims of the study. We modified the checklists and added a table with definitions and instructions related to suggested responses about external and internal validity (Appendix A). We included the option to mark more than one applicable response and added an option to specify each criterion when suggested responses did not include applicable information.

Table 1. Rating the quality criteria of observational studies

Incidence Studies	Mean ± Standard Deviation	Risk Factor Studies	Mean ± Standard Deviation
Aim of study	5±2	Aim of study	**6**±3
Study design	**6±**2	Objectives	5±3
		*Hypothesis***	5±3
		Study design	**7**±2
External Validity*			
1. Sampling of the subjects by the investigators	**6±**3	1. Sampling of the subjects by the investigators	**7**±3
2. Assessment of sampling bias	**6±**3	2. Assessment of sampling bias	**6**±2
3. Estimation of sampling bias	**6±**2	3. Estimation of sampling bias	**6**±2
4. Exclusion rate from the analysis	**6±**2	4. Exclusion rate from the analysis	**6**±3
5. Sampling bias is addressed in the analysis	5±3	5. Sampling bias is addressed in the analysis	**6**±3
6. Subject flow	**6±**2	6. Subject flow	**6**±2
Internal Validity*			
1. Source to measure outcomes	**7±**2	1. Source to measure outcomes	**7**±2
2. Definition of outcomes	**6±**3	2. Definition of outcomes	**6**±2
3. Measurements of outcomes	**6±**3	3. Measurements of outcomes	**6**±2
4. Outcomes in race, ethnic, age, or gender subpopulations	5±3	4. Definition of the exposure	**6**±2
5. Reporting of outcomes	4±2	5. Measurements of the exposure	**6**±3

Table 1. Rating the quality criteria of observational studies (continued)

Incidence Studies	Mean ± Standard Deviation	Risk Factor Studies	Mean ± Standard Deviation
*Outcomes in groups with risk factors of the outcomes***	5±3	6. Confounding factors	**7**±2
		7. Loss of followup	**7**±2
		8. Masking of exposure status	**6**±2
		9. Statistical analysis	5±1
		Assessment of temporality***	5±2
		10. Appropriateness of statistical models	5±2
		Dose response with exposure***	**6**±2
		11. Reporting of tested hypothesis	**6**±2
		12. Precision of the estimates	5±2
		13. Sample size justification	5±2
*Overall estimation of internal validity***	5±2		5±2
*Overall estimation of external validity***	5±2		**6**±2
*Estimation the level of evidence of the association with risk factors***	5±2		4±2

*good credibility when score mean was ≥6
**criteria were removed from the checklist
***aspects of causality

The experts voted that overall assessments of internal and external validity should not be part of this project, based on the recommendation of the experts (Table 2). We deleted overall conclusions about external and internal validity, did not use qualitative categories of applicability or internal validity (e.g., high, low), but intended to distinguish the studies at least on the basis of the presence of important flaws that could affect the results. We did not propose numerical scaling of quality, quantitative values for criteria, or numerical weighting of flaws.

Table 2. Credibility of the checklists to assess quality of observational studies (evaluation by participating experts)

Quality Criteria	Incidence Mean ± STD	Risk Factors Mean ± STD
Is this criterion essential to evaluate quality of the studies?	**6**±2	**7**±2
Is this criterion important to explain differences in estimations across studies?	**8**±1	**7**±2
Is this criterion important to evaluate level of evidence?	5±2	**6**±2
Is this criterion applicable to questions (prevalence or incidence or risk factors)?	**8**±3	**8**±1
Is this criterion important for different audiences (clinicians, public health professionals, populations, policymakers)?	**8**±1	**7**±2
Is this criterion simple to evaluate?	5±2	4±2
Are the instructions adequate?	4±2	5±2
Is a subjective decision needed for this criterion?	4±2	4±2
Are responses comprehensive?	**7**±1	**7**±2
Should this item be weighted to detect major flaws in quality?	2±1	2±2
Is number of response options adequate?	**7**±1	**7**±2
Discrimination power for reporting quality vs. study quality	4±1	4±1
Are assessments reproducible?	**7**±1	**7**±1
Overall Estimation of Validity*		
Wide applicability for different areas of research	**6**±1	**6**±0
Can be useful for various groups (clinicians, public health professionals, consumers, policymakers)	**7**±2	**9**±0
Overall estimation is clear and simple to understand	**6**±2	**6**±3
Program conclusion based on detected major flaws is adequate (can be specified depending on the area of research)	**6**±2	5±2
Is necessary information usually available to estimate validity?	5±2	4±2
Is subjective decision needed to estimate validity?	5±2	**6**±2

Table 2. Credibility of the checklists to assess quality of observational studies (evaluation by participating experts) (continued)

Quality Criteria	Incidence Mean ± STD	Risk Factors Mean ± STD
Likelihood of bias in applicability of the tool	5±2	**6**±1
Is single domain of validity comprehensive?	5±2	5±0
Are redundant items present in the tool?	**6**±2	3±0
Are item weights adequate to detect major flaws in validity?	**6**±2	5±0
Are the numbers of conclusions optional?	5±2	5±0
Does the tool have discrimination power for levels of validity?	5±2	5±0
Does the tool have discrimination power for reporting vs. study quality?	5±1	**6**±1

*good credibility when score mean was ≥6

Pilot Testing of Reliability of Modified Checklists. The experts were asked to read the articles and complete appraisals for the four studies of incidence or prevalence and six studies of risk factors of chronic conditions. Seven experts completed the evaluations.

In order to identify potential actions to improve reliability, we compared raw agreement, Fleiss' and generalized kappa, and AC1 statistics to detect levels of agreement for each quality component, and evaluated possible reasons for disagreement. The estimation of reliability differed across three statistics. Overall, Fleiss' kappa was nonsignificant, ranging from 0 to 85 percent (Table 3). We detected a paradox in kappa values when high agreement was joined with low kappa values.[115] For example, nongeneral population based sampling methods by self-selection were marked with a high level of agreement; the statistical tests demonstrated a negative nonsignificant kappa (-0.04±0.95) but a significant AC1 statistic (0.92±0.08) (Appendix B, Tables 1-2). Standard errors of calculated generalized kappa and Fleiss' kappa were generally largely attributed to the small number of articles. Kappa statistics can be less reliable for categorical, not mutually exclusive responses, and more than one answer was allowed, including an open-ended response for each quality component.

Table 3. Agreement summaries

Incidence or Prevalence Studies			Risk Factors Studies		
Agreement	Fleiss Kappa (95%CI)	Generalized Kappa	Agreement	Fleiss Kappa (95%CI)	Weighted Kappa
Funding					
0.9	0.79 (-0.08; 1)	1	0.5	0.22 (-0.06; 0.39)	0.288
Role of the funding organization in data analysis and interpretations of the results					
0.8	-0.07 (-0.19; 0.01)	0.02	0.8	-0.098 (-0.168; -0.039)	
Conflict of interest					
0.7	0.6 (-0.13; 0.79)	0.64	0.8	-0.082 (-0.139; -0.039)	-0.037
Country					
0.9	0.86 (0.36; 1)		0.9	0.853 (0.553; 1)	
Ethics approval					
0.9	0.83 (-0.08; 1)	0.83	0.6	0.376 (-0.145; 0.616)	0.425
Aim of the study					
0.3	0.03 (-0.15; 0.05)	0.21	0.3	0.105 (-0.095; 0.263)	0.155
Study design					
0.9	0.83 (-.12; 1)	1	0.4	0.302 (-.011; 0.595)	0.34
Objectives					
			0.4	0.005 (-0.125; 0.135)	0.025
		External Validity			
General population based sampling					
0.1	-0.04 (-0.4; -0.03)	0.1	0.2	0.058 (-0.092; 0.105)	0.042
Nongeneral population based sampling method					
0.2	0 (-0.21; 0.04)	0	0.1	-0.156 (-0.299; -0.103)	-0.085
Nongeneral population based sampling frame					
0.8	0.81 (0.2;1)	0.74	0.4	0.158 (-0.131; 0.385)	0.215

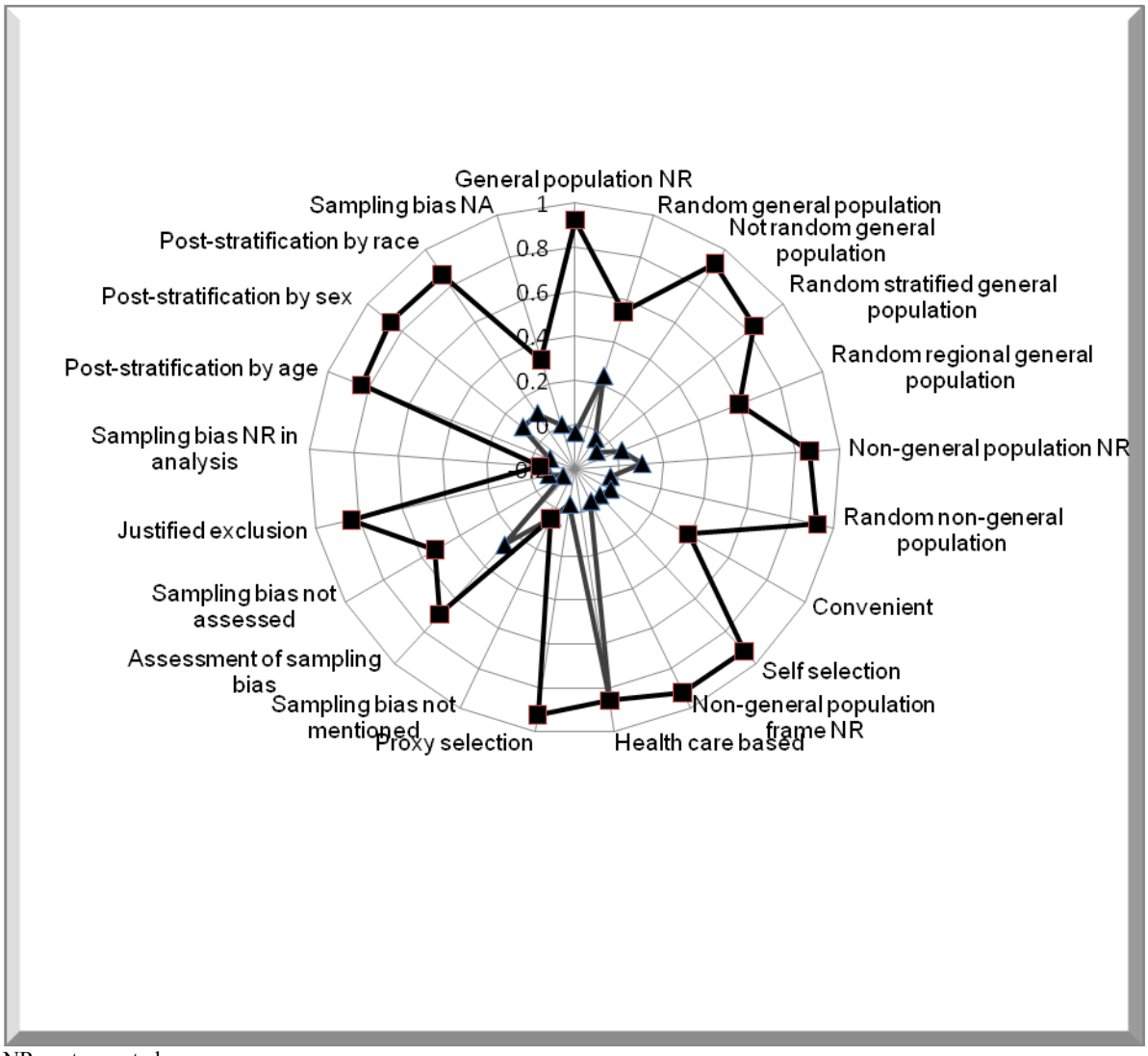
Table 3. Agreement summaries (continued)

Incidence or Prevalence Studies			Risk Factors Studies		
Agreement	Fleiss Kappa (95%CI)	Generalized Kappa	Agreement	Fleiss Kappa (95%CI)	Weighted Kappa
Assessment of sampling bias					
0.2	-0.01 (-0.14; 0.07)	0.04	0.2	0.021 (-0.106; 0.114)	-0.021
Sampling bias is addressed in the analysis					
0.3	0 (-0.14; 0.09)	-0.02	0.3	0.005 (-0.139; 0.091)	0.06
Exclusion rate from the analysis					
0.6	0.34 (-0.18; 0.69)		0.4	0.054 (-0.135; 0.161)	
Exclusion rate from the analysis in exposed and not exposed					
			0.3	0.008 (-0.168; 0.202)	0.074
For case control studies					
			0.6	0.182 (-0.2; 0.403)	0.114
Subject flow					
0.3	0.08 (-0.22; 0.14)		0.2	-0.045 (-0.266; -0.004)	
		Internal Validity			
Source to measure dependent variables (target, outcomes)					
0.2	-0.03 (-0.2; -0.02)	0.27	0.1	0.051 (-0.068; 0.128)	0.042
Definition of the dependent variable-reference period					
0.2	-0.01 (-0.14; 0)	-0.02	0.2	-0.01 (-0.101; 0.006)	-0.004
Severity					
0.4	0.08 (-0.14; 0.18)	0.21	0.4	0.074 (-0.106; 0.202)	0.077
Frequency of the symptoms					
0.4	0.15 (-0.09; 0.22)	0.14	0.5	0.237 (0.005; 0.357)	0.224
Measurements of the incidence/prevalence of chronic disease(s)					
0.3	0.12 (-0.12; 0.22)	0.12	0.3	0.024 (-0.105; 0.118)	0.06
Reliability of the estimates					
0.5	0.18 (-0.15; 0.4)	0.26	0.3	-0.028 (-0.134; 0.03)	0.058
Confounding factors, or the factors that can modify the association between risk factor and disease					
			0.3	-0.099 (-0.168; -0.079)	-0.072
Measure of confounding factors					
			0.3	0.018 (-0.104; 0.103)	0.058
Loss of followup					
			0.6	0.48 (0.099; 0.802)	
Masking of exposure status for investigators who measured dependent variables					
			0.4	0.018 (-0.09; 0.055)	0.07
Statistical analysis					
			0.5	0.02 (-0.097; 0.124)	0.076
Appropriateness of statistical model to reduce research specific bias					
			0.5	0.072 (-0.094; 0.201)	0.039
Sample size justification					
			0.5	-0.092 (-0.154; -0.073)	-0.067

Checklists of studies of incidence or prevalence of chronic conditions. Generalized kappa for each article demonstrated fair agreement for two studies of incontinence (generalized kappa 0.38; 95% CI 0.32; 0.43;[96] 0.21; 95% CI 0.12; 0.30[98]) and two studies of depression (generalized kappa 0.34; 95% CI 0.26; 0.42;[97] 0.28; 95% CI 0.22; 0.34[99]). The differences in agreement for each quality criteria in generalized kappa and AC1 statistics are presented in Figure 3, which displays means of both statistics relative to a scale from 0 (center point) to 1 (perimeter area, perfect agreement). Subject flow with calculated eligibility, enrollment, and recruitment fractions had intra-class correlations of 0.86, 0.97, and 1.00, respectively. Because the studies differed by order of magnitude, we also computed the intra-class correlations on the logarithm of the responses; these were 0.94, 0.99, and 1.00, respectively.

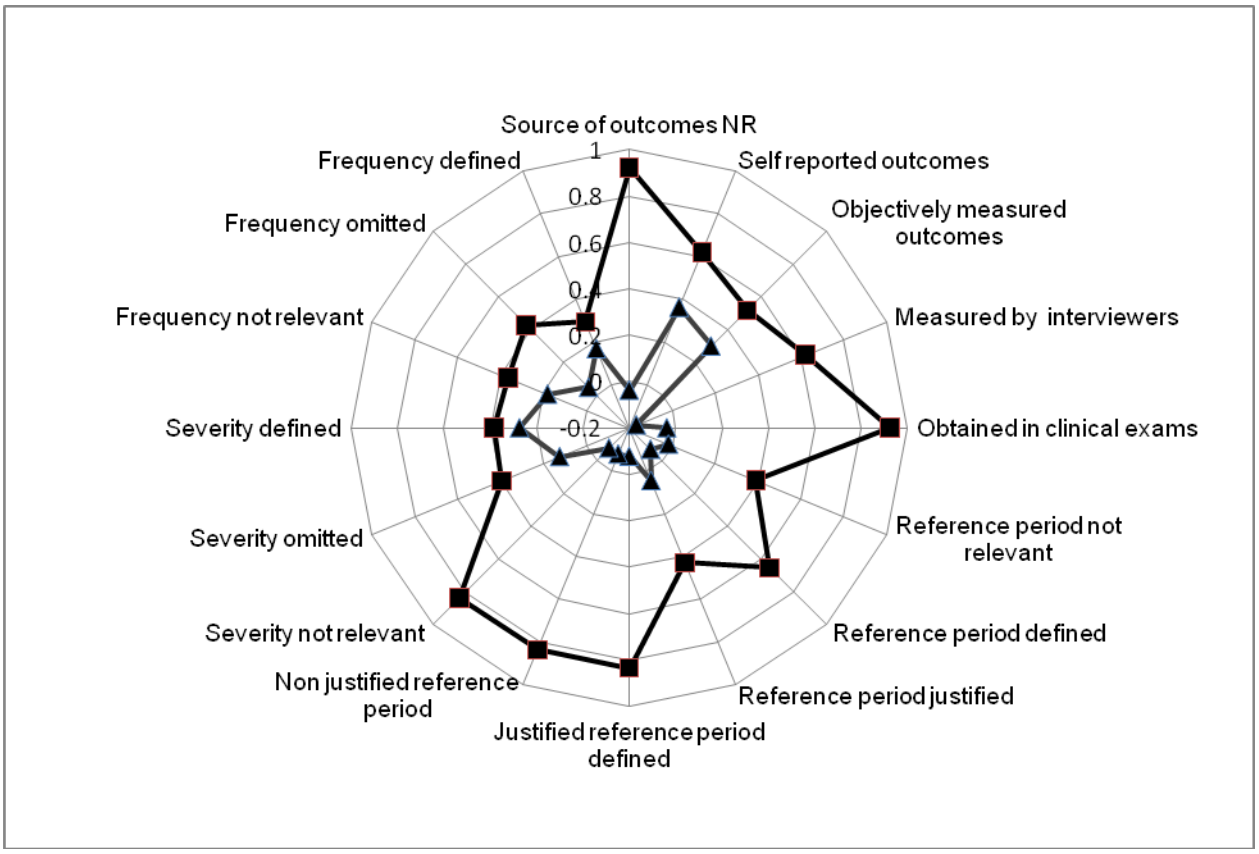
Figure 3. General kappa (triangle symbols) and AC1 statistics in observational nontherapeutic studies of population incidence or prevalence of chronic diseases (based on pilot reliability testing of four articles by seven expert groups)

