

Development and Implementation of the Standards for Evaluating and Reporting Epidemiologic Studies on Chronic Disease Incidence or Prevalence

Tatyana Shamliyan^{1,10,*}, Mohammed T. Ansari², Gowri Raman³, Nancy Berkman⁴, Mark Grant⁵, Gail Janes⁶, Margaret Maglione⁷, David Moher², Mona Nasser⁸, Karen Robinson⁹, Jodi Segal⁹, Sophia Tsouros²

¹Division of Health Policy and Management, University of Minnesota School of Public Health, Minneapolis

²Clinical Epidemiology Methods Centre, Ottawa Health Research Institute, Ottawa

³Tufts University Medical Center, Boston

⁴RTI International – University of North Carolina, Chapel Hill

⁵Blue Cross and Blue Shield Association, Chicago

⁶Centers for Disease Control and Prevention, Atlanta

⁷Southern California EPC; RAND Corporation, Santa Monica

⁸University of Plymouth, Peninsula Dental School, Plymouth, UK

⁹Johns Hopkins University, Baltimore

¹⁰Elsevier Clinical Solutions, Senior Director, Quality assurance

*Corresponding author: t.shamliyan@elsevier.com

Received January 8, 2013; Revised August 14, 2013; Accepted August 16, 2013

Abstract We aimed to develop quality checklists for observational non-therapeutic studies. Based on a systematic review of current practices of quality assessment of observational studies, collaborating co-authors from Evidence-based Practice Centers and the Centers for Disease Control and Prevention developed a new checklist for studies examining incidence and prevalence of chronic conditions, evaluated face and content validity, and discrimination validity to distinguish reporting from methodological quality. This new checklist is available in text format or as a relational database to produce standardized reports with flaws in reporting quality, external (six criteria), and internal (five criteria) validity of the studies. Study and hypotheses (subgroups) level analyses are possible with predetermined in protocol templates criteria of major and minor flaws. Consensus around justified research specific methodological standards and reliability tests should precede quality evaluation of primary studies to assure confidence in quality assessment. To be effective, policy decisions should be made based on comprehensive systematic evidence reviews that include transparent, standardized quality appraisals. Implementation of the developed checklists would increase transparency and quality of research leading to effective informed decisions in health care.

Keywords: risk factors, morbidity, reproducibility of results, validation studies, bias (epidemiology), quality control, review literature as topic

Cite This Article: Tatyana Shamliyan, Mohammed T. Ansari, Gowri Raman, Nancy Berkman, Mark Grant, Gail Janes, Margaret Maglione, David Moher, Mona Nasser, Karen Robinson, Jodi Segal, and Sophia Tsouros, "Development and Implementation of the Standards for Evaluating and Reporting Epidemiologic Studies on Chronic Disease Incidence or Prevalence." *American Journal of Public Health Research* 1, no. 7 (2013): 183-190. doi: 10.12691/ajphr-1-7-7.

1. Introduction

Decision makers in public and health care settings need comprehensive critically appraised synthesis of evidence about incidence and prevalence of chronic diseases [1]. Evidence based decision making process involves thorough systematic appraisal of internal and external validity in individual studies and body of evidence [2,3,4,5]. In the US chronic diseases cost an estimated \$1.5 trillion annually [6]. The large number of systematic reviews to summarize incidence or prevalence reflects a

growing interest in such estimations [7,8]. Estimates, however, vary across the studies. For example, the estimated prevalence of dementia in the US varies from 6 to 10% in older adults and from 40 to 58% in elderly persons [9]. Estimated age-standardized incidence of dementia varied from 10.5 to 15.6 per 100000 in men and from 15.2 to 19.4 per 100000 in women [10]. Quality of the primary studies may contribute to differences in the estimates and should be carefully appraised with predefined validated tools[1].

Prevalence or incidence of chronic diseases can be evaluated only in observational studies which are prone to increased risk of bias [11]. Thus, assessing the quality of

observational studies is essential in conducting systematic reviews and evidence based reports [1,12]. While several tools have been validated for therapeutic studies [13,14,15], an extensive review of available quality appraisals for observational studies concluded a need for reliable quality ratings for non-therapeutic studies of incidence and prevalence of chronic diseases [12,16,17,18].

We conducted a comprehensive review of the published checklists and scales for quality assessment of observational studies [16,17]. We analyzed 145 systematic reviews of observational nontherapeutic studies [17] and 84 publications that described 96 tools to assess quality of observational studies [16]. We defined observational nontherapeutic studies as observations of patient outcomes that did not examine procedures concerned with the remedial treatment or prevention of diseases [19]. We examined how systematic reviews [20,21,22] appraised quality of the primary studies [23], which tools they used (checklist [24] or scale [20]), information about content and previous validation and reliability of the tools, domains of quality assessment (external and internal validity, level of evidence), and how systematic reviews incorporated quality assessment into the synthesis of evidence. We use the term tools interchangeably for the checklists and scales for quality assessment. We concluded that available tools require subjective judgments about “appropriateness” of study design and execution or “adequacy” of the reducing bias strategies that vary substantially depending on specific areas of research [16,17]. Available tools did not discriminate various quality criteria; for example, the same score would be given for prospective study design or using valid outcome measurement [21]. The available tools did not discriminate reporting quality with internal or external validity [22,25,26,27].

To address this gap in quality assessment of non-therapeutic observational studies we aimed to develop valid and reliable quality criteria for observational studies that examine incidence and prevalence of chronic diseases. Our objectives included testing the validity and reliability of the checklist to achieve agreement around criteria for the design, reporting standards, and assessment of nontherapeutic observational studies in systematic reviews and evidence-based reports. Developed criteria ought to improve quality of systematic reviews and informed evidence based decision making [1].

2. Methods

Our analytical framework included several steps. First, based on our systematic literature review we developed a checklist that is available in the format of a relational database (Access) and in text format with the manual and instructions. Then we organized a collaborating effort to test the credibility and content validity of the checklist. We conducted a pilot reliability test of these quality appraisals by participating experts. After that we finalized content and interface of the checklist and identified directions for checklist implementation and future research.

The protocol for the development of the checklist to evaluate quality of nontherapeutic studies 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 [28]. We analyzed actual published tools using previously published criteria [29] and evaluated each criterion by applicability to incidence or prevalence studies and by relevance to examine external or internal validity [11,12,30,31]. We created the tool to assess the quality of studies of incidence/prevalence that included all validated quality components of external and internal validity.

We defined external validity as the extent to which the results of the study can be generalized to the target population. [11] Applicability may differ from external validity by the definition of the target population; for instance, well-designed studies from different countries with good external validity can have low applicability to the U.S. population. The definition of the target population is not a quality criterion; however, the extent to which the results can be generalizable to the target population (external validity) is. We defined internal validity as the extent to which results of the studies are correct for the study subjects and the associations detected in the study are truly caused by exposure. [11] We addressed risk of bias in primary studies but avoided labeling the biases in the quality evaluation because of differences in definitions of biases among scholars. For example, selection bias was defined as “the introduction of error due to systematic differences in the characteristics between those selected and those not selected for a given study” [32] or “systematic differences in comparison groups” [13,33] as a result of selective nonrandom treatment assignment. Selection of the criteria was designed to avoid duplication in the evaluation process.

We discriminated reporting quality from methodological quality of the studies by having the option of “not reported” for all quality criteria. We discriminated flaws in external and internal validity with two different reports; one with the list of poorly reported or flawed quality criteria of external, and another of internal validity. We used pre-specified major and minor flaws in external and internal validity. The standard reports separated for internal and external validity of the study have been developed to list major and minor flaws without formal scaling of criteria or summarizing them into global arithmetic score or obscured nontransparent quality rank. Incidence or prevalence estimates, therefore, can be compared across the studies with different reporting and methodological quality. The investigators of systematic reviews can incorporate reporting or methodological quality into sensitivity analyses and overall synthesis of evidence.

The co-authors from EPCs and from the CDC judged face content validity [28,34,35] and discriminant validity; and conducted pilot reliability testing.

We conducted a pilot test to examine inter-rater reliability by the participating experts.[36] We used Landis & Koch's measure of inter-rater agreement for multiple raters, with papers (studies) in place of subjects, when different studies were rated by different groups of raters. [37,38] We also calculated generalized kappa [39] and AC1 statistics for each quality component and each article [40,41,42] using Excel [39] and SAS [41] software .Since none of the statistical tests for reliability

of nominal multi-rater responses using checklists is ideal [28], 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, and 0.40-0.59 as moderate, 0.60-0.79 as substantial, and 0.80-1.00 as almost perfect agreement.