A. Sampling and assessment of sampling bias



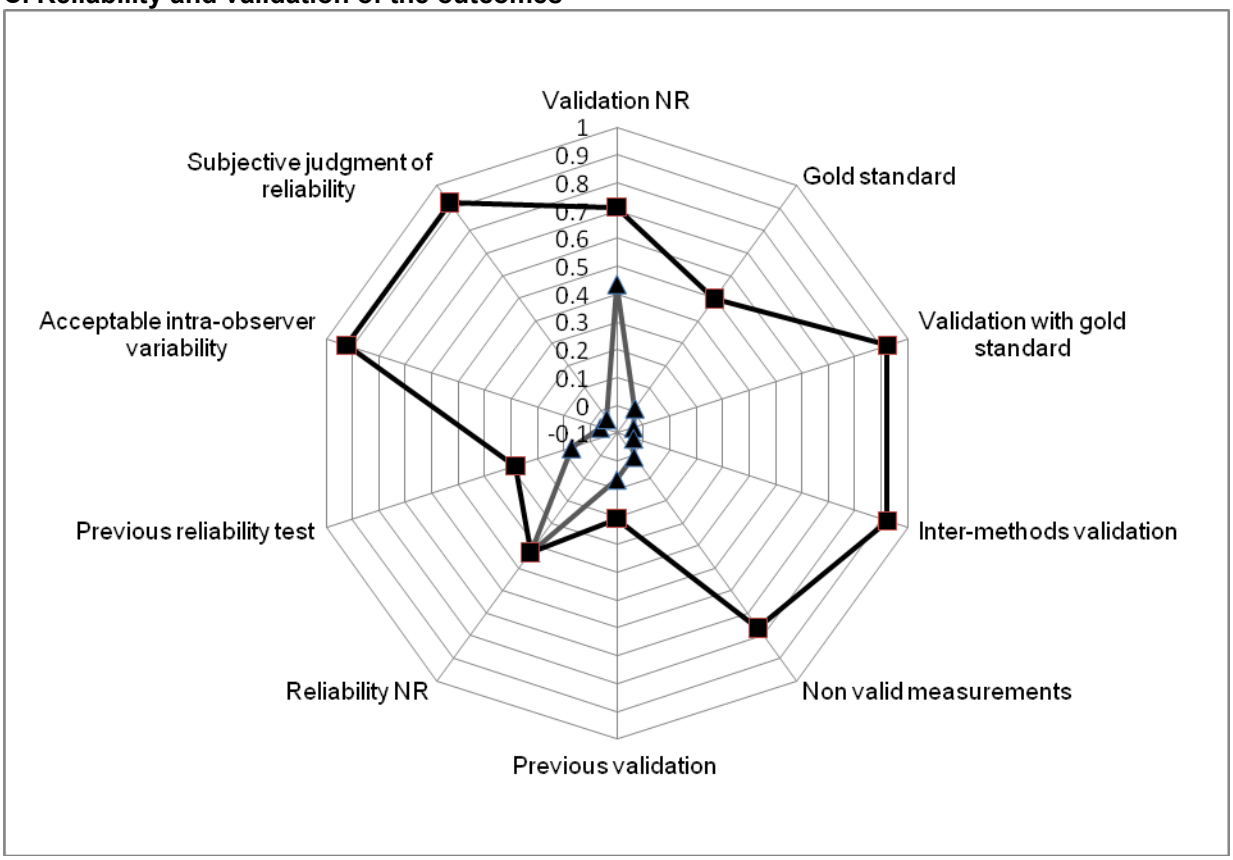
NR=not reported

B. Measurement and definitions of the outcomes



NR=not reported

C. Reliability and validation of the outcomes



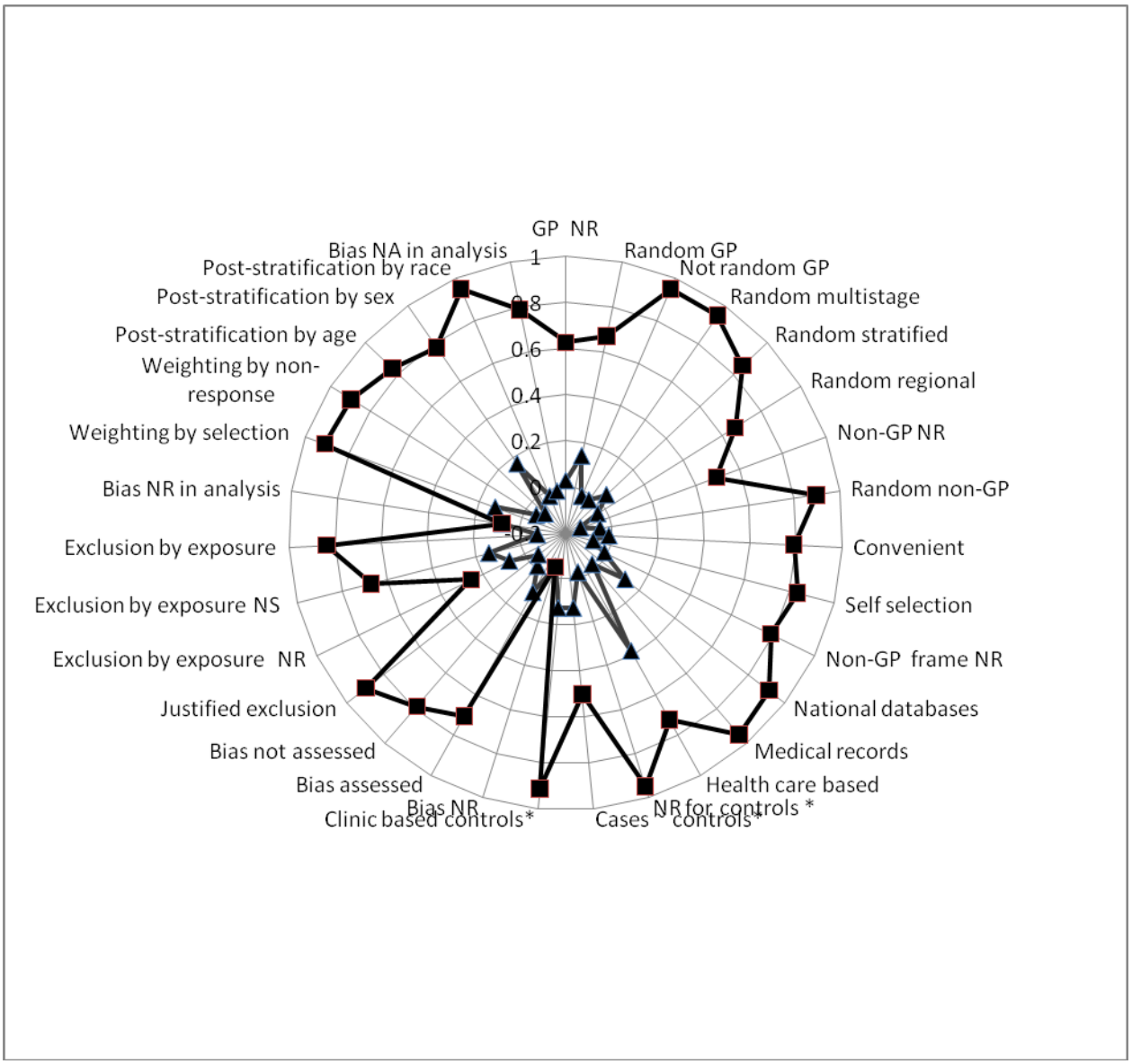
NR=not reported

Radar plot displays mean of generalized kappa and AC1 statsitics relaitve to a scale from 0 (center point) to 1 (perimeter area, perfect agreement), standard errors and significance are presented in Appendix B, Table 1

Checklists of studies of risk factors for chronic conditions. Generalized kappa for each article demonstrated fair agreement for studies of genetic risk factors (0.23, 95% CI 0.16; 0.30),[102] insomnia (0.27, 95% CI 0.21; 0.32),[101] and coronary heart diseases among women (0.26, 95% CI 0.18; 0.34).[105].Agreement was only by chance for articles of risk factors of depression[100] and violence among adolescents.[104] The differences in agreement for each quality criterion in generalized kappa and AC1 statistic are shown in Figure 4. Subject flow with calculated eligibility, enrollment, and recruitment fractions had the intra-class correlations 0.98, 0.76, and 0.73. Because the studies differed by order of magnitude, we also computed the intra-class correlations on the logarithm of the responses; these were 0.99, 0.91, and 0.70, respectively**.**

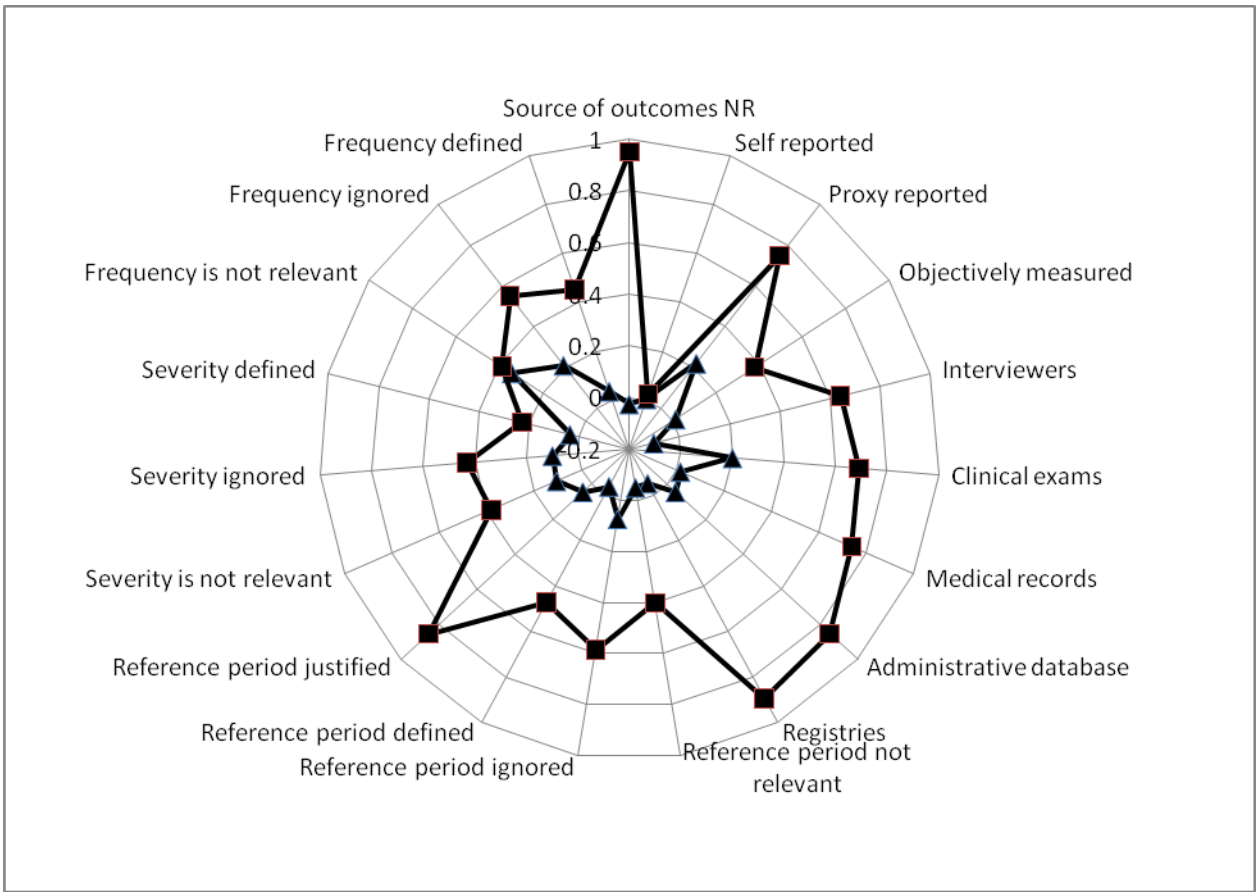
Figure 4. General kappa (triangle symbols) and AC1 statistics in observational nontherapeutic studies of risk factors of chronic diseases (based on pilot reliability testing of six articles by seven expert groups)

A. Sampling and assessment of sampling bias



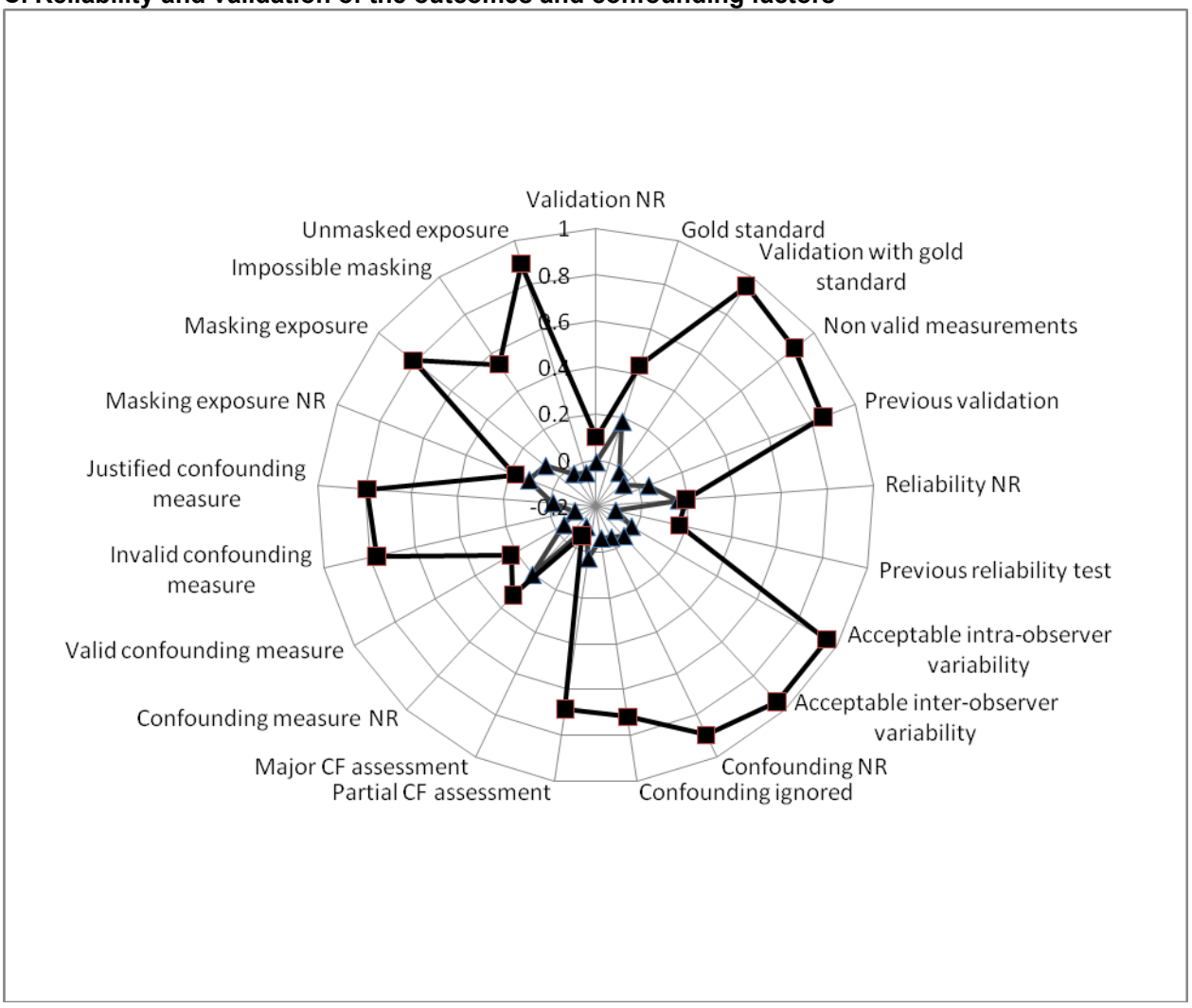
* specific for case control studies
NR=not reported
GP=general population
NA=not addressed

B. Sources to measure and definitions of the outcomes



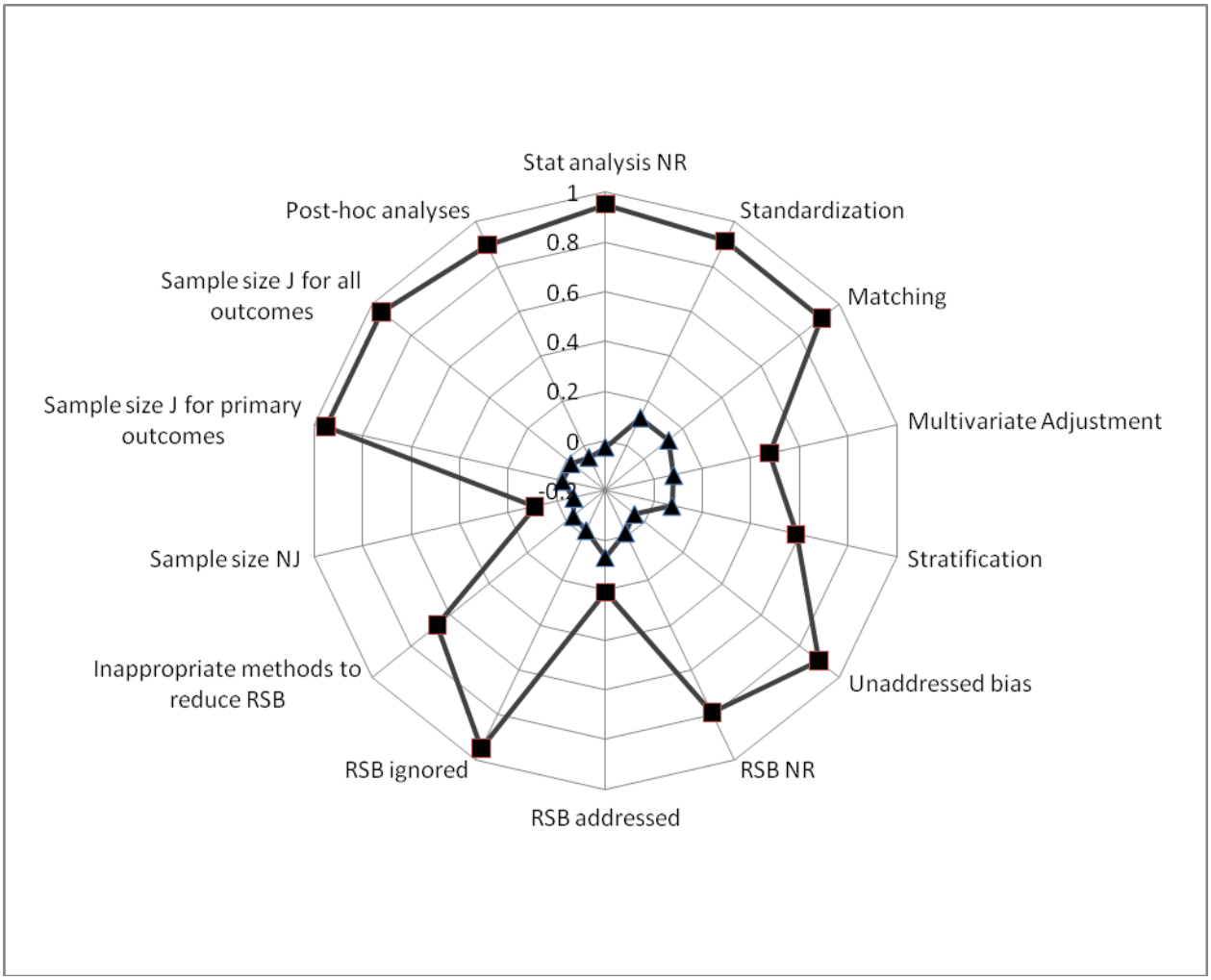
NR=not reported

C. Reliability and validation of the outcomes and confounding factors



NR=not reported
CF=confounding factors

D. Strategies to reduce bias and type II error



NR=not reported
J=justified
NJ=not justified
RSB=research specific bias
Radar plot displays mean of generalized kappa and AC1 statsitics relaitve to a scale from 0 (center point) to 1 (perimeter area, perfect agreement), standard errors and significance are presented in Appendix B, Table 2

Validation of the checklists. We examined discriminant validity by testing the hypothesis that our checklists can discriminate quality across studies and discriminate reporting vs. methodological quality. Mean scores differed for the articles with different numbers of methodological flaws (Table 4). We concluded that our checklists discriminated the quality of studies. Total scores for poor reporting were significantly greater for the majority of the articles when compared to the scores for methodological flaws (Table 4). We concluded that the checklists discriminated poor reporting from low methodological quality in the studies based on the presence of major flaws (Figure 5) and minor flaws (Figure 6).

Table 4. Discriminant validity of the checklists to detect differences in reporting quality and in methodological flaws of the studies

Study	Reporting Quality		Minor Flaws		Major Flaws	
	Mean	95% CI	Mean	95% CI	Mean	95% CI
Studies of incidence or prevalence of chronic conditions						
Luutonen, 2002[97]	0.19*‡	0.10; 0.29	0.09*†‡**	0.03; 0.15	0.17	0.06; 0.27
Lauzon, 2003[99]	0.25*	0.15; 0.34	0.08*†‡**	0.02; 0.14	0.17	0.06; 0.27
Hunskaar, 2004[98]	0.18*‡	0.09; 0.28	0.11	0.05 ;0.17	0.12	0.01; 0.23
Waetjen, 2007[96]	0.29	0.19; 0.38	0.13	0.07; 0.19	0.12*	0.01; 0.23
Studies of risk factors of chronic conditions						
Lesperance, 2002[100]	0.18*†	0.10; 0.27	0.07†	0.03; 0.11	0.06	0.01; 0.11
Leppavuori, 2002[101]	0.36†‡	0.28; 0.44	0.13*‡	0.08; 0.17	0.13*	0.08; 0.18
Nuotio, 2003[103]	0.21*‡	0.13; 0.29	0.09	0.05; 0.13	0.08	0.03; 0.14
Crew, 2007[102]	0.29	0.21; 0.37	0.08‡	0.04; 0.12	0.02	-0.03; 0.08
Vilbergsson, 1998[105]	0.39*	0.31; 0.47	0.11*	0.07; 0.15	0.04	-0.02; 0.09
Dishion, 1997[104]	0.28	0.20; 0.36	0.05**	0.01; 0.09	0.05*	0.00 0.10

*,†,‡,** - significant differences in reporting or methodological quality

Figure 5. Discrimination of methodological vs. reporting quality of the studies by the checklists: differences in major flaw and poor reporting

Checklists discriminated reporting vs. methodological qualities of the examined studies when 95% confidence interval (CI) of mean difference in the total score between reporting and methodological qualities does not include 0.

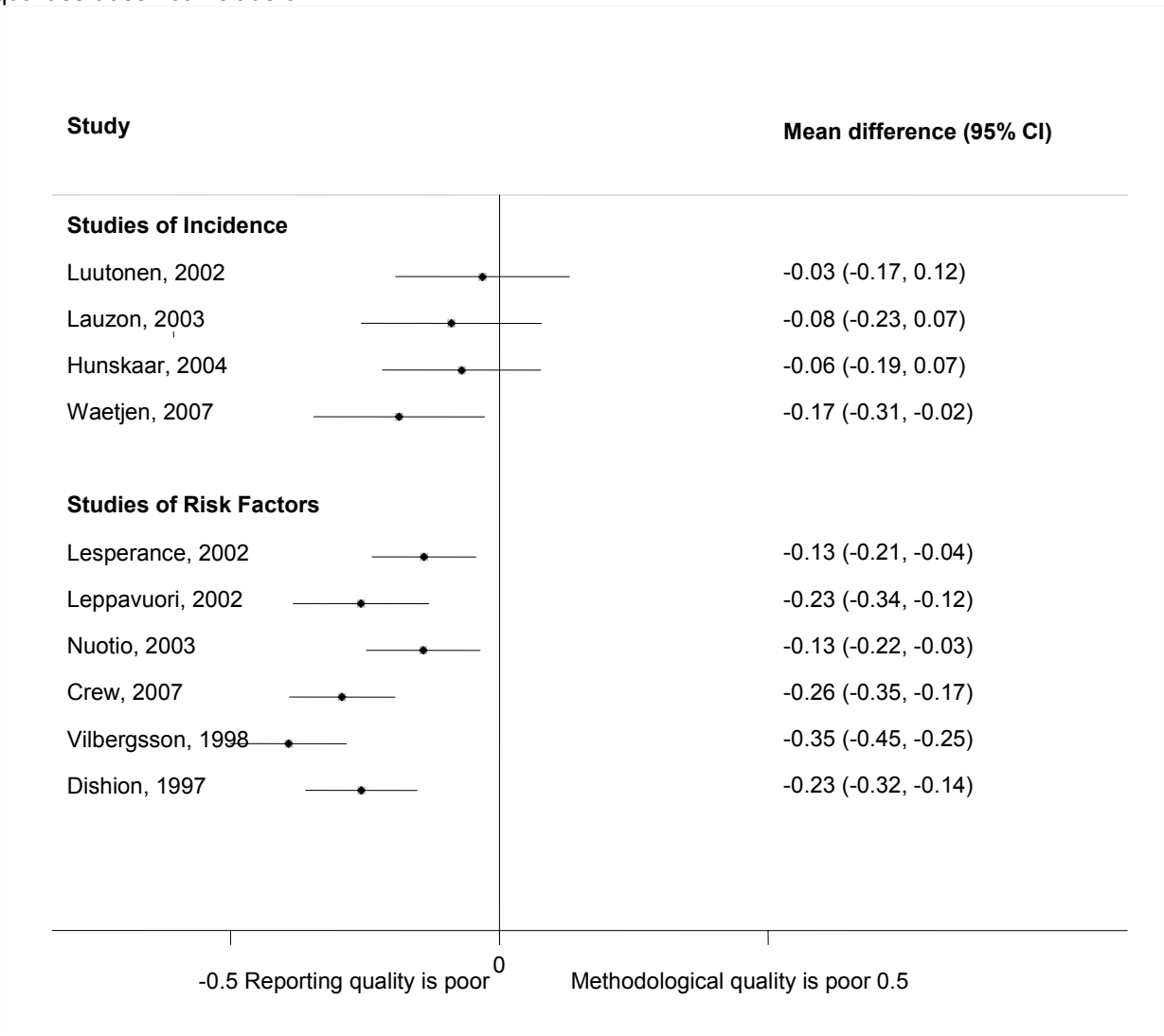
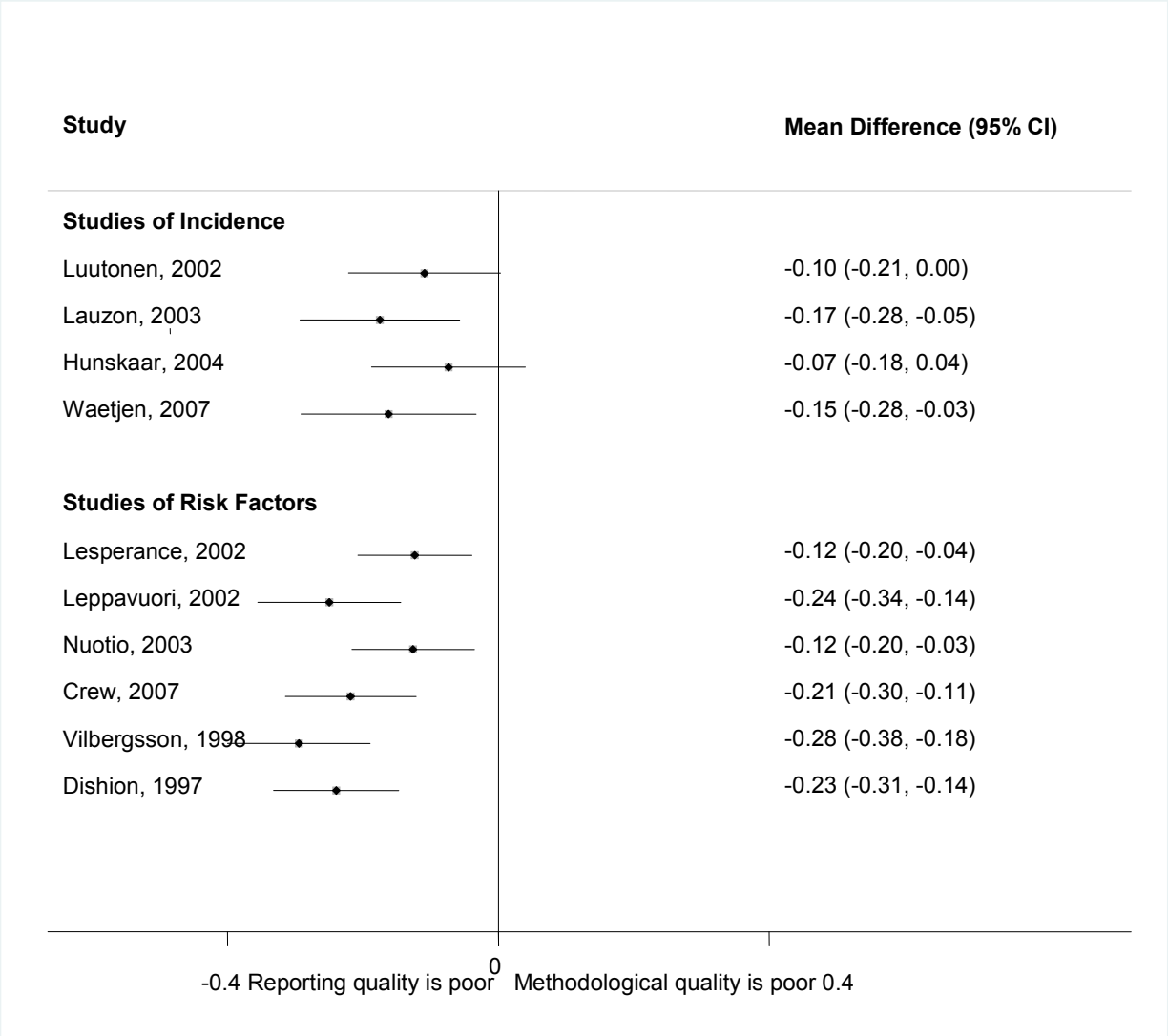


Figure 6. Discrimination of methodological vs. reporting quality of the studies by the checklists: differences in minor flaw and poor reporting

Checklists discriminated reporting vs. methodological qualities of the examined studies when 95% confidence interval of mean difference in the total score between reporting and methodological qualities does not include 0.



We concluded that the checklists have reasonable validity but poor overall agreement according to kappa statistics. We also concluded that the structure and allowing more than one response resulted in lower than expected kappa. We detected areas of major disagreement that can be improved by a priori consensus around appropriate definitions of the target population, population subgroups, risk factors, and the reference methods of the measurements.

Finalizing the Checklists

Modifications based on the tests for validity and reliability. We designed a template for the preplanned protocols of quality evaluation that should include justified consensus about target

populations, availability of a gold standard to measure outcomes, acceptable definitions of exposure and outcomes, and the most appropriate strategies to reduce bias (Appendix A).

The quality assessment for each study includes separate evaluations of external and internal validity, lists major and minor flaws; and poorly reported details about the methodology of the study. The assessments are available in both text format (Access reports) and Excel spreadsheets that can be analyzed by statistical software to incorporate quality criteria to quantitative analysis of evidence.

Interpretation of quality assessments. The finalized checklists were used to create an Access database which produces standardized quality reports (Appendix A). Each report has separate evaluations of external and internal validity. Each evaluation lists poorly reported quality criteria and methodological flaws in the study. Quality reports in spreadsheet format should be used to synthesize evidence including quantitative meta-analyses and meta-regression. The extent to which the quality of individual studies can explain heterogeneity in the results would depend on the specific area of the research.

Discussion

We present the results of our pilot collaborative project to develop checklists for quality assessment of observational nontherapeutic studies. Comparing our checklists with previously published scales and checklists, we could find only one that was developed for studies of incidence or prevalence of health conditions.[50] The scale has three domains, including the validity of study methods and interpretation and applicability of the results. A maximum score of 8 does not differentiate domains; for example, the same score of 7 was given to studies with biased sampling and to studies with low diagnostic values of the tests to measure the outcome.[50] The appropriate study design or sampling method would receive the same score of 1.[34,50] Quality assessments based on detected flaws in external and internal validity rather than an arithmetic score can result in standardized and transparent appraisals of studies. Our checklists address and discriminate both reporting and methodological quality, which is more efficient than performing separate assessments.

We believe that our checklists can provide comprehensive quality assessments of incidence and prevalence or risk factors studies of any chronic disease. Growing numbers of publications address research specific reporting and methodological standards,[120] including appropriate adjustment for confounding or residual confounding.[121] For example, recently published reporting standards for studies of genetic associations[122] emphasized the role of population stratification bias, Hardy-Weinberg equilibrium, or genotyping errors as well as treatment effects in quantitative traits. We incorporated research specific quality standards in the checklists, including one question that directly addresses research specific appropriate methods to reduce bias that should be predefined in the protocols of quality evaluations.

Several experts noted that completing the quality assessment was time consuming. Poor quality studies with major flaws required more time than those that were well designed. Investigators of evidence-based reports must evaluate the quality of all eligible studies, and flawed studies are very common in biomedical research. To address this issue, we developed three levels of quality assessment by the presence of major methodological flaws or poor reporting of the most important quality criteria (Appendix A).We modified both checklists for different audiences, depending on the goals of quality appraisals: retrospective quality assessment of individual studies for systematic reviews, and quality assessment of submitted manuscripts.

Limitations

In the absence of a gold standard, a formal test for criterion validity was not feasible. Previously published appraisals tested the ability of the checklists or scales to discriminate poor quality from well designed studies.[74]The investigators of such validity tests judged the quality of studies before actual testing. In contrast, we did not select studies by internal or external validity to examine discriminant validity of the checklists when compared to well-designed population-based surveys or cohort studies. We tested the hypothesis that our checklists can detect differences in reporting and methodological quality among the studies that are typical in EPC reports.

We included a very small number of studies in the pilot reliability test. Pilot reliability tests demonstrated agreement by chance for most quality criteria and stronger agreement about different quality components. Allowing more than one response may reduce statistical estimates

of agreement. Three reliability estimates (Fleiss' kappa, generalized kappa, and AC1 statistics) showed different direction, magnitude, and significance of agreement. We selected studies that evaluated different chronic diseases; there may have been better agreement if the articles examined incidence and risk factors for the same disease. A previously published reliability test for a measurement tool for AMSTAR involved 99 paper-based and 52 electronic systematic reviews.[123] A pilot test of the GRADE system was conducted evaluating 12 examples of systematic reviews and meta-analyses.[4] The authors of the previously developed and validated appraisals did not report how they achieved good reliability, other than via prior consensus about research specific quality criteria and training of the examiners.[41,49-50,74,90] In our pilot we detected major areas of disagreement, including adequacy of sampling, use of gold standards to measure outcomes, and validity of subgroup analyses; we propose to pre-specify consensus around those criteria in the protocols of quality evaluation.

We did not test reliability in the same situation as a review team would; namely, a dual independent review with a consensus resolution. Poor reliability evaluating studies from different research areas precludes our recommendations for widespread use of our checklists. Fully powered reliability testing should be conducted by the authors of systematic reviews with a consensus resolution around research specific quality standards.

Recommendations for Future Research

Future research should involve a broader audience of methodologists and other stakeholders to evaluate proposed quality criteria. Our preliminary testing should continue and result in the development of reliable tools that can be used with confidence.

For future research we propose using our checklists to assess the quality of studies of incidence and prevalence of chronic diseases (MORE, Appendix A) or risk factors (MEVORECH, Appendix A) in systematic reviews. Protocols of systematic reviews of nontherapeutic observational studies should include justified definitions of research specific quality components and methodological flaws, and preplanned reliability testing of appraisals. Systematic reviews should incorporate quality into the synthesis of evidence to estimate how it affects the results of primary studies and conclusions of the review. The evaluation of the level of evidence from several observational nontherapeutic studies was beyond our present goals and should be studied in future research.

Journal policies, including implementation of quality assessments of manuscripts, play a significant role in improving completeness and transparency of reports.[124] Journal editors should make publication decisions and revisions based on standardized valid quality evaluations of submitted manuscripts. The peer review process of grant proposals does not include standard, transparent, and valid prospective estimation of the expected quality of studies.[125-126] No evidence about the effects of peer review on the quality of funded research is presently available.[127] Future research using the newly developed checklists should evaluate the effects of prospective quality assessment on the validity and applicability of funded research. Quality evaluations should be conducted in all stages of biomedical research, from funding to publication, and to the decisionmaking process. Journal editors should require clarifications of poorly reported quality criteria, can require additional justifications that flaws could not be avoided or bias cannot be reduced, or can reject the manuscript based on quality evaluation. Reviewers of grant proposals should reject both poorly written applications and proposals of studies with expected major or minor flaws in external and internal validity.

References

1. Fox DM. Evidence of evidence-based health policy: the politics of systematic reviews in coverage decisions. Health Aff (Millwood) 2005 Jan-Feb;24(1):114-122.
2. Lavis J, Davies H, Oxman A, et al. Towards systematic reviews that inform health care management and policy-making. J Health Serv Res Policy 2005 Jul;10(Suppl 1):35-48.
3. Atkins D, Eccles M, Flottorp S, et al. Systems for grading the quality of evidence and the strength of recommendations I: critical appraisal of existing approaches The GRADE Working Group. BMC Health Serv Res 2004 Dec 22;4(1):38.
4. Atkins D, Briss PA, Eccles M, et al. Systems for grading the quality of evidence and the strength of recommendations II: pilot study of a new system. BMC Health Serv Res 2005 Mar 23;5(1):25.
5. Aschengrau A, Seage GR. Essentials of Epidemiology in Public Health. Sudbury, Mass: Jones and Bartlett; 2003.
6. Bero LA, Jadad AR. How consumers and policymakers can use systematic reviews for decision making. Ann Intern Med 1997 Jul 1;127(1):37-42.
7. Chan KS, Morton SC, Shekelle PG. Systematic reviews for evidence-based management: how to find them and what to do with them. Am J Manag Care 2004 Nov;10(11 Pt 1):806-812.
8. Briss PA, Zaza S, Pappaioanou M, et al. Developing an evidence-based Guide to Community Preventive Services—methods. The Task Force on Community Preventive Services. Am J Prev Med 2000 Jan;18(1 Suppl):35-43.
9. Agency for Healthcare Research and Quality. University of California-San Francisco-Stanford Evidence-Based Practice Center. Systems to Rate the Strength of Scientific Evidence: Rockville, MD; 2002.
10. Shamliyan T, Minnesota Evidence-based Practice Center, United States. Agency for Healthcare Research and Quality. Prevention of urinary and fecal incontinence in adults. Agency for Healthcare Research and Quality. Available at: http://purl.access.gpo.gov/GPO/LPS88803.
11. Feinstein AR. Clinimetrics. New Haven: Yale University Press; 1987.
12. Horwitz RI, Feinstein AR. Methodologic standards and contradictory results in case-control research. Am J Med 1979 Apr;66(4):556-564.
13. How to read clinical journals: IV. To determine etiology or causation. Can Med Assoc J 1981 Apr 15;124(8):985-990.
14. Krogh CL. A checklist system for critical review of medical literature. Med Educ 1985 Sep;19(5):392-395.
15. Gardner MJ, Machin D, Campbell MJ. Use of check lists in assessing the statistical content of medical studies. Br Med J (Clin Res Ed) 1986 Mar 22;292(6523):810-812.
16. Mulrow CD, Lichtenstein MJ. Blood glucose and diabetic retinopathy: a critical appraisal of new evidence. J Gen Intern Med 1986 Mar-Apr;1(2):73-77.
17. Esdaile JM, Horwitz RI. Observational studies of cause-effect relationships: an analysis of methodologic problems as illustrated by the conflicting data for the role of oral contraceptives in the etiology of rheumatoid arthritis. J Chronic Dis 1986;39(10):841-852.
18. Lichtenstein MJ, Mulrow CD, Elwood PC. Guidelines for reading case-control studies. J Chronic Dis 1987;40(9):893-903.
19. Longnecker MP, Berlin JA, Orza MJ, et al. A meta-analysis of alcohol consumption in relation to risk of breast cancer. JAMA 1988 Aug 5;260(5):652-656.
20. Zola P, Volpe T, Castelli G, et al. Is the published literature a reliable guide for deciding between alternative treatments for patients with early cervical cancer? Int J Radiat Oncol Biol Phys 1989 Mar;16(3):785-797.
21. Reisch JS, Tyson JE, Mize SG. Aid to the evaluation of therapeutic studies. Pediatrics 1989 Nov;84(5):815-827.
22. Spitzer WO, Lawrence V, Dales R, et al. Links between passive smoking and disease: a best-evidence synthesis. A report of the Working Group on Passive Smoking. Clin Invest Med 1990 Feb;13(1):17-42; discussion 3-6.