3. Results

We formulated the requirements for the checklist to assess quality of the studies of incidence or prevalence of chronic diseases.

We aimed to develop a comprehensive tool. The tool should include an exhaustive range of criteria and possible responses plus the option of open questions. Definitions of research specific biases should be pre-specified in the protocols of quality evaluation.

The tool should have mutually exclusive responses to avoid ambiguity in evaluations. The interface should have options to choose the best response, mark all applicable responses, or specify each quality component using access interface.

We aimed to develop a tool with realistic quality evaluation. The tool should define the best (gold standard) methodology that CDC uses to conduct Public Health Surveillance for Chronic Conditions for incidence/prevalence studies. The reviewers should have the flexibility to define biases that can be specific for research questions.

We aimed to develop a tool that discriminates overall quality estimation. We suggested that the proposed checklist includes predefined major flaws that must be pre-specified depending on the research topic. We decided to seek a balance between rigorous quality assessment and flexible applicability of the tools in different areas of research. Quality assessment would require transparent and justified definitions of the flaws that are planned in the protocol of the systematic reviews. The tool can't evaluate the exact probability of bias in external or internal validity since "true universal association" is unknown in most cases of observational nontherapeutic research. The report should contain a conclusion of applicability of the results to the general population or specific subpopulations and a conclusion of validity of the estimated incidence/prevalence.

We aimed to develop a tool with hypothesis level analyses of quality. No one published tool gave an opportunity to assess more than one hypothesis examined in the study [16]. However, subgroup analyses are preferable to make individualized decisions but at the same time are most vulnerable to bias. We proposed that the checklist must be able to evaluate validity of incidence or prevalence estimates overall and in subpopulations.

We decided that the grading the level of evidence should require additional information about consistency in results across the studies and should not be part of the standard report for individual studies.

We aimed to develop a tool with coherent quality evaluation. Basic knowledge in epidemiology should be required to complete the tool. Judgment about appropriateness of strategies to reduce bias should be standardized with minimal subjectivity in the evaluation.

We evaluated all components of the published tools for applicability to assess external or internal validity of observational studies. Then we generated the bank of criteria by applicability to observational studies of incidence/prevalence and by assessment of external or internal validity. Finally, we selected components relevant to studies of incidence/prevalence of chronic conditions. The draft checklist included an exhaustive range of criteria and possible responses plus the option of open questions. Definitions of research specific biases were pre-specified prior to development of the draft checklist. In this case, we used the CDC definitions used in conducting Public Health Surveillance for Chronic Conditions for incidence/prevalence studies.

A detailed description of the development of the checklists, validation and pilot reliability testing is reported elsewhere. [36] We then evaluated the face and content validity of the checklist (content, definitions of the flaws, and internal algorithm for the reports) and agreed upon six criteria for assessing external validity and five criteria for assessing internal validity. Pilot testing demonstrated face and content validities and discrimination of reporting vs. methodological qualities. [36] Inter-rater agreement was poor with a lower than expected kappa.

In order to improve reliability, we analyzed the reasons for poor reliability and proposed explicit operational definitions of the research specific quality standards. We detected areas of disagreement due to multiple response options for each question. Lack of clarity around research specific quality standards was the major area of disagreement. We recommend *a priori* discussion and consensus around appropriate definitions of the target population, population subgroups, or the reference methods of the measurements. [36] 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.

The finalized checklist has descriptive information about the study, six criteria of external and five criteria of internal validity (can be downloaded from https://netfiles.umn.edu/xythoswfs/webui/_xy-17471658_1-t_aRG151Im). The checklist is available in the format of an Access database that produces standardized reports categorizing criteria by reporting quality as well as by major and minor flaws in external and internal validity (can be downloaded from https://netfiles.umn.edu/xythoswfs/webui/_xy-17471658_1-t_aRG151Im). The reports are available in text format (Access reports) and spreadsheets that can be analyzed by statistical software to incorporate quality criteria to an overall strength of evidence grade.

The instruction manual provides examples and definitions of the quality components and examples from previously published research. These instructions and examples are also available as "help" files in the Access database. A template is also available to help reviewers achieve consensus about target populations and availability of gold standard to measure outcomes.