23. Berlin JA, Colditz GA. A meta-analysis of physical activity in the prevention of coronary heart disease. Am J Epidemiol 1990 Oct;132(4):612-628.

24. Stock SR. Workplace ergonomic factors and the development of musculoskeletal disorders of the neck and upper limbs: a meta-analysis. Am J Ind Med 1991;19(1):87-107.

25. Oxman AD, Guyatt GH. Validation of an index of the quality of review articles. J Clin Epidemiol 1991;44(11):1271-1278.

26. Fowkes FG, Fulton PM. Critical appraisal of published research: introductory guidelines. Bmj 1991 May 11;302(6785):1136-1140.

27. Carruthers SG, Larochelle P, Haynes RB, et al. Report of the Canadian Hypertension Society Consensus Conference: 1. Introduction. CMAJ 1993 Aug 1;149(3):289-293.

28. Carson CA, Fine MJ, Smith MA, et al. Quality of published reports of the prognosis of community-acquired pneumonia. J Gen Intern Med 1994 Jan;9(1):13-19.

29. Avis M. Reading research critically. II. An introduction to appraisal: assessing the evidence. J Clin Nurs 1994 Sep;3(5):271-277.

30. Gyorkos TW, Tannenbaum TN, Abrahamowicz M, et al. An approach to the development of practice guidelines for community health interventions. Can J Public Health 1994 Jul-Aug; 85(Suppl 1):S8-S13.

31. Cho MK, Bero LA. Instruments for assessing the quality of drug studies published in the medical literature. JAMA 1994 Jul 13;272(2):101-104.

32. Levine M, Walter S, Lee H, et al. Users' guides to the medical literature. IV. How to use an article about harm. Evidence-Based Medicine Working Group. JAMA 1994 May 25;271(20):1615-1619.

33. Goodman SN, Berlin J, Fletcher SW, et al. Manuscript quality before and after peer review and editing at Annals of Internal Medicine. Ann Intern Med 1994 Jul 1;121(1):11-21.

34. DuRant RH. Checklist for the evaluation of research articles. J Adolesc Health 1994 Jan;15(1):4-8.

35. Campos-Outcalt D, Senf J, Watkins AJ, et al. The effects of medical school curricula, faculty role models, and biomedical research support on choice of generalist physician careers: a review and quality assessment of the literature. Acad Med 1995 Jul;70(7):611-619.

36. Margetts BM, Thompson RL, Key T, et al. Development of a scoring system to judge the scientific quality of information from case-control and cohort studies of nutrition and disease. Nutr Cancer 1995;24(3):231-239.

37. Cowley DE. Prostheses for primary total hip replacement. A critical appraisal of the literature. Int J Technol Assess Health Care 1995 Fall;11(4):770-778.

38. Garber BG, Hebert PC, Yelle JD, et al. Adult respiratory distress syndrome: a systemic overview of incidence and risk factors. Crit Care Med 1996 Apr;24(4):687-695.

39. Anders JF, Jacobson RM, Poland GA, et al. Secondary failure rates of measles vaccines: a metaanalysis of published studies. Pediatr Infect Dis J 1996 Jan;15(1):62-66.

40. Hadorn DC, Baker D, Hodges JS, et al. Rating the quality of evidence for clinical practice guidelines. J Clin Epidemiol 1996 Jul;49(7):749-754.

41. Jabbour M, Osmond MH, Klassen TP. Life support courses: are they effective? Ann Emerg Med 1996 Dec;28(6):690-698.

42. Ciliska D, Hayward S, Thomas H, et al. A systematic overview of the effectiveness of home visiting as a delivery strategy for public health nursing interventions. Can J Public Health 1996 May-Jun;87(3):193-198.

43. Solomon DH, Bates DW, Panush RS, et al. Costs, outcomes, and patient satisfaction by provider type for patients with rheumatic and musculoskeletal conditions: a critical review of the literature and proposed methodologic standards. Ann Intern Med 1997 Jul 1;127(1):52-60.

44. Littenberg B, Weinstein LP, McCarren M, et al. Closed fractures of the tibial shaft. A meta-analysis of three methods of treatment. J Bone Joint Surg Am 1998 Feb;80(2):174-183.

45. Spencer-Green G. Outcomes in primary Raynaud phenomenon: a meta-analysis of the frequency, rates, and predictors of transition to secondary diseases. Arch Intern Med 1998 Mar 23;158(6):595-600.

46. Kreulen CM, Creugers NH, Meijering AC. Meta-analysis of anterior veneer restorations in clinical studies. J Dent 1998 May;26(4):345-353.

47. Jadad AR, Moher D, Klassen TP. Guides for reading and interpreting systematic reviews: II. How did the authors find the studies and assess their quality? Arch Pediatr Adolesc Med 1998 Aug;152(8):812-817.

48. Borghouts JA, Koes BW, Bouter LM. The clinical course and prognostic factors of non-specific neck pain: a systematic review. Pain 1998 Jul;77(1):1-13.

49. Downs SH, Black N. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. J Epidemiol Community Health 1998 Jun;52(6):377-384.

50. Loney PL, Chambers LW, Bennett KJ, et al. Critical appraisal of the health research literature: prevalence or incidence of a health problem. Chronic Dis Can 1998;19(4):170-176.

51. Silman A, Symmons D. Reporting requirements for longitudinal observational studies in rheumatology. J Rheumatol 1999 Feb;26(2):481-483.

52. van Rooyen S, Black N, Godlee F. Development of the review quality instrument (RQI) for assessing peer reviews of manuscripts. J Clin Epidemiol 1999 Jul; 52(7):625-629.

53. Angelillo IF, Villari P. Residential exposure to electromagnetic fields and childhood leukaemia: a meta-analysis. Bull World Health Organ 1999; 77(11):906-915.

54. Corrao G, Bagnardi V, Zambon A, et al. Exploring the dose-response relationship between alcohol consumption and the risk of several alcohol-related conditions: a meta-analysis. Addiction 1999 Oct;94(10):1551-1573.

55. Cullum N. Critical appraisal. Finding and appraising cohort studies for causation and prognosis. NT Learn Curve 1999 Sep 1;3(7):8-10.

56. Nguyen QV, Bezemer PD, Habets L, et al. A systematic review of the relationship between overjet size and traumatic dental injuries. Eur J Orthod 1999 Oct;21(5):503-515.

57. Cameron I, Crotty M, Currie C, et al. Geriatric rehabilitation following fractures in older people: a systematic review. Health Technol Assess 2000;4(2):i-iv, 1-111.

58. Ariens GA, van Mechelen W, Bongers PM, et al. Physical risk factors for neck pain. Scand J Work Environ Health 2000 Feb;26(1):7-19.

59. Zeegers MP, Tan FE, Dorant E, et al. The impact of characteristics of cigarette smoking on urinary tract cancer risk: a meta-analysis of epidemiologic studies. Cancer 2000 Aug 1;89(3):630-639.

60. Zaza S, Wright-De Aguero LK, Briss PA, et al. Data collection instrument and procedure for systematic reviews in the Guide to Community Preventive Services. Task Force on Community Preventive Services. Am J Prev Med 2000 Jan;18(1 Suppl):44-74.

61. van der Windt DA, Thomas E, Pope DP, et al. Occupational risk factors for shoulder pain: a systematic review. Occup Environ Med 2000 Jul;57(7):433-442.

62. Steinberg EP, Eknoyan G, Levin NW, et al. Methods used to evaluate the quality of evidence underlying the National Kidney Foundation-Dialysis Outcomes Quality Initiative Clinical Practice Guidelines: description, findings, and implications. Am J Kidney Dis 2000 Jul;36(1):1-11.

63. Harris EC, Barraclough BM. Suicide as an outcome for medical disorders. Medicine (Baltimore) 1994 Nov;73(6):281-296.

64. Harbour R, Miller J. A new system for grading recommendations in evidence based guidelines. BMJ 2001 Aug 11;323(7308):334-336.

65. Macfarlane TV, Glenny AM, Worthington HV. Systematic review of population-based epidemiological studies of oro-facial pain. J Dent 2001 Sep;29(7):451-467.

66. Pilote L, Tager IB. Outcomes research in the development and evaluation of practice guidelines. BMC Health Serv Res 2002 Mar 25;2(1):7.

67. Jain A, Concato J, Leventhal JM. How good is the evidence linking breastfeeding and intelligence? Pediatrics 2002 Jun;109(6):1044-1053.

68. Bhutta AT, Cleves MA, Casey PH, et al. Cognitive and behavioral outcomes of school-aged children who were born preterm: a meta-analysis. JAMA 2002 Aug 14;288(6):728-737.

69. Al-Jader LN, Newcombe RG, Hayes S, et al. Developing a quality scoring system for epidemiological surveys of genetic disorders. Clin Genet 2002 Sep;62(3):230-234.

70. Carneiro AV. Critical appraisal of prognostic evidence: practical rules. Rev Port Cardiol 2002 Jul-Aug;21(7-8):891-900.

71. Elwood M. Forward projection--using critical appraisal in the design of studies. Int J Epidemiol 2002 Oct;31(5):1071-1073.

72. Campbell H, Rudan I. Interpretation of genetic association studies in complex disease. Pharmacogenomics J 2002; 2(6):349-360.

73. Manchikanti L, Singh V, Vilims BD, et al. Medial branch neurotomy in management of chronic spinal pain: systematic review of the evidence. Pain Physician 2002 Oct;5(4):405-418.

74. Slim K, Nini E, Forestier D, et al. Methodological index for non-randomized studies (minors): development and validation of a new instrument. ANZ J Surg 2003 Sep;73(9):712-716.

75. Scholten-Peeters GGM, Verhagen AP, Bekkering GE, et al. Prognostic factors of whiplash-associated disorders: a systematic review of prospective cohort studies. Pain 2003; 104(1):303-322.

76. Rangel SJ, Kelsey J, Colby CE, et al. Development of a quality assessment scale for retrospective clinical studies in pediatric surgery. J Pediatr Surg 2003 Mar;38(3):390-396; discussion 396.

77. Meijer R, Ihnenfeldt DS, van Limbeek J, et al. Prognostic factors in the subacute phase after stroke for the future residence after six months to one year. A systematic review of the literature. Clin Rehabil 2003 Aug;17(5):512-520.

78. Centre for Evidence Based Mental Health (Oxford England). CEBMH : Critical Appraisal Forms. [S.l.]: Centre for Evidence Based Mental Health; 2000.

79. Federal Focus Inc. Epidemiologic data in regulatory risk assessments : recommendations for implementing the "London principles" and for risk assessment guidance. Washington, D.C.: Federal Focus, Inc. (11 Dupont Circle, Ste. 700, Washington DC 20036); 1996.

80. Scottish Intercollegiate Guidelines Network, Harbour RT, Forsyth L. Sign 50. A guideline developer's handbook. Rev. ed. Ediburgh, Scotland: Scottish Intercollegiate Guidelines Network; 2008.

81. Wells GA SB, O'Connell D, Peterson J, Welch V, Losos M,, P T. Quality Assessment Scales for Observational Studies.: Ottawa Health Research Institute; 2004.

82. Woodbury MG, Houghton PE. Prevalence of pressure ulcers in Canadian healthcare settings. Ostomy Wound Manage 2004 Oct; 50(10):22-24, 6, 8, 30, 2, 4, 6-8.

83. Tooth L, Ware R, Bain C, et al. Quality of reporting of observational longitudinal research. Am J Epidemiol 2005 Feb 1;161(3):280-288.

84. Moja LP, Telaro E, D'Amico R, et al. Assessment of methodological quality of primary studies by systematic reviews: results of the metaquality cross sectional study. BMJ 2005 May 7;330(7499):1053.

85. Pavia M, Pileggi C, Nobile CG, et al. Association between fruit and vegetable consumption and oral cancer: a meta-analysis of observational studies. Am J Clin Nutr 2006 May;83(5):1126-1134.

86. de Boer AG, Verbeek JH, van Dijk FJ. Adult survivors of childhood cancer and unemployment: A metaanalysis. Cancer 2006 Jul 1;107(1):1-11.

87. Shea B, Boers M, Grimshaw JM, et al. Does updating improve the methodological and reporting quality of systematic reviews? BMC Med Res Methodol 2006;6:27.

88. Bornhoft G, Maxion-Bergemann S, Wolf U, et al. Checklist for the qualitative evaluation of clinical studies with particular focus on external validity and model validity. BMC Med Res Methodol 2006;6:56.

89. Moher D. Reporting research results: a moral obligation for all researchers. Can J Anaesth 2007 May;54(5):331-315.

90. Genaidy AM, Lemasters GK, Lockey J, et al. An epidemiological appraisal instrument—a tool for evaluation of epidemiological studies. Ergonomics 2007 Jun;50(6):920-960.

91. Eichler K, Puhan MA, Steurer J, et al. Prediction of first coronary events with the Framingham score: a systematic review. Am Heart J 2007 May;153(5):722-731, 31 e1-e8.

92. Hirtz D, Thurman DJ, Gwinn-Hardy K, et al. How common are the "common" neurologic disorders? Neurology 2007 Jan 30;68(5):326-337.

93. Tricco AC, Tetzlaff J, Sampson M, et al. Few systematic reviews exist documenting the extent of bias: a systematic review. J Clin Epidemiol 2008 May;61(5):422-434.

94. Lundh A, Gotzsche PC. Recommendations by Cochrane Review Groups for assessment of the risk of bias in studies. BMC Med Res Methodol 2008;8:22.

95. Conde-Agudelo A, Rosas-Bermudez A, Kafury-Goeta AC. Effects of birth spacing on maternal health: a systematic review. Am J Obstet Gynecol 2007 Apr;196(4):297-308.

96. Waetjen LE, Liao S, Johnson WO, et al. Factors associated with prevalent and incident urinary incontinence in a cohort of midlife women: a longitudinal analysis of data: study of women's health across the nation. Am J Epidemiol 2007 Feb 1;165(3):309-318.

97. Luutonen S, Holm H, Salminen JK, et al. Inadequate treatment of depression after myocardial infarction. Acta Psychiatr Scand 2002 Dec; 106(6):434-439.

98. Hunskaar S, Lose G, Sykes D, et al. The prevalence of urinary incontinence in women in four European countries. BJU Int 2004 Feb;93(3):324-330.

99. Lauzon C, Beck CA, Huynh T, et al. Depression and prognosis following hospital admission because of acute myocardial infarction. CMAJ 2003 Mar 4;168(5):547-552.

100. Lesperance F, Frasure-Smith N, Talajic M, et al. Five-year risk of cardiac mortality in relation to initial severity and one-year changes in depression symptoms after myocardial infarction. Circulation 2002 Mar 5;105(9):1049-1053.

101. Leppavuori A, Pohjasvaara T, Vataja R, et al. Insomnia in ischemic stroke patients. Cerebrovasc Dis 2002;14(2):90-97.

102. Crew KD, Gammon MD, Terry MB, et al. Polymorphisms in nucleotide excision repair genes, polycyclic aromatic hydrocarbon-DNA adducts, and breast cancer risk. Cancer Epidemiol Biomarkers Prev 2007 Oct;16(10):2033-2041.

103. Nuotio M, Jylha M, Luukkaala T, et al. Urinary incontinence in a Finnish population aged 70 and over. Prevalence of types, associated factors and self-reported treatments. Scand J Prim Health Care 2003 Sep;21(3):182-187.

104. Dishion TJ, Eddy JM, Haas E, et al. Friendships and violent behavior during adolescence. Social Development 1997;6(2):207-223.

105. Vilbergsson S, Sigurdsson G, Sigvaldason H, et al. Coronary heart disease mortality amongst non-insulin-dependent diabetic subjects in Iceland: the independent effect of diabetes. The Reykjavik Study 17-year follow up. J Intern Med 1998 Oct;244(4):309-316.

106. Grady D, University of California San Francisco-Stanford Evidence-Based Practice Center, Agency for Healthcare Research and Quality. Diagnosis and Treatment of Coronary Heart Disease in Women; Systematic Reviews of Evidence on Selected Topics. Evidence Report/Technology Assessment No. 81. Rockville, MD: U.S. Dept. of Health and Human Services, Public Health Service, Agency for Healthcare Research and Quality; 2003.

107. Chan LS, Agency for Healthcare Research and Quality. Preventing Violence and Related Health-Risking Social Behaviors in Adolescents. Evidence Report/Technology Assessment. Summary No. 107. Rockville, MD: Agency for Healthcare Research and Quality; 2004.

108. Bush DE, Agency for Healthcare Research and Quality, Johns Hopkins University Evidence-based Practice Center. Post-Myocardial Infarction Depression. Evidence Report/Technology Assessment No. 123. Rockville, MD: Agency for Healthcare Research and Quality; 2005.

109. Buscemi N, Agency for Healthcare Research and Quality, University of Alberta Evidence-based Practice Center. Manifestations and Management of Chronic Insomnia in Adults. Evidence Report/Technology Assessment No. 125. Rockville, MD: Agency for Healthcare Research and Quality; 2005.

110. King JE. Software solutions for obtaining a kappa-type statistic for use with multiple raters. Paper presented at: The annual meeting of the Southwest Educational Research Association, 2004; Dallas, TX.

111. Fleiss JL. Statistical methods for rates and proportions. Second Edition: John Wiley and Sons, Inc. New York, NY; 1981.

112. Efron B, Tibshirani R. An introduction to the bootstrap. New York: Chapman & Hall; 1993.

113. Gwet K. Inter-rater reliability: dependency on trait prevalence and marginal homogeneity. Statistical Methods for Inter-Rater Reliability Assessment Series 2002;2:1-9.

114. Gwet K. Computing inter-rater reliability with the SAS system. Stat Methods Inter-Rater Reliability Assess 2002;3:1-16.

115. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol 2008 May;61(Pt 1):29-48.

116. Walter SD, Eliasziw M, Donner A. Sample size and optimal designs for reliability studies. Stat Med 1998 Jan 15;17(1):101-110.

117. R: A language and environment for statistical computing [computer program]. Version. Vienna, Austria; 2008.

118. R Project for Statistical Computing, R Foundation for Statistical Computing, Technische Universität Wien. Institut für Statistik Wahrscheinlichkeitstheorie und Versicherungsmathematik., et al. R news—the newsletter of the R Project. Wien: Technische Universität Wien, Institut für Statistik, Wahrscheinlichkeitstheorie, und Versicherungsmathematik; 2001.

119. SAS Institute. SAS version 9.1.3. Chicago, IL: SAS Institute Inc.; 2008: 6 CD-ROMs.

120. Vandenbroucke JP. STREGA, STROBE, STARD, SQUIRE, MOOSE, PRISMA, GNOSIS, TREND, ORION, COREQ, QUOROM, REMARK... and CONSORT: for whom does the guideline toll? J Clin Epidemiol 2009 Jan 30.

121. Mullner M, Matthews H, Altman DG. Reporting on statistical methods to adjust for confounding: a cross-sectional survey. Ann Intern Med 2002 Jan 15;136(2):122-126.

122. Little J, Higgins JP, Ioannidis JP, et al. STrengthening the REporting of Genetic Association Studies (STREGA): an extension of the STROBE statement. PLoS Med 2009 Feb 3;6(2):e22.

123. Shea BJ, Grimshaw JM, Wells GA, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. BMC Med Res Methodol 2007;7:10.

124. Needleman I, Worthington H, Moher D, et al. Improving the completeness and transparency of reports of randomized trials in oral health: the CONSORT statement. Am J Dent 2008 Feb;21(1):7-12.

125. Marsh HW, Jayasinghe UW, Bond NW. Improving the peer-review process for grant applications: reliability, validity, bias, and generalizability. Am Psychol 2008 Apr;63(3):160-168.

126. Johnson VE. Statistical analysis of the National Institutes of Health peer review system. Proc Natl Acad Sci USA 2008 Aug 12;105(32):11076-11080.

127. Demicheli V, Di Pietrantonj C. Peer review for improving the quality of grant applications. Cochrane Database Syst Rev 2007;(2):MR000003.

Appendix A. Methodological Evaluation of Observational Research

Methodological Evaluation of Observational REsearch (MORE)—Observational Studies of Incidence or Prevalence of Chronic Diseases

Please define the protocol specific for your research quality components:

1. Define and justify **target population** _____
   Define and justify population subgroups if applicable, race_____, gender_____, other_____
2. **Response rate.** Justify acceptable response rate:_____ and rate that can be defined as a major flaw of the study_____ in the total sample and in race, gender, and other subgroups if applicable.
3. **Exclusion rate from the analysis** - define in the protocol ranges specific for your research_____ and rate that can be defined as a major flaw of the study_____ in the total sample and in race, gender, and other subgroups if applicable.
4. **Source of measure incidence/prevalence of chronic diseases.** Define and justify minor flaws specific for the nature of the condition:

Sources	
Suggested Minor Flaws	
Self reported (collected for the study)	
Proxy reported (collected for the study)	Minor flaw
Objectively measured with diagnostic methods for the purpose of the study (independent of health care)	
Measured by interviewers for the study	
Obtained during clinical exam for the purpose of the study	
Obtained from medical records (mining of the data collected for health care purposes)	Minor flaw
Obtained from administrative database (mining of the data collected for health care purposes)	Minor flaw
Obtained from registries or administrative databases (collected for epidemiologic evaluation independent of health care)	
Other (please specify)	

1. **Reference period** (time of occurrence) in a definition of the outcome. Define and justify reference period specific for the nature of the outcomes_____
2. **Severity** (degree of the symptoms of the chronic disease) in a definition of the outcome. Define and justify severity if applicable for the nature of the outcomes_____
3. **Frequency of the symptoms of the chronic disease in definition of the outcome.** Define and justify importance of frequency per day, week, or month specific for the nature of the disease_____
4. **Dependent variable (outcomes) in subpopulations.** Define and justify the major flaw in assessment of the variables in subpopulations, if applicable _____
5. **Gold standard to measure the outcomes.** Define and justify gold standard (if known) to measure outcomes _____
6. **Reliability of the estimates.** Define and justify acceptable intra-observer variability_____ and inter-observer reliability _____

Instructions about the survey forms in Access format:

(1) If you are using Office 2007, probably you'll see an "Option" button right above this window. Please click on the button and choose "Enable this context."
(2) For a questions ending with a minor flaw symbol, please provide at least one response.
(3) When you are typing in a textbox, your input is not saved until you click on any other textbox or checkbox.
(4) You can exit the program at anytime and then resume the survey later by selecting the same Article ID.
(5) Help is available by clicking on the word Help next to the item you see.
(6) Though a textbox for "Other (please specify)" shows only about 2 lines of text, it can contain more than 6,000 words. This is just like a small window to see a big world.

Descriptive

Article ID (file name) _____

Journal of publication_____ Year of publication_____

Funding of study (Mark one best (*) and all applicable responses):

A	Not reported	Poor reporting
B	Industry	
C	Grant	
D	Combined industry + grant	
E	Other (specify)	

Role of funding organization in data analysis and interpretations of the results (mark one best (*) and all applicable responses):

A	Not reported	Poor reporting
B	Sponsoring organization participated in data analyses	
C	Other (please specify)	
D	Sponsoring organization did not participate in data analyses and interpretation	

Conflict of interest (Mark one best (*) and all applicable responses):

A	Disclosure not reported	Poor reporting
B	Reported not having conflict of interest	
C	Reported having conflict of interest	
D	Other (please specify)	

Country_____

Ethical approval of the study (Mark one best (*) and all applicable responses):

A	Not reported	Poor reporting
B	Study was approved by Ethical Committee	
C	Other (please specify)	

Aim of study (Mark one best (*) and all applicable responses):

A	Aim was not stated	Poor reporting
B	Included prevalence estimation in the general population	
C	Included prevalence estimation in racial subgroups	
D	Included prevalence estimation in sex subgroups	
E	Included prevalence estimation in other population subgroups (define)	
F	Included prevalence estimation without clear target population	Minor flaw
G	Included Incidence estimation in the general population	
H	Included Incidence estimation in racial subgroups	
I	Included Incidence estimation in sex subgroups	
J	Included Incidence estimation in other population subgroups (define)	
K	Included Incidence estimation without clear target population	Minor flaw

Study Design (Mark one best (*) and all applicable responses):

A	Not clear statement	Poor reporting
B	Cross-sectional	
C	Retrospective	
D	Prospective	
E	Other (please specify)	

External Validity

Sampling of the subjects by the investigators. General population based (Mark one best (*) and all applicable responses):

A	Not reported	Poor reporting
B	Random population based	
C	Non-random population based	
D	Random multistage population based	
E	Random stratified population based	
F	Random sampling restricted to geographic area (minor flaw if the aim was to examine incidence/prevalence in the general population without place restrictions)	Minor flaw
G	Other sampling of the general population (please specify)	

Nongeneral population based sampling method (Mark one best (*) and all applicable responses):

A	Not reported	Poor reporting
B	Random	
C	Convenient	Minor flaw
D	Self selection	Minor flaw
E	Other (specify)	

Nongeneral population based sampling frame (Mark one best (*) and all applicable responses):

A	Not reported	
B	Sampling within nationally representative registries or databases	
C	Medical records	Major flaw
D	Insurance claims	Major flaw
E	Work place	Major flaw
F	Health care based (clinics, hospitals)	Major flaw
G	Proxy selection (parents, relatives, legal representatives, care takers...)	
H	Other (please specify)	

Assessment of sampling bias - failure to ensure that all members of the reference population have a known chance of selection in the sample (Mark one best (*) and all applicable responses)

A	No information about sampling bias	Poor reporting
B	Sampling bias was assessed by the authors - differences in study population vs. target population are reported	
C	The authors did not assess sampling bias	Minor flaw
D	The authors did not assess sampling bias but justified exclusion of the subjects from the sampling or analysis	
E	Other (please specify)	

Estimate bias

Response rate in total sample: define the protocol ranges specific for research area. Please note that included ranges are simply illustrative; they need to be justified and vary with each systematic review. (Mark one best (*) and all applicable responses).

A	Not reported	Poor reporting
B	>60%	
C	<40%	Major flaw
D	40-60%	
E	Other (specify)	

Response rate in race subgroups (if applicable): define the protocol ranges specific for the research area. Please note that included ranges are simply illustrative; they need to be justified and vary with each systematic review.

A	Not reported	Poor reporting
B	>60%	
C	<40%	Major flaw
D	40-60%	
E	Other (specify)	

Response rate in gender subgroups (if applicable)—define the protocol ranges specific for research area. Please note that included ranges are simply illustrative; they need to be justified and vary with each systematic review.

A	Not reported	Poor reporting
B	>60%	
C	<40%	Major flaw
D	40-60%	
E	Other (specify)	

One study could examine incidence or prevalence in the total sample and in population subgroups with different probability of bias/error. Please decide if quality assessment is needed for each population subgroup.
If yes, abstract information adding evaluation tables for as many subgroups as you need. Specify definition of each subgroup.

Response rate in other subgroups - define the protocol ranges specific for research area. Please note that included ranges are simply illustrative; they need to be justified and vary with each systematic review.

A	Not reported	Poor reporting
B	>60%	
C	<40%	Major flaw
D	40-60%	
E	Other (specify)	

Exclusion rate from the analysis - define the protocol ranges specific for research area. Please note that included ranges are simply illustrative; they need to be justified and vary with each systematic review.

A	Not reported	Poor reporting
B	>10%	Major flaw
C	0-5%	
D	6-10%	
E	Other (please specify)	

Exclusion rate in subgroups (if applicable):

A	Not reported	Poor reporting
B	>10%	Major flaw
C	0-5%	
D	6-10%	
E	Different exclusion rate in evaluated subgroups (specify)	

Address Bias

Sampling bias is addressed in the analysis:

A	Not reported	Poor reporting
B	Weighting of the estimates by probability of selection	
C	Weighting of the estimates by non-response adjustment within sampling subgroups	
D	Post-stratification by age	
E	Post-stratification by sex	
F	Post-stratification by race	
G	Not addressed in analysis	Minor flaw
H	Other (please specify0	

Subject flow (define in the protocol the acceptable ranges specific for the area of research):

A	Not applicable for study design	
B	Number screened	
C	Number of screened not reported	Poor reporting
D	Number of eligible	
E	Number eligible not reported	Poor reporting
F	Number enrolled	
G	Number of enrolled not reported	Poor reporting

Recruitment fractions (automatically calculated):

Eligibility fraction: # eligible / # screened
Enrollment fraction: # enrolled / # eligible
Recruitment fraction: # enrolled / # screened
Number needed to screen: 1 / recruitment fraction

Internal Validity

Source of measure incidence/prevalence of chronic diseases (dependent variables) (define in the protocol flaws specific for the nature of the condition). (Mark one best (*) and all applicable responses)

A	Not reported	Poor reporting
B	Self reported (collected for the study)	
C	Proxy reported (collected for the study)	Minor flaw
D	Objectively measured with diagnostic methods for the purpose of the study (independent on health care)	
E	Measured by interviewers for the study	
F	Obtained during clinical exam for the purpose of the study	
G	Obtained from medical records (mining of the data collected for health care purposes)	Minor flaw
H	Obtained from administrative database (mining of the data collected for health care purposes)	Minor flaw
I	Obtained from registries or administrative databases (collected for epidemiologic evaluation independent of health care)	
J	Other (please specify)	

Reference period (time of occurrence) (defined in the protocol reference period specific for the nature of the outcomes). (Mark one best (*) and all applicable responses)

A	Reference period not relevant for the nature of the outcome	
B	Reference period may be relevant but not included in definition of the outcome (define relevance specific for research question)	Minor flaw
C	Reference period recommended by the CDC or guidelines (12 months for chronic diseases) is included in definition of the outcome	
D	Reference period different from recommended is justified and included in the definition	
E	Reference period different from recommended and not justified	Minor flaw
F	Other please (specify)	

Severity (degree of the symptoms of the chronic disease) (define importance of severity specific for the nature of the disease). (Mark one best (*)and all applicable responses)

A	Severity is not relevant for the outcome	
B	Severity can be relevant but not assessed in the study	Major flaw
C	Definition of the outcomes included severity of conditions	
D	Other (please specify)	

Frequency of the symptoms of the chronic disease (define in the protocol importance of frequency per day, week, or month specific for the nature of the disease). (Mark one best (*) and all applicable responses)

A	Frequency is not relevant for the outcome	
B	Frequency can be relevant but not assessed in the study	Minor flaw
C	Definition of the outcomes included frequency of diagnostic criterion of chronic conditions	
D	Other (please specify)	

Validation. (Mark one best (*) and all applicable responses):

A	No information about validation	Poor reporting
B	Variables were measured using known “gold standard” (define specific for the outcomes)	
C	Methods to measure outcomes were validated with gold standard	
D	The authors reported inter-methods validation (one method vs. another)	Minor flaw
E	The authors did not validate the methods to measure dependent variables (nonvalid methods were obtained)	Major flaw
F	The authors justified validity of the used methods from previously published research	
G	Other (please specify)	

Reliability of the estimates. (Mark one best (*) and all applicable responses)

A	Not reported	Poor reporting
B	Reliability assumed acceptable according to previous published analyses (medical coding, insurance claims)	
C	Intra-observer variability is within acceptable for the outcome standards (define acceptable variability specific for the nature of the outcome)	
D	Intra-observer variability is reported with subjective judgment of reliability	Minor flaw
E	Inter-observer variability is within acceptable for the outcome standards (define acceptable variability specific for the nature of the outcome)	
F	Inter-observer variability is reported with subjective judgment of reliability	Minor flaw
G	Other (please specify)	

Dependent variable (outcomes) in subpopulations (if applicable). (Mark applicable responses)

A	Measurements of the outcomes in subpopulations were not clarified	Poor reporting
B	The same methods were used to measure outcome in the total sample and in subgroups	
C	Outcomes in subpopulations were measured differently (define in the protocol the major flaw in assessment of the variables in subpopulations in applicable)	Minor flaw
D	Other (please specify)	

Reporting of prevalence: Type (Mark the best responses)

A	Not clear	Poor reporting
B	Point prevalence	Minor flaw
C	Period prevalence	
D	Other (please specify)	

Precision of estimate (error, 95% CI). (Mark one best (*) and all applicable responses)

A	Omitted	Poor reporting
B	Reported	
C	Other (please specify)	

Prevalence in total sample. (Mark the best responses):

A	Crude prevalence in total sample	Minor flaw
B	Age adjusted prevalence in total sample	
C	Other (specify)	

Prevalence in population subgroup (define relevant subgroups specific for research question). (Mark one best (*) and all applicable responses)

A	Stated as aim of the study but nor reported	Poor reporting
B	Crude prevalence in age subgroups	
C	Crude prevalence in race groups	Minor flaw
D	Crude prevalence in gender groups	Minor flaw
E	Crude prevalence other subgroups	Minor flaw
F	Age adjusted prevalence in race subpopulations	
G	Age adjusted prevalence in gender subpopulations	
H	Standardized estimation of prevalence by age and gender	
I	Age adjusted prevalence in other subgroups	
J	Other (please specify)	

Reporting of Incidence: Incidence Type. (Mark one best (*) and all applicable responses)

A	Not clear	Poor reporting
B	Cumulative incidence	
C	Incidence rate	
D	Other (specify)	

Precision of estimation (error, 95% CI). (Mark one best (*) and all applicable responses)

A	Omitted	Poor reporting
B	Reported	
C	Other (specify)	

Incidence in total sample. (Mark one best (*) and all applicable responses)

A	Crude incidence in total sample	Minor flaw
B	Age adjusted incidence in total sample	
C	Other (specify)	

Incidence in population subgroups (define relevant subgroups specific for research question). Mark one best (*) and all applicable responses

A	Stated in the aim of the study but not reported	Poor reporting
B	Crude incidence in age subgroups	
C	Crude incidence in race groups	Minor flaw
D	Crude incidence in gender groups	Minor flaw
E	Age adjusted incidence in race subpopulations	
F	Age adjusted incidence in gender subpopulations	
G	Standardized estimation of incidence by age and gender	
H	Crude incidence in other subgroups	Minor flaw
I	Age adjusted incidence in other subgroups	
J	Other (specify)	

Example of Quality Validity Report

Item	Issue
Article: _____	
Evaluator: _____	
External Validity	
Not reported	
Estimation of sampling bias: Exclusion rate from the analysis	Not reported
Estimation of sampling bias: Response rate in total sample	Not reported
Sampling: Assessment of sampling bias	No information about sampling bias
Sampling: Sampling method, Not general population based	Not reported
Estimation of sampling bias: Addressing sampling bias	Not reported
Internal Validity	
Minor	
Definition of incidence/prevalence: Frequency of symptoms	Can be relevant but not assessed in the study
Not Reported	
Measurements of incidence/prevalence: Reliability	Not reported
Article: _____	
Evaluator: _____	
External Validity	
Major	
Estimation of sampling bias: Exclusion rate from the analysis	>10%
Sampling: Sampling method: Nongeneral population based	Health care based (clinics, hospitals)
Minor	
Sampling: Sampling method: Nongeneral population based	Convenient
Not reported	
Estimation of sampling bias: Subject flow	Number of screened not reported
Estimation of sampling bias: Addressing sampling bias	Not reported
Sampling: Assessment of sampling bias	No information about sampling bias
Article: _____	
Evaluator: _____	
External Validity	
Major	
Sampling: Sampling frame: Nongeneral population based	Health care based (clinics, hospitals)
Minor	
Sampling: Sampling method: Nongeneral population based	Convenient
Not reported	
Estimation of sampling bias: Addressing sampling bias	Not reported
Estimation of sampling bias: Exclusion rate from the analysis	Not reported
Estimation of sampling bias: Subject flow	Number of eligible not reported
Estimation of sampling bias: Subject flow	Number of screened not reported
Sampling: Assessment of sampling bias	No information about sampling bias
Internal Validity	
Not Reported	
Measurements of incidence/prevalence: Reliability	Not reported

Methodological Evaluation of Observational Research (MEVORECH)—Observational Studies of Risk Factors of Chronic Diseases

Please define in the protocol specific for your research quality components:

1. Define and justify **target population** _____
   Define and justify population subgroups if applicable, race_____, gender_____, other_____
2. Define and justify exposure (risk factors) _____
3. **Response rate.** Justify acceptable response rate:_____ and rate that can be defined as a major flaw of the study_____ in the total sample and in race, gender, and other subgroups if applicable.
4. **Exclusion rate from the analysis** - define in the protocol ranges specific for your research_____ and rate that can be defined as a major flaw of the study_____ in the total sample and in race, gender, and other subgroups if applicable
5. **Source of measure outcomes. D**efine and justify minor flaws specific for the nature of the condition:

Sources	**Suggested minor flaws**
Self reported (collected for the study)	
Proxy reported (collected for the study)	Minor flaw
Objectively measured with diagnostic methods for the purpose of the study (independent on health care)	
Measured by interviewers for the study	
Obtained during clinical exam for the purpose of the study	
Obtained from medical records (mining of the data collected for health care purposes)	Minor flaw
Obtained from administrative database (mining of the data collected for health care purposes)	Minor flaw
Obtained from registries or administrative databases (collected for epidemiologic evaluation independent of health care)	
Other (please specify)	

1. **Reference period** (time of occurrence) in a definition of the outcome. Define and justify reference period specific for the nature of the outcomes_____
2. **Severity** (degree of the symptoms of the chronic disease) in a definition of the outcome. Define and justify severity is applicable for the nature of the outcomes_____
3. **Frequency of the symptoms of the chronic disease in a definition of the outcome. De**fine and justify importance of frequency per day, week, or month specific for the nature of the disease_____
4. **Gold standard to measure the outcomes.** Define and justify gold standard (if known) to measure outcomes _____
5. **Reliability of the estimates.** Define and justify acceptable Intra-observer variability_____ and inter-observer reliability _____
6. **Source of measure exposure. D**efine and justify minor flaws specific for the nature of the condition:

Source	**Suggested minor flaw**
Self reported (collected for the study)	
Proxy reported (collected for the study)	Minor flaw
Objectively measured with diagnostic methods for the purpose of the study (independent on health care)	
Measured by interviewers for the study	
Obtained during clinical exam for the purpose of the study	
Obtained from medical records (mining of data collected for health care purposes)	Minor flaw
Obtained from administrative database (mining of data collected for health care purposes)	Minor flaw
Obtained from registries (collected for epidemiologic evaluation independent of health care)	
Other (please specify)	

Reference period (time of occurrence) in a definition of the Exposure. Define and justify reference period specific for the nature of exposure _____
Length of exposure when applicable in the definition/assessment of exposure**.** Define and justify a length of exposure that was established by consensus of the experts or in guidelines _____
Intensity/dose of exposure. Define and justify importance of dose specific for the nature of the exposure (list for each risk factor_____
Measure of exposure. Define and justify gold standards to measure risk factors:
Factor _____ known gold standard _____
Confounding factors or factors that can modify the association between risk factor and disease. Define and justify set of major confounding factors specific for the association of the interest_____
Measure of confounding factors. Define and justify gold standards to measure primary confounding factors.
Factor _____ known gold standard _____
Loss of followup. Define and justify acceptable cutoff for loss of followup_____
Appropriateness of statistical model to reduce research specific bias. Define and justify the most appropriate methods specific for research questions_____

Instructions about the survey forms in Access format:

(1) If you are using Office 2007, probably you'll see an "Option" button right above this window. Please click on the button and choose "Enable this context."
(2) For a questions ending with a Minor flaw symbols, please provide at least one response.
(3) When you are typing in a textbox, your input is not saved until you click on any other textbox or checkbox.
(4) You can exit the program at anytime, and then resume the survey later by selecting the same Article ID.
(5) Help is available by clicking on the word Help next to the item you see.
(6) Though a textbox for "Other (please specify)" shows only about 2 lines of text, it can contain more than 6,000 words. This is just like a small window to see a big world.

Descriptive

Journal of publication_____
Year of publication_____
Funding of study (Mark one best (*) and all applicable responses)

A	Not reported	Poor reporting
B	Industry	
C	Grant	
D	Combined industry + grant	
E	Other (please specify)	

Role of funding organization in data analysis and interpretations of the results (Mark all applicable responses):

A	Not reported	Poor reporting
B	Sponsoring organization participated in data analyses	
C	Other (specify)	
D	Sponsoring organization did not participate in data analyses and interpretation	

Conflict of interest (Mark all applicable responses):

A	Disclosure not reported	Poor reporting
B	Reported not having conflict of interest	
C	Reported having conflict of interest	
D	Other (specify)	

Country_____

Ethical approval of the study (Mark all applicable responses):

A	Not reported	Poor reporting
B	Study was approved by Ethical Committee	
C	Other (specify)	

Aim

Aim of the study. (Mark one best (*) and all applicable responses)

A	Aim was not stated	Poor reporting
B	Included association with risk factors in the general population	
C	Included association with risk factors in race subgroups	
D	Included association with risk factors in gender subgroups	
E	Included association with risk factors in other population subgroups (define: diseases, specific demographics, socio-economic, or legal status, access to health insurance...)	
F	Included association with risk factors without clear definition of the target population	Minor flaw
G	Other (please specify)	

Objectives

(Mark one best (*) and all applicable responses)

A	Not clear statement	Poor reporting
B	Estimation of the association with prevalence of chronic conditions	
C	Estimation of the association with incidence of chronic conditions	
D	Other (please specify)	

Design

Study Design (Mark one best (*) and all applicable responses)

A	Not clear statement about the study design	Poor reporting
B	Cross-sectional	
C	Cohort (prospective) study with concurrent controls	
D	Cohort (retrospective) study with concurrent controls	
E	Case-controlled (retrospective) study	
F	Cohort (prospective) study with historical controls	
G	Nested case-control	
H	Other (please specify)	

External Validity

Sampling of the subjects by investigators

General population based (Mark one best (*) and all applicable responses)

A	Not reported	Poor reporting
B	Random population based	
C	Nonrandom population based	
D	Random multistage population based	
E	Random stratified population based	
F	Random sampling restricted to geographic area	
G	Other sampling of the general population (please specify)	

Nongeneral population based sampling method (Mark one best (*) and all applicable responses)

A	Not reported	Poor reporting
B	Random	
C	Convenient	Minor flaw
D	Self selection	Minor flaw
E	Other (please specify)	

Nongeneral population-based sampling frame (Mark one best (*) and all applicable responses)

A	Not reported	
B	Sampling within nationally representative registries or databases	
C	Medical records	Major flaw
D	Insurance claims	Major flaw
E	Work place	Major flaw
F	Health care based (clinics, hospitals)	Major flaw
G	Proxy selection (parents, relatives, legal representatives, caretakers...)	
H	Other (please specify)	

For case-control studies. (Mark one best (*) and all applicable responses)

A	Sampling of controls are not clearly reported	Poor reporting
B	Sampling of controls from the sample population as cases	
C	Sampling of controls from different population as cases	Major flaw
D	Sampling of controls from health care related sources (out-clinic or in-clinics, health care claims)	Minor flaw
E	Sampling of controls from work-related sources	
F	Sampling of controls from multiple sources	
G	Other (please specify)	

Assess bias

Assessment of sampling bias (failure to ensure that all members of the reference population have a known chance of selection in the sample). (Mark one best (*) and all applicable responses)

A	No information about sampling bias	Poor reporting
B	Sampling bias was assessed by the authors - differences in study population vs. target population are reported	
C	The authors did not assess sampling bias	Minor flaw
D	The authors did not assess sampling bias but justified exclusion of the subjects from the sampling or analysis	
E	Other (please specify)	

Estimate bias

Response rate in total sample - define in the protocol ranges specific for research area. Please note that included ranges are simply illustrative; they need to be justified and vary with each systematic review. (Mark one best (*) and all applicable responses)

A	Not reported	Poor reporting
B	>40 %	
C	<10-20%	Major flaw
D	21-40%	
E	Other (please specify)	

Exclusion rate from the analysis in total sample (define in the protocol acceptable ranges specific for research question). (Mark one best (*) and all applicable responses)

A	Not reported	Poor reporting
B	>10%	Major flaw
C	0-5%	
D	6-10%	
E	Other (specify)	

Exclusion rate from the analysis in exposed and not exposed (Mark one best (*) and all applicable responses)

A	Exclusion from the analyses was not reported separately for exposed and nonexposed	Poor reporting
B	Reasons to exclude from the analyses were the same for exposed and not exposed	
C	Reasons to exclude from the analyses differ for exposed and not exposed	Major flaw
D	Specify reasons for exclusion	

Address Bias

Sampling bias is addressed in the analysis. (Mark one best (*) and all applicable responses)

A	Not reported	Poor reporting
B	Weighting of the estimates by probability of selection	
C	Weighting of the estimates by nonresponse adjustment within sampling subgroups	
D	Post-stratification by age	
E	Post-stratification by sex	
F	Post-stratification by race	
G	Not addressed in analysis	Minor flaw
H	Other (please specify)	

Subject flow (define in the protocol the acceptable ranges specific for the area of research) (Mark one best (*) and all applicable responses)

A	Not applicable for study design	
B	Number of screened	
C	Not reported	Poor reporting
D	Number eligible	
E	Not reported	Poor reporting
F	Number enrolled	
G	Not reported	Poor reporting

Calculations with query Recruitment fractions (Insert calculated number, %)	
A	Eligibility fraction: # eligible / # screened
C	Enrollment fraction: # enrolled / # eligible
E	Recruitment fraction: # enrolled / # screened
G	Number needed to screen: 1 / recruitment fraction

Internal Validity

Source to measure dependent variables (target, outcomes) (define in the protocol flaws specific for the nature of the condition). (Mark one best (*) and all applicable responses)

A	Not reported	Poor reporting
B	Self reported (collected for the study)	
C	Proxy reported (collected for the study)	Minor flaw
D	Objectively measured with diagnostic methods for the purpose of the study (independent on health care)	
E	Measured by interviewers for the study	
F	Obtained during clinical exam for the purpose of the study	
G	Obtained from medical records (mining of data collected for health care purposes)	Minor flaw
H	Obtained from administrative database (mining of data collected for health care purposes)	Minor flaw
I	Obtained from registries (collected for epidemiologic evaluation independent of health care)	
J	Other-please specify	

Dependent variable
Reference period, time of occurrence of the disease (define reference period specific for the nature of the outcomes). (Mark one best (*) and all applicable responses)

A	Reference period not relevant for the nature of the outcome	
B	Reference period may be relevant but not included in definition of the outcome (define relevance specific for research question)	Minor flaw
C	Reference period recommended by the CDC or guidelines (12 months for chronic diseases) is included in definition of the outcome	
D	Reference period different from recommended is justified and included in the definition	
E	Reference period different from recommended and not justified	Minor flaw
F	Other (please specify)	

Severity, degree of the symptoms of the chronic condition (define importance of severity specific for the nature of the outcomes). (Mark one best (*) and all applicable responses)

A	Severity is not relevant for the outcome	
B	Severity can be relevant but not assessed in the study	Major flaw
C	Definition of the outcomes included severity of conditions	
D	Other (please specify)	

Frequency of the symptoms (define importance of frequency per day, week, or month specific for the nature of the outcomes). (Mark one best (*) and all applicable responses)

A	Frequency is not relevant for the outcome	
B	Frequency can be relevant but not assessed in the study	Minor flaw
C	Definition of the outcomes included frequency of diagnostic criterion of chronic conditions	
D	Other (please specify)	

Validation (Mark one best (*) and all applicable responses)

A	No information about validation	Poor reporting
B	Variables were measured using known “gold standard” (define specific for the outcomes)	
C	Methods to measure outcomes were validated with gold standard	
D	The authors reported inter-methods validation (one method vs. another)	Minor flaw
E	The authors did not validate the methods to measure dependent variables (nonvalid methods were obtained)	Major flaw
F	The authors justified validity of the used methods from previously published research	
G	Other (please specify)	

Reliability of the estimates (Mark one best (*) and all applicable responses)

A	Not reported	Poor reporting
B	Reliability assumed acceptable according to previous published analyses (medical coding, insurance claims)	
C	Intra-observer variability is within acceptable for the outcome standards (define acceptable variability specific for the nature of the outcome)	
D	Intra-observer variability is reported with subjective judgment of reliability	Minor flaw
E	Inter-observer variability is within acceptable for the outcome standards (define acceptable variability specific for the nature of the outcome)	
F	Inter-observer variability is reported with subjective judgment of reliability	Minor flaw
G	Other (please specify)	

When one study reported several risk factors with different probability of bias/error among tested hypotheses, please decide if quality assessment is needed for each risk factor.
If yes, abstract information adding as many risk factors as you need. Define risk factor or list risk factors for which quality assessment would be the same.
Define risk factor or list risk factors for which quality assessment would be the same:

Source to measure exposure
Hypothesis specific: complete for each risk factor. Source to measure exposure (risk factors, independent variables, input). (Mark one best (*) and all applicable responses)

A	Not reported	Poor reporting
B	Self reported (collected for the study)	
C	Proxy reported (collected for the study)	Minor flaw
D	Objectively measured with diagnostic methods for the purpose of the study (independent on health care)	
E	Measured by interviewers for the study	
F	Obtained during clinical exam for the purpose of the study	
G	Obtained from medical records (mining of data collected for health care purposes)	Minor flaw
H	Obtained from administrative database (mining of data collected for health care purposes)	Minor flaw
I	Obtained from registries (collected for epidemiologic evaluation independent of health care)	
J	Other (please specify)	

Define exposure
Definition of the exposure (risk factors, independent variables) (specific for research questions)
Hypothesis specific: complete for each risk factor.
Reference period/length of exposure (define reference period specific for the nature of the exposure risk factors, independent variables). (Mark one best (*) and all applicable responses)

A	Reference period/length of exposure not relevant for the nature of exposure	
B	Reference period/length of exposure may be relevant but not included in definition of the exposure (define relevance specific for research question)	Minor flaw
C	Reference period/length of exposure recommended by guidelines is included in definition of exposure	
D	Reference period/length of exposure different from recommended is justified and included in the definition	
E	Reference period/length of exposure different from recommended and not justified	Minor flaw
F	Other (please specify)	

Hypothesis specific: complete for each risk factor. Intensity/dose (define importance of dose specific for the nature of the exposure (risk factors, independent variables). (Mark one best (*) and all applicable responses)

A	Intensity/dose is not relevant for exposure	
B	Intensity/dose can be relevant but not assessed in the study	Minor flaw
C	Definition of the exposure (risk factors, independent variables) included intensity/dose	
D	Other (please specify)	

Measure exposure
Measurements of the exposure (risk factors, independent variables).
Hypothesis specific: complete for each risk factor. Validation. (Mark one best (*) and all applicable responses)

A	Not reported	Poor reporting
B	Exposure (risk factors, independent variables) were measured using known “gold standard” (define specific for the exposure)	
C	Methods to measure exposure (risk factors, independent variables) were validated with gold standard	
D	The authors reported inter-methods validation (one method vs. another)	Minor flaw
E	The authors did not validate the methods to measure exposure (risk factors, independent variables)	Major flaw
F	The authors justified validity of the used methods from previously published research	
G	Other (please specify)	

Hypothesis specific: complete for each risk factor. Reliability of the estimates. (Mark one best (*) and all applicable responses)

A	Not reported	Poor reporting
B	Reliability assumed acceptable according to previous published analyses	
C	Intra-observer variability is acceptable for exposure standards (define acceptable variability specific for the nature of exposure)	
D	Intra-observer variability is reported with subjective judgment of reliability	Minor flaw
E	Inter-observer variability is within acceptable for exposure standards (define acceptable variability specific for the nature of exposure)	
F	Inter-observer variability is reported with subjective judgment of reliability	
G	Other (please specify)	

Design specific. For case-control studies. (Mark one best (*) and all applicable responses)

A	The same methods were used to measure exposure risk factors, independent variable) in cases and controls	
B	The authors did not state that the same methods were used to measure exposure risk factors, independent variable) in cases and controls	Minor flaw
C	The authors used different methods to measure exposure (risk factors, independent variable) in cases and controls	Major flaw
D	Other (please specify)	

Confounding factors or factors that can modify the association between risk factor and disease (define in the protocol the primary confounding factors specific for the association of the interest). Mark one best (*) and all applicable responses

A	Not reported	Poor reporting
B	Major confounding factors/effect modifiers were not assessed	Major flaw
C	Major confounding factors /effect modifiers were assessed partially	Minor flaw
D	Major confounding factors/effect modifiers were assessed (known sets of confounders specific for research questions)	
E	Other (please specify)	

Measure of confounding factors (define the protocol gold standards to measure primary confounding factors specific for the research question). (Mark one best (*) and all applicable responses)

A	Not reported	Poor reporting
B	Valid measurements of major confounding factors	
C	Unknown validity to measure confounding factors	Minor flaw
D	Non valid methods to measure confounding factors	Major flaw
E	The authors justified validity of the used methods from previously published research	
F	Other (please specify)	

Followup

Loss of followup (define acceptable important cut off specific for research question). (Mark one best (*) and all applicable responses)

A	Not reported
B	% in total sample
C	% among exposed and not exposed
D	Not applicable (no followup in the study)
E	Loss of followup is larger than acceptable
F	Other (please specify)

Design specific for case-control studies. (Mark one best (*) and all applicable responses)

A	Not reported	Poor reporting
B	% of nonresponse among cases the same as for controls	
C	% of nonresponse differed among cases and controls	Minor flaw
D	% of nonresponse reported for cases only	Minor flaw
E	Other (please specify)	

Mask Exposure

Masking of exposure status for investigators who measured dependent variables (outcomes)

A	Not reported	Poor reporting
B	Was stated	
C	Was not possible	
D	Was possible but not obtained	Minor flaw
E	Was stated and assessed	
F	Other (please specify)	

Statistics

Statistical analysis. (Mark one best (*) and all applicable responses)

A	Not reported	Poor reporting
B	Standardization	
C	Matching	
D	Adjustment in multivariate model	
E	Stratification	
F	Propensity scoring	
G	The authors did not obtain methods to reduce bias	Major flaw
H	Several methods to reduce bias	
I	Other methods were justified and obtained to reduce bias (please specify)	

Temporality

For cohort studies.

Design and hypothesis specific. Assessment of temporality. (Mark one best (*) and all applicable responses)

A	Not reported	Poor reporting
B	Demonstration that exposure preceded the outcome (the disease of interest was not present at start of study)	
C	Other (specify)	

Appropriateness

Appropriateness of statistical model to reduce research specific bias (define in the protocol the most appropriate methods specific for research questions). (Mark one best (*) and all applicable responses)

A	Strategies to reduce research specific bias not reported	Poor reporting
B	Authors justified using appropriate statistical models to reduce research specific bias	
C	Authors did not use statistical models that may be the most appropriate according to the published literature (examples may include population stratification bias in case-control studies of genetic association, odds ratio in cohort studies of common diseases, missing data, large loss of followup)	Minor flaw
D	Authors did not justify choice of statistical models to reduce research specific bias	Minor flaw
E	Authors attempted to reduce bias in post hoc statistical adjustment	Minor flaw
F	Other (please specify)	

Dose response

Hypothesis specific: complete for each risk factor. Dose response with exposure. (Mark one best (*) and all applicable responses)

A	Not relevant for research question	
B	May be relevant but not reported	Poor reporting
C	Reported as significant	
D	Reported as nonsignificant	
E	Other (please specify)	

Report

Hypothesis specific. Reporting of tested hypothesis. (Mark one best (*) and all applicable responses)

A	Unclear reporting of the estimates (unclear model, reference level, set of confounding factors…)	Poor reporting
B	Crude estimates	Major flaw if C is not marked
C	Authors reported estimates of primary and secondary hypotheses adjusted for confidential sources of bias	
D	Incomplete selective reporting of the tested hypotheses (compared to aim and objectives)	Minor flaw
E	Other (please specify)	

Precision

Hypothesis specific. Precision of the estimates (Mark one best (*) and all applicable responses)

A	Mean with 95% CI reported	
B	Mean and standard error of estimates reported	
C	Numeric value of estimates not reported (p value only, significance or non significance only)	Minor flaw
D	Mean only reported without p value or variance	Poor reporting
E	Other (please specify)	

Sample Size
Sample size justification. (Mark one best (*) and all applicable responses)

A	Not reported	Poor reporting
B	Justified for primary outcome	
C	Justified for secondary outcomes	
D	Justification by authors is incomplete or inaccurate	Minor flaw
E	Post-hoc analyses	Minor flaw
F	Other (please specify)	

Example of Quality Validity Report

Item	Issue
Article: _____	
Evaluator: _____	
External Validity	
Not reported	
Estimation of sampling bias: Addressing sampling bias	Not reported
Estimation of sampling bias: Response rate in total sample	Not reported
Estimation of sampling bias: Subject flow	Number of eligible not reported
Assessment of sampling bias	No information about sampling bias
Internal Validity	
Major	
Measurement of dependent variable (target=outcomes): Validation	Did not validate the methods to measure dependent variables (nonvalid methods were obtained)
Minor	
Measure of confounding factors	Unknown validity to measure confounding factors
Not reported	
Masking of exposure status for investigators who measured dependent variables (outcomes)	Not reported
Measurements of dependent variable (target=outcomes): Reliability	Not reported
Article: _____	
Evaluator: _____	
External Validity	
Minor	
Sampling: For case control study	Sampling of controls from health care related sources (out clinic or in clinics, health care claims)
Not reported	
Estimation of sampling bias: Subject flow	Number of screened not reported
Estimation of sampling bias: Subject flow	Number of enrolled not reported
Estimation of sampling bias: Subject flow	Number of eligible not reported
Estimation of sampling bias: Response rate in total sample	Not reported
Estimation of sampling bias: Exclusion rate from analysis	Not reported
Sampling: Nongeneral population based sampling method	Not reported
Assessment of sampling bias	No information about sampling bias
Estimation of sampling bias: Addressing sampling bias	Not reported
Internal Validity	
Minor	
Measure of confounding factors	Unknown validity to measure confounding factors
Definition of the dependent variable (target=outcome): Reference period	May be relevant but not included in definition of the outcome
Not reported	

Loss of followup	Not reported
Masking of exposure status for investigators who measured dependent variables (outcomes)	Not reported
Article: _____	
Evaluator: _____	
External Validity	
Not reported	
Estimation of sampling bias: Subject flow	Number of eligible not reported
Estimation of sampling bias: Subject flow	Number of screened not reported
Estimation of sampling bias: Exclusion rate from the analysis	Not reported
Estimation of sampling bias: Addressing sampling bias	Not reported
Assessment of sampling bias	No information about sampling bias
Sampling: General population based	Not reported
Internal Validity	
Minor	
Confounding factors or the factors that can modify the association: risk factor and disease	Major confounding factors/effect modifiers were assessed partially
For cohort study: Appropriateness of statistical model to reduce research specific bias	Did not justify choice of statistical models to reduce research specific bias
Not reported	
Measurements of dependent variable (target=outcomes): Validation	No information about validation
Loss of followup	Not reported
Masking of exposure status for investigators who measured dependent variables (outcomes)	Not reported
Measure of confounding factors	Not reported
Measurements of dependent variable (target=outcomes): Reliability	Not reported

Methodological Evaluation of Observational REsearch (MORE)—Observational Studies of Incidence or Prevalence of Chronic Diseases

Instructions:
Please review the checklist and mark with X quality items that are not reported and flaws in external or internal validity if present.

Descriptive
Funding of study
Role of funding organization in data analysis and interpretations of the results
Conflict of interest
Ethical approval of the study

Aim of study

Not reported	☐ Poor reporting
Included prevalence estimation without clear target population	☐ Minor flaw
Included Incidence estimation without clear target population	☐ Minor flaw

External Validity

Sampling of the subjects by the investigators
General population based

Not reported	☐ Poor reporting
Random sampling restricted to geographic area (minor flaw if the aim was to examine incidence/prevalence in the general population without place restrictions)	☐ Minor flaw

Nongeneral population based sampling method

Not reported	☐	Poor reporting
Convenient	☐	Minor flaw
Self selection	☐	Minor flaw

Nongeneral population based sampling frame

Not reported	☐ Poor reporting
Medical records	☐ Major flaw
Insurance claims	☐ Major flaw
Work place	☐ Major flaw
Health care based (clinics, hospitals)	☐ Major flaw

Assessment of sampling bias - failure to ensure that all members of the reference population have a known chance of selection in the sample

Not reported	☐	Poor reporting
The authors did not assess sampling bias	☐	Minor flaw

Estimate bias
Response rate in total sample (cut off of acceptable response rate depend on the target population)

Not reported	☐ Poor reporting
<40 (or less than cut off specific for the target population) %	☐ Major flaw

Response rate in race subgroups (if applicable)

Not reported	☐ Poor reporting
<40 (or less than cut off specific for the target population) %%	☐ Major flaw

Response rate in other subgroups (if applicable)

Not reported	☐	Poor reporting
<40 (or less than cut off specific for the target population) %	☐	Major flaw

Exclusion rate from the analysis

Not reported	☐	Poor reporting
>10%	☐	Major flaw

Exclusion rate in subgroups (if applicable)

Not reported	☐	Poor reporting
>10%	☐	Major flaw

Address Bias
Sampling bias is addressed in the analysis

Not reported	☐	Poor reporting
Not addressed in analysis	☐	Minor flaw

Subject flow (the acceptable ranges can be specific for the area of research)

Not applicable for study design	☐	
Number screened not reported	☐	Poor reporting
Number eligible not reported	☐	Poor reporting
Number enrolled not reported	☐	Poor reporting

Internal Validity

Source of measure incidence/prevalence of chronic diseases

Not reported	☐	Poor reporting
Proxy reported (collected for the study)	☐	Minor flaw
Obtained from medical records (mining of the data collected for health care purposes)	☐	Minor flaw
Obtained from administrative database (mining of the data collected for health care purposes)	☐	Minor flaw

Reference period (time of occurrence) if applicable

Reference period not relevant for the nature of the outcome	☐	
Reference period may be relevant but not included in definition of the outcome (define relevance specific for research question)	☐	Minor flaw
Reference period different from recommended and not justified	☐	Minor flaw

Severity (degree of the symptoms of the chronic disease)

Severity is not relevant for the outcome	☐	
Severity can be relevant but not assessed in the study	☐	Major flaw

Frequency of the symptoms of the chronic disease

Frequency is not relevant for the outcome	☐	
Frequency can be relevant but not assessed in the study	☐	Minor flaw

Validation of outcomes measurements

No information about validation	☐	Poor reporting
The authors reported inter-methods validation (one method vs. another)	☐	Minor flaw
The authors did not validate the methods to measure dependent variables (nonvalid methods were obtained)	☐	Major flaw

Reliability of the estimates (mark one best (*) and all applicable responses)

Not reported	☐	Poor reporting
Intra-observer variability is reported with subjective judgment of reliability	☐	Minor flaw
Inter-observer variability is reported with subjective judgment of reliability	☐	Minor flaw

Dependent variable (outcomes) in subpopulations (if applicable)		
Measurements of the outcomes in subpopulations were not clarified	☐	Poor reporting
Outcomes in subpopulations were measured differently (define in the protocol the major flaw in assessment of the variables in subpopulations in applicable)	☐	Minor flaw
Reporting of prevalence		
Not clear	☐	Poor reporting
Point prevalence	☐	Minor flaw
Precision of estimate (error, 95% CI).		
Omitted	☐	Poor reporting
Prevalence in total sample		
Crude prevalence in total sample	☐	Minor flaw
Prevalence in population subgroup if applicable		
Stated as aim of the study but not reported	☐	Poor reporting
Crude prevalence in race groups	☐	Minor flaw
Crude prevalence in gender groups	☐	Minor flaw
Crude prevalence other subgroups	☐	Minor flaw
Reporting of Incidence: Incidence type		
Not clear	☐	Poor reporting
Precision of estimation (error, 95% CI)		
Omitted	☐	Poor reporting
Incidence in total sample (mark one best (*) and all applicable responses)		
Crude incidence in total sample	☐	Minor flaw
Incidence in population subgroups if applicable		
Stated in the aim of the study but not reported	☐	Poor reporting
Crude incidence in race groups	☐	Minor flaw
Crude incidence in gender groups	☐	Minor flaw
Crude incidence in other subgroups	☐	Minor flaw

Quality Validity Report
(Access reports are generated based on responses above)

Item	Decision
Manuscript: _____	
Reviewer: _____	
External Validity	
Not reported	☐ Require reporting
Major flaws	☐ 1. Require justification that flaws could not be avoided or bias cannot be reduced ☐ 2. Reject manuscript
Minor flaws	☐ 1. Require justification that flaws could not be avoided or bias cannot be reduced ☐ 2. Reject manuscript

Internal Validity

Not reported	☐ Require reporting
Major flaws	☐ 1. Require justification that flaws could not be avoided or bias cannot be reduced ☐ 2. Reject manuscript
Minor flaws	☐ 1. Require justification that flaws could not be avoided or bias cannot be reduced ☐ 2. Reject manuscript

Methodological Evaluation of Observational Research (MEVORECH)—Observational Studies of Risk Factors of Chronic Diseases

Instructions :
Please review the checklist and mark quality items that not reported and flaws in external or internal validity if present.

Descriptive
Journal of publication _____
Year of publication _____
Funding of study
Role of funding organization in data analysis and interpretations of the results
Conflict of interest
Ethical approval of the study

Aim of the study

Aim was not stated	☐	Poor reporting
Included association with risk factors without clear definition of the target population	☐	Minor flaw

Objectives

Not clear statement	☐	Poor reporting
Estimation of the association with prevalence of chronic conditions	☐	
Estimation of the association with incidence of chronic conditions	☐	

Design

Not clear statement about the study design	☐	Poor reporting

External Validity

Sampling of the subjects by investigators
General population based

Not reported	☐	Poor reporting

Nongeneral population based sampling method

Not reported	☐	Poor reporting
Convenient	☐	Minor flaw
Self selection	☐	Minor flaw

Nongeneral population-based sampling frame

Not reported	☐	
Medical records	☐	Major flaw
Insurance claims	☐	Major flaw
Work place	☐	Major flaw
Health care based (clinics, hospitals)	☐	Major flaw

For case-control studies

Sampling of controls are not clearly reported	☐	Poor reporting
Sampling of controls from different population as cases	☐	Major flaw
Sampling of controls from health care related sources (out-clinic or in-clinics, health care claims)	☐	Minor flaw

Assess bias

Assessment of sampling bias (failure to ensure that all members of the reference population have a known chance of selection in the sample)

No information about sampling bias	☐	Poor reporting
The authors did not assess sampling bias	☐	Minor flaw

Response rate in total sample (cut off of acceptable response rate depend on the target population)

Not reported	☐	Poor reporting
<40 (or less than cut off specific for the target population) %	☐	Major flaw

Response rate in race subgroups (if applicable)

Not reported	☐	Poor reporting
<40 (or less than cut off specific for the target population) %	☐	Major flaw

Response rate in other subgroups (if applicable)

Not reported	☐	Poor reporting
<40 (or less than cut off specific for the target population) %	☐	Major flaw

Exclusion rate from the analysis

Not reported	☐	Poor reporting
>10%	☐	Major flaw

Exclusion rate in subgroups (if applicable)

Not reported	☐	Poor reporting
>10%	☐	Major flaw

Subject flow (the acceptable ranges can be specific for the area of research)

Not applicable for study design	☐	
Number screened not reported	☐	Poor reporting
Number eligible not reported	☐	Poor reporting
Number enrolled not reported	☐	Poor reporting

Exclusion rate from the analysis in exposed and not exposed

Exclusion from the analyses was not reported separately for exposed and nonexposed	☐	Poor reporting
Reasons to exclude from the analyses differ for exposed and not exposed	☐	Major flaw

Address Bias

Sampling bias is addressed in the analysis. (Mark one best (*) and all applicable responses)

Not reported	☐	Poor reporting
Not addressed in analysis	☐	Minor flaw

Internal Validity

Source to measure dependent variables (target, outcomes)

Not reported	☐	Poor reporting
Proxy reported (collected for the study)	☐	Minor flaw
Obtained from medical records (mining of data collected for health care purposes)	☐	Minor flaw
Obtained from administrative database (mining of data collected for health care purposes)	☐	Minor flaw

Dependent variable

Reference period, time of occurrence of the disease

Reference period may be relevant but not included in definition of the outcome (define relevance specific for research question)	☐	Minor flaw
Reference period different from recommended and not justified	☐	Minor flaw

Severity, degree of the symptoms of the chronic condition

Severity is not relevant for the outcome	☐	
Severity can be relevant but not assessed in the study	☐	Major flaw

Frequency of the symptoms (decide importance of frequency per day, week, or month specific for the nature of the outcomes)

Frequency is not relevant for the outcome	☐	
Frequency can be relevant but not assessed in the study	☐	Minor flaw

Validation of outcomes measurements

No information about validation	☐	Poor reporting
The authors reported inter-methods validation (one method vs. another)	☐	Minor flaw
The authors did not validate the methods to measure dependent variables (nonvalid methods were obtained)	☐	Major flaw

Reliability of the estimates

Not reported	☐	Poor reporting
Intra-observer variability is reported with subjective judgment of reliability	☐	Minor flaw
Inter-observer variability is reported with subjective judgment of reliability	☐	Minor flaw

Source to measure exposure (can be completed for more than one risk factor)

Not reported	☐	Poor reporting

Proxy reported (collected for the study)	☐	Minor flaw
Obtained from medical records (mining of data collected for health care purposes)	☐	Minor flaw
Obtained from administrative database (mining of data collected for health care purposes)	☐	Minor flaw

Definition of the exposure (risk factors, independent variables)
Reference period/length of exposure)

Reference period/length of exposure not relevant for the nature of exposure	☐	
Reference period/length of exposure may be relevant but not included in definition of the exposure (define relevance specific for research question)	☐	Minor flaw
Reference period/length of exposure different from recommended and not justified	☐	Minor flaw

Intensity/dose

Intensity/dose is not relevant for exposure	☐	
Intensity/dose can be relevant but not assessed in the study	☐	Minor flaw

Measure exposure
Measurements of the exposure (can be completed for more than one risk factor)

Not reported	☐	Poor reporting
The authors reported inter-methods validation (one method vs. another)	☐	Minor flaw
The authors did not validate the methods to measure exposure (risk factors, independent variables)	☐	Major flaw

Reliability of exposure estimates

Not reported	☐	Poor reporting
Intra-observer variability is reported with subjective judgment of reliability	☐	Minor flaw

For case-control studies

The authors did not state that the same methods were used to measure exposure risk factors, independent variable) in cases and controls	☐	Minor flaw
The authors used different methods to measure exposure (risk factors, independent variable) in cases and controls	☐	Major flaw

Confounding factors or factors that can modify the association between risk factor and disease

Not reported	☐	Poor reporting
Major confounding factors/effect modifiers were not assessed	☐	Major flaw
Major confounding factors /effect modifiers were assessed partially	☐	Minor flaw

Measure of confounding factors		
Not reported	☐	Poor reporting
Unknown validity to measure confounding factors	☐	Minor flaw
Non valid methods to measure confounding factors	☐	Major flaw
Followup		
Loss of followup (acceptable important cut off can be specific for research question)		
Not reported	☐	Poor reporting
Not applicable (no followup in the study)	☐	
Loss of followup is larger than acceptable	☐	
For case-control studies		
Not reported	☐	Poor reporting
% of nonresponse differed among cases and controls	☐	Minor flaw
% of nonresponse reported for cases only	☐	Minor flaw
Masking of exposure status for investigators who measured dependent variables (outcomes)		
Not reported	☐	Poor reporting
Was possible but not obtained	☐	Minor flaw
Statistical analysis		
Not reported	☐	Poor reporting
The authors did not obtain methods to reduce bias	☐	Major flaw
Temporality (for cohort studies)		
Assessment of temporality		
Not reported	☐	Poor reporting
Appropriateness of statistical model to reduce research specific bias		
Strategies to reduce research specific bias not reported	☐	Poor reporting
Authors did not use statistical models that may be the most appropriate according to the published literature (examples may include population stratification bias in case-control studies of genetic association, odds ratio in cohort studies of common diseases, missing data, large loss of followup)	☐	Minor flaw
Authors did not justify choice of statistical models to reduce research specific bias	☐	Minor flaw
Authors attempted to reduce bias in post hoc statistical adjustment	☐	Minor flaw
Dose response with exposure		
Not relevant for research question	☐	
May be relevant but not reported	☐	Poor reporting
Reporting of tested hypothesis		

Unclear reporting of the estimates (unclear model, reference level, set of confounding factors…)	☐	Poor reporting
Crude estimates	☐	Major flaw
Incomplete selective reporting of the tested hypotheses (compared to aim and objectives)	☐	Minor flaw
Precision of the estimates		
Numeric value of estimates not reported (p value only, significance or non significance only)	☐	Minor flaw
Mean only reported without p value or variance	☐	Poor reporting
Sample size justification		
Not reported	☐	Poor reporting
Justification by authors is incomplete or inaccurate	☐	Minor flaw
Post-hoc analyses	☐	Minor flaw

Quality Validity Report
(Access reports are generated based on responses above)

Item	Decision
Manuscript: _____	
Reviewer: _____	
External Validity	
Not reported	☐ Require reporting
Major flaws	☐ 1. Require justification that flaws could not be avoided or bias cannot be reduced ☐ 2. Reject manuscript
Minor flaws	☐ 1. Require justification that flaws could not be avoided or bias cannot be reduced ☐ 2. Reject manuscript
Internal Validity	
Not reported	☐ Require reporting
Major flaws	☐ 1. Require justification that flaws could not be avoided or bias cannot be reduced ☐ 2. Reject manuscript
Minor flaws	☐ 1. Require justification that flaws could not be avoided or bias cannot be reduced ☐ 2. Reject manuscript

Methodological Evaluation of Observational REsearch (MORE)—Observational Studies of Population Incidence or Prevalence of Chronic Diseases

Suggested criteria for Level A exclusion from synthesis or Level C separate limited synthesis if major flaws detected

External Validity

Sampling of the subjects by the investigators

Nongeneral population based sampling frame

Not reported	Poor reporting	Level C
Medical records	Major flaw	Level A
Insurance claims	Major flaw	Level A
Work place	Major flaw	Level A
Health care based (clinics, hospitals)	Major flaw	Level A

Response rate in total sample (Cut off of acceptable response rate depend on the target population)

Not reported	Poor reporting	Level C
<40 (or less than cut off specific for the target population)%	Major flaw	Level A

Response rate in race or other subgroups (if applicable)

Not reported	Poor reporting	Level C
<40 (or less than cut off specific for the target population)%	Major flaw	Level A

Exclusion rate from the analysis

Not reported	Poor reporting	Level C
>10%	Major flaw	Level A

Exclusion rate in subgroups (if applicable):

Not reported	Poor reporting	Level C
>10%	Major flaw	Level A

Internal Validity

Source of measure incidence/prevalence of chronic diseases

Not reported	Poor reporting	Level C
Proxy reported (collected for the study)	Minor flaw	Level C
Obtained from medical records (mining of the data collected for health care purposes)	Minor flaw	Level C
Obtained from administrative database (mining of the data collected for health care purposes)	Minor flaw	Level C

Severity (degree of the symptoms of the chronic disease)

Severity can be relevant but not assessed in the study	Major flaw	Level A

Validation of outcomes measurements

No information about validation	Poor reporting	Level C
The authors did not validate the methods to measure dependent variables (nonvalid methods were obtained)	Major flaw	Level A

Incidence or prevalence in total sample

Crude prevalence in total sample	Minor flaw	Level C

Prevalence in population subgroup if applicable

Stated as aim of the study but not reported	Poor reporting	Level C
Crude prevalence in race groups	Minor flaw	Level C
Crude prevalence in gender groups	Minor flaw	Level C
Crude prevalence other subgroups	Minor flaw	Level C

Methodological Evaluation of Observational Research (MEVORECH)—Observational Studies of Risk Factors of Chronic Diseases

Stopping Rules

External Validity

Sampling of the subjects by investigators

Nongeneral population-based sampling frame

Not reported		Level C
Medical records	Major flaw	Level A
Insurance claims	Major flaw	Level A
Work place	Major flaw	Level A
Health care based (clinics, hospitals)	Major flaw	Level A

For case-control studies

Sampling of controls are not clearly reported	Poor reporting	Level C
Sampling of controls from different population as cases	Major flaw	Level A
Sampling of controls from health care related sources (out-clinic or in-clinics, health care claims)	Minor flaw	Level C

Response rate in total sample (Cut off of acceptable response rate depend on the target population)

Not reported	Poor reporting	Level C
<40 (or less than cut off specific for the target population) %	Major flaw	Level A

Response rate in race or other subgroups (if applicable)

Not reported	Poor reporting	Level C
<40 (or less than cut off specific for the target population) %	Major flaw	Level A

Exclusion rate from the analysis

Not reported	Poor reporting	Level C
>10%	Major flaw	Level A

Exclusion rate in subgroups (if applicable)

Not reported	Poor reporting	Level C
>10%	Major flaw	Level A

Exclusion rate from the analysis in exposed and not exposed

Exclusion from the analyses was not reported separately for exposed and nonexposed	Poor reporting	Level C
Reasons to exclude from the analyses differ for exposed and not exposed	Major flaw	Level C

Internal Validity

Source to measure dependent variables (target, outcomes)

Not reported	Poor reporting	Level C
Proxy reported (collected for the study)	Minor flaw	Level C
Obtained from medical records (mining of data collected for health care purposes)	Minor flaw	Level C
Obtained from administrative database (mining of data collected for health care purposes)	Minor flaw	Level C

Severity, degree of the symptoms of the chronic condition

Severity can be relevant but not assessed in the study	Major flaw	Level A

Validation of outcomes measurements

No information about validation	Poor reporting	Level C
The authors did not validate the methods to measure dependent variables (nonvalid methods were obtained)	Major flaw	Level A

Source to measure exposure (can be completed for more than one risk factor)

Not reported	Poor reporting	Level C
Proxy reported (collected for the study)	Minor flaw	Level C
Obtained from medical records (mining of data collected for health care purposes)	Minor flaw	Level C
Obtained from administrative database (mining of data collected for health care purposes)	Minor flaw	Level C

Measure exposure

Measurements of the exposure (can be completed for more than one risk factor)

Not reported	Poor reporting	Level C
The authors reported inter-methods validation (one method vs. another)	Minor flaw	Level C
The authors did not validate the methods to measure exposure (risk factors, independent variables)	Major flaw	Level A

For case-control studies		
The authors did not state that the same methods were used to measure exposure risk factors, independent variable) in cases and controls	Minor flaw	Level C
The authors used different methods to measure exposure (risk factors, independent variable) in cases and controls	Major flaw	Level A
Confounding factors or factors that can modify the association between risk factor and disease		
Not reported	Poor reporting	Level C
Major confounding factors/effect modifiers were not assessed	Major flaw	Level A
Major confounding factors /effect modifiers were assessed partially	Minor flaw	Level C
Statistical analysis		
Not reported	Poor reporting	Level C
The authors did not obtain methods to reduce bias	Major flaw	Level A
Appropriateness of statistical model to reduce research specific bias		
Strategies to reduce research specific bias not reported	Poor reporting	Level C
Authors did not use statistical models that may be the most appropriate according to the published literature (examples may include population stratification bias in case-control studies of genetic association, odds ratio in cohort studies of common diseases, missing data, large loss of followup)	Minor flaw	Level C
Authors did not justify choice of statistical models to reduce research specific bias	Minor flaw	Level C
Authors attempted to reduce bias in post hoc statistical adjustment	Minor flaw	Level C
Reporting of tested hypothesis		
Unclear reporting of the estimates (unclear model, reference level, set of confounding factors)	Poor reporting	Level C
Crude estimates	Major flaw	Level A
Incomplete selective reporting of the tested hypotheses (compared to aim and objectives)	Minor flaw	Level C

Appendix B. Reliability Testing of the Developed Checklists

Exhibit B1. Distribution of *n* subjects by rater and response

Rater B	Rater A		Total
	+	-	
+	a	b	**B+**
-	c	d	**B-**
Total	**A+**	**A-**	**n**

$P_{\alpha} = (a+d)/n$

Cohen's kappa statistics[1] is given by:

$K = [P_{\alpha} - P_{e(\kappa)}] / [1 - P_{e(\kappa)}]$

$P_{e(\kappa)} = [(P_{A+})*(P_{B+})] + [(P_{A-})*(P_{B-})]$

$P_{A+} = A+/n$; $P_{A-} = A-/n$; $P_{B+} = B+/n$; $P_{B-} = B-/n$.

The AC1 statistic record #1092 is given by:

$AC1 = [P_{\alpha} - P_{e(\lambda)}] / [1 - P_{e(\lambda)}]$,

where $P_{e(\lambda)}$ is defined as follows:

$P_{e(\lambda)} = (2P_{+}) * (1 - P_{+})$

$P_{+} = [(A+ + B+)/2]/n$.

Exhibit B2. Fleiss' and Gwet's generalized kappa for the each article quality components using excel software[2]

$$K = \frac{p_o - p_e}{1 - p_e},$$

where $p_o = \sum_{i=1}^{k} p_{ii}$, $p_e = \sum_{i=1}^{k} p_{i.} p_{.i}$, and p = the proportion of ratings by two raters on a scale having k categories.

Fleiss' extension of kappa (called the generalized kappa):

$$K = 1 - \frac{nm^2 - \sum_{i=1}^{n} \sum_{j=1}^{k} x_{ij}^2}{nm(m-1)\sum_{j=1}^{k} \overline{p}_j \overline{q}_j},$$

where k = the number of categories, n = the number of subjects rated, m = the number of raters, $\overline{p}_j$ = the mean proportion for category j, and $\overline{q}_j$ = 1 – the mean proportion for category j.

Fleiss' standard error:

$$SE(K) \approx \sqrt{\frac{2}{Nm(m-1)} \times \frac{P(E) - (2m-3)[P(E)]^2 + 2(m-2)\sum p_j^3}{[1-P(E)]^2}},$$

where $P(E) = \sum_{j=1}^{m} p_j^2$ and $\sum p_j^3 = \sum_{j=1}^{m} p_j^3$.

Fleiss, Nee, and Landis corrected the standard error:

$$SE(K) \approx \frac{\sqrt{2}}{\sum_{j=1}^{k} \overline{p}_j \overline{q}_j \sqrt{nm(m-1)}} \times \sqrt{\left(\sum_{j=1}^{k} \overline{p}_j \overline{q}_j\right)^2 - \sum_{j=1}^{k} \overline{p}_j \overline{q}_j (\overline{q}_j - \overline{p}_j)}.$$

Table B1. General kappa and AC1 statistics: pilot reliability testing of quality assessment of four studies of incidence/prevalence of chronic diseases by seven expert groups

Quality Item	Response	Kappa	Standard Error	AC1 Statistics	Standard Error
External validity					
General population based sampling	Not reported	-0.04	0.95	0.92	0.08*
	Random population based	0.24	0.24	0.54	0.26*
	Nonrandom population based	-0.04	0.95	0.92	0.08*
	Random stratified population based	-0.08	0.64	0.84	0.10*
	Random sampling restricted to geographic area	0.03	0.34	0.60	0.17*
Nongeneral population based sampling method	Not reported	0.10	0.64	0.86	0.14*
	Random	-0.04	0.95	0.92	0.08*
	Convenient	-0.02	0.24	0.39	0.21*
	Self selection	-0.04	0.95	0.92	0.08*
Nongeneral population based sampling frame	Not reported	-0.04	0.95	0.92	0.08*
	Health care based (clinics, hospitals)	0.86	0.11	0.86	0.14*
	Proxy selection (parents, relatives, legal representatives, caretakers...)	-0.04	0.95	0.92	0.08*
Assessment of sampling bias	No information about sampling bias	0.04	0.11	0.05	0.13
	Sampling bias was assessed by the authors –- differences in study population vs. target population are reported	0.27	0.34	0.70	0.19*
	The authors did not assess sampling bias	-0.14	0.34	0.53	0.07*
	The authors did not assess sampling bias but justified exclusion of the subjects from the sampling or analysis	-0.08	0.64	0.84	0.10*
Sampling bias is addressed in the analysis	Not reported	-0.05	0.11	-0.04	0.05
	Post-stratification by age	-0.08	0.64	0.84	0.10*
	Post-stratification by sex	0.10	0.64	0.86	0.14*
	Post-stratification by race	0.10	0.64	0.86	0.14*
	Not addressed in analysis	0.01	0.21	0.32	0.23
Internal validity					
Source to measure outcomes	Not reported	-0.04	0.95	0.92	0.08*
	Self reported (collected for the study)	0.37	0.24	0.62	0.22*
	Objectively measured with diagnostic methods for the purpose of the study (independent on health care)	0.30	0.21	0.52	0.28*
	Measured by interviewers for the study	-0.17	0.40	0.62	0.00*
	Obtained during clinical exam for the purpose of the study	-0.04	0.95	0.92	0.08*
Reference period in definition	Reference period not relevant for the nature of the outcome	-0.02	0.24	0.39	0.21*
	Reference period may be relevant but not included in definition of the outcome (define relevance specific for research question)	-0.07	0.40	0.65	0.13*

Quality Item	Response	Kappa	Standard Error	AC1 Statistics	Standard Error
	Reference period recommended by the CDC or guidelines (12 months for chronic diseases) is included in definition of the outcome	0.05	0.24	0.43	0.22*
	Reference period different from recommended is justified and included in the definition	-0.08	0.64	0.84	0.10*
	Reference period different from recommended and not justified	-0.08	0.64	0.84	0.10*
Severity in definition	Severity is not relevant for the outcome	-0.08	0.64	0.84	0.10*
	Severity can be relevant but not assessed in the study	0.13	0.21	0.40	0.23*
	Definition of the outcomes included severity of conditions	0.27	0.16*	0.38	0.23*
Frequency of the symptoms	Frequency is not relevant for the outcome	0.18	0.18	0.37	0.24
	Frequency can be relevant but not assessed in the study	0.05	0.24	0.43	0.22*
	Definition of the outcomes included frequency of diagnostic criterion of chronic conditions	0.17	0.16	0.30	0.24
Validation of outcomes measure	No information about validation	0.43	0.28	0.71	0.18*
	Variables were measured using known “gold standard” (define specific for the outcomes)	0.01	0.28	0.50	0.19*
	Methods to measure outcomes were validated with gold standard	-0.04	0.95	0.92	0.08*
	The authors reported inter-methods validation (one method vs. another)	-0.04	0.95	0.92	0.08*
	The authors did not validate the methods to measure dependent variables (nonvalid methods were obtained)	0.00	0.49	0.76	0.14*
	The authors justified validity of the used methods from previously published research	0.07	0.16	0.21	0.26
Reliability of the estimates	Not reported	0.43	0.11*	0.43	0.33
	Reliability assumed acceptable according to previous published analyses (medical coding, insurance claims)	0.07	0.18	0.28	0.24
	Intra-observer variability is acceptable for the outcome standards (define acceptable variability specific for the nature of the outcome)	-0.04	0.95	0.92	0.08*
	Intra-observer variability is reported with subjective judgment of reliability	-0.04	0.95	0.92	0.08*

*significant agreement at 95% CI

Table B2. General kappa and AC1 statistics: pilot reliability testing of quality assessment of six studies of risk factors of chronic diseases by seven expert groups

Quality Item	Response	Kappa	Standard Error	AC1 Statistics	Standard Error
General population based sampling	Not reported	0.03	0.29	0.63	0.13*
	Random population based	0.14	0.29	0.67	0.16*
	Nonrandom population based	-0.02	0.97	0.95	0.05*
	Random multistage population based	-0.02	0.97	0.95	0.05*
	Random stratified population based	0.04	0.52	0.85	0.10*
	Random sampling restricted to geographic area	-0.04	0.33	0.66	0.12*
Nongeneral population based sampling method	Not reported	-0.13	0.26	0.50	0.06*
	Random	-0.05	0.66	0.90	0.07*
	Convenient	-0.01	0.43	0.79	0.10*
	Self selection	-0.08	0.52	0.84	0.07*
Nongeneral population based sampling frame	Not reported	-0.01	0.43	0.79	0.10*
	Sampling within nationally representative registries or databases	0.13	0.66	0.91	0.09*
	Medical records	-0.02	0.97	0.95	0.05*
	Health care based (clinics, hospitals)	0.38	0.26	0.72	0.17*
Sampling for case-control studies	Sampling of controls are not clearly reported	-0.02	0.97	0.95	0.05*
	Sampling of controls from the sample population as cases	0.13	0.21	0.50	0.17*
	Sampling of controls from health care related sources (out-clinic or in-clinics, health care claims)	0.13	0.66	0.91	0.09*
Assessment of sampling bias	No information about sampling bias	-0.05	0.09	-0.05	0.04
	Sampling bias was assessed by the authors—differences in study population vs. target population are reported	0.09	0.33	0.71	0.14*
	The authors did not assess sampling bias	-0.01	0.43	0.79	0.10*
	The authors did not assess sampling bias but justified exclusion of the subjects from the sampling or analysis	-0.05	0.66	0.90	0.07*
Exclusion rate from the analysis in exposed and not exposed	Exclusion from the analyses was not reported separately for exposed and not exposed	0.07	0.14	0.26	0.17
	Reasons to exclude from the analyses were the same for exposed and not exposed	0.14	0.29	0.67	0.16*
	Reasons to exclude from the analyses differ for exposed and not exposed	-0.08	0.52	0.84	0.07*
Sampling bias is addressed in the analysis	Not reported	0.08	0.09	0.08	0.19
	Weighting of the estimates by probability of selection	0.13	0.66	0.91	0.09*
	Weighting of the estimates by nonresponse adjustment within sampling subgroups	-0.05	0.66	0.90	0.07*
	Post-stratification by age	-0.08	0.52	0.84	0.07*
	Post-stratification by sex	0.17	0.37	0.78	0.14*
	Post-stratification by race	-0.02	0.97	0.95	0.05*
	Not addressed in analysis	-0.01	0.43	0.79	0.10*

Quality Item	Response	Kappa	Standard Error	AC1 Statistics	Standard Error
Source to measure dependent variables	Not reported	-0.02	0.97	0.95	0.05*
	Self reported (collected for the study)	0.01	0.09	0.02	0.09
	Proxy reported (collected for the study)	0.22	0.33	0.75	0.13*
	Objectively measured with diagnostic methods for the purpose of the study (independent on health care)	0.01	0.19	0.38	0.14*
	Measured by interviewers for the study	-0.10	0.33	0.64	0.08*
	Obtained during clinical exam for the purpose of the study	0.20	0.29	0.69	0.15*
	Obtained from medical records (mining of the data collected for health care purposes)	0.02	0.37	0.74	0.12*
	Obtained from administrative database (mining of the data collected for health care purposes)	0.04	0.52	0.85	0.10*
	Obtained from registries (collected for epidemiologic evaluation independent of health care)	-0.05	0.66	0.90	0.07*
Reference period, the time of the occurrence of the disease	Reference period not relevant for the nature of the outcome	-0.05	0.21	0.40	0.13*
	Reference period may be relevant but not included in definition of the outcome (define relevance specific for research question)	0.07	0.26	0.59	0.15*
	Reference period recommended by the CDC or guidelines (12 months for chronic diseases) is included in definition of the outcome	-0.04	0.23	0.47	0.13*
	Reference period different from recommended is justified and included in the definition	0.04	0.52	0.85	0.10*
Severity, the degree of the symptoms of the chronic condition	Severity is not relevant for the outcome	0.11	0.17	0.38	0.13*
	Severity can be relevant but not assessed in the study	0.10	0.19	0.43	0.18*
	Definition of the outcomes included severity of conditions	0.04	0.14	0.23	0.16
Frequency of the symptoms	Frequency is not relevant for the outcome	0.34	0.11*	0.39	0.16*
	Frequency can be relevant but not assessed in the study	0.21	0.21	0.55	0.16*
	Definition of the outcomes included frequency of diagnostic criterion of chronic conditions	0.04	0.21	0.45	0.13*
Validation	No information about validation	-0.01	0.12	0.10	0.08
	Variables were measured using known “gold standard” (define specific for the outcomes)	0.18	0.17	0.44	0.19*
	Methods to measure outcomes were validated with gold standard	-0.02	0.97	0.95	0.05*
	The authors did not validate the methods to measure dependent variables (nonvalid methods were obtained)	-0.05	0.66	0.90	0.07*

Quality Item	Response	Kappa	Standard Error	AC1 Statistics	Standard Error
	The authors justified validity of the used methods from previously published research	0.04	0.52	0.85	0.10*
Reliability of the estimates	Not reported	0.16	0.10	0.19	0.11
	Reliability assumed acceptable according to previous published analyses (medical coding, insurance claims)	-0.11	0.16	0.17	0.07
	Intra-observer variability is acceptable for the outcome standards (define acceptable variability specific for the nature of the outcome)	-0.02	0.97	0.95	0.05*
	Inter-observer variability is acceptable for the outcome standards (define acceptable variability specific for the nature of the outcome)	-0.02	0.97	0.95	0.05*
Confounding factors	Not reported	-0.05	0.66	0.90	0.07*
	Major confounding factors/effect modifiers were not assessed	-0.06	0.37	0.72	0.10*
	Major confounding factors/effect modifiers were assessed partially	0.03	0.33	0.68	0.12*
	Major confounding factors/effect modifiers were assessed (known sets of confounders specific for research questions)	-0.10	0.10	-0.06	0.04
Measure of confounding factors	Not reported	0.20	0.13	0.32	0.17*
	Valid measurements of major confounding factors	-0.04	0.16	0.22	0.09
	Unknown validity to measure confounding factors	-0.11	0.43	0.77	0.07*
	The authors justified validity of the used methods from previously published research	-0.01	0.43	0.79	0.10*
Masking of exposure status	Not reported	0.11	0.11	0.17	0.09
	Was stated	0.08	0.43	0.81	0.12*
	Was not possible	-0.03	0.26	0.54	0.12*
	Was possible but not obtained	-0.05	0.66	0.90	0.07*
Statistical analysis	Not reported	-0.02	0.97	0.95	0.05*
	Standardization	0.13	0.66	0.91	0.09*
	Matching	0.13	0.66	0.91	0.09*
	Adjustment in multivariate model	0.08	0.21	0.48	0.14*
	Stratification	0.07	0.26	0.59	0.15*
	The authors did not obtain methods to reduce bias	-0.05	0.66	0.90	0.07*
Appropriateness of statistical model to reduce research specific bias	Strategies to reduce research specific bias not reported	-0.01	0.43	0.79	0.10*
	Authors justified using appropriate statistical models to reduce research specific bias	0.07	0.13	0.21	0.09

Quality Item	Response	Kappa	Standard Error	AC1 Statistics	Standard Error
	The authors did not use statistical models that may be the most appropriate according to the published literature (examples may include population stratification bias in case-control studies of genetic association, odds ratio in cohort studies of common diseases, missing data, large loss of followup)	-0.02	0.97	0.95	0.05*
Appropriateness of statistical model to reduce research specific bias	The authors did not justify choice of statistical models to reduce research specific bias	-0.04	0.33	0.66	0.12*
Sample size justification	Not reported	-0.07	0.13	0.09	0.08
	Justified for primary outcome	-0.02	0.97	0.95	0.05*
	Justified for secondary outcomes	-0.02	0.97	0.95	0.05*
	Post-hoc analyses	-0.05	0.66	0.90	0.07*

*significant agreement at 95% CI

References

1. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. The British Journal of Mathematical and Statistical Psychology. 2008 May;61(Pt 1):29-48.

2. King JE. Software solutions for obtaining a kappa-type statistic for use with multiple raters. The annual meeting of the Southwest Educational Research Association; 2004; Dallas, TX; 2004