The checklist assesses external validity by assessing the sampling, inclusion, and exclusion of subjects from the study, and the differences between target and study population. [11] Studies which maintain participants

through each stage reduce the risk of a sampling bias and increase the probability that eligible subjects from the target population would be selected to the study (Table 1). Sampling bias is, thus, defined as failure to ensure that all members of the reference population have a known chance of selection in the sample. Each eligible individual in the target population can have the same (random population based samples) or different (nonrandom samples) probability of selection into the study [43,44,45,46]. The checklist suggests sampling strategies based on the practice of the CDC to use random multistage population based sampling. [43,44,45,46] We define random sampling restricted to geographic areas as a minor flaw if the aim of the study was to examine incidence/prevalence in the general population. Such restrictions may lead to false

estimations of incidence; for example, age adjusted incidence of prostate cancer per 100,000 male population varied from 360 or more in New Jersey and the District of Columbia to less than 300 in ten other states. [47] We defined a major flaw of a sampling frame as those derived from non-population based environments such as place of health care or employment, or symptom based inclusion criteria because prevalence of chronic diseases among specialty clinic or hospital or a working population, or among those with pre-defined symptoms may differ substantially compared to the general population. For example, the prevalence of fecal incontinence varied from approximately 0.7 -5-8% in a community based studies [48,49] to 12-19% among adults visiting primary care physicians or gastroenterologists [50].

Table 1. Methodological Evaluation of Observational Research (MORE) – observational studies of incidence or prevalence of chronic diseases

Quality criteria	Descriptor	Reporting Quality/ Methodological Quality- Presence of Major and Minor Flaws
Descriptive information about the study	Article identification number Journal of publication Year of publication Country	
Funding of study	Industry if funded by one or more corporate sponsors; Grant if funded from one or more not-for-profit sponsors; Combined industry + Grant if funded from one or more corporate sponsors and one or more not-for-profit sponsors	
Role of funding organization in data analysis and interpretations of the results	Sponsor participation in data analysis and interpretation of the results	
Conflict of interest	Disclosure of conflict of interest (at least one author)	
Ethical approval of the study	Approval of the study by ethical committees	
Aim of study	Incidence of prevalence estimation in the general population, race or ethnic, gender or sex, or other defined population subgroups by demographic, biological, health, socio-economic status, or other characteristics	Minor flaw if target population was not well defined
Study design	Cross-sectional Retrospective Prospective	
External Validity		
Sampling the subjects		
Sampling subjects from the general population	Random population based Non-random population based Random multistage population based Random stratified population based. Random sampling restricted to geographic area	Minor flaw :sampling restricted to geographic area if the aim was to examine incidence/prevalence in the general population without place restrictions
Nongeneral population sampling method	Random Convenient Self selection	Minor flaw: Convenient or self-selection sampling methods
Nongeneral population based sampling frame	Sampling within nationally representative registries or databases Health care based, medical records Insurance claims Work place Proxy selection	Major flaws: sampling based on medical records, insurance claims, work place, health care based (clinics, hospitals) if the study aimed to estimate incidence or prevalence of chronic condition or disease in the general population
Assessment of sampling bias— failure to ensure that all members of the reference population have a known chance of selection in the sample	Possible sources for sampling bias may include: failure to adhere to the random sampling procedures; omission of specific subgroups of the population from the sampling frame and therefore from the sample; non-response to a survey by specific subgroups of the population; nonrandom exclusion the subjects from specific subgroups of the population that are relevant to the study goals and objectives.	Minor flaw if the authors did not assess sampling bias
Estimation of sampling bias	Response rate in the total sample, race, age, gender, and other subgroups. The ranges need to be justified and vary in specific research areas, should be predefined before quality evaluation	Major flaw if response rate <40% or less than acceptable in a specific subpopulation
Exclusion rate from the analysis	Exclusion rate in the total sample, race, age, gender, and other subgroups. The ranges need to be justified and vary in specific research areas, should be predefined before quality evaluation	Major flaw if more than 10% of eligible subjects were excluded from the analyses or more than acceptable in a specific subpopulation
Sampling bias is addressed in the analysis	The goal is to adjust the results for violations of the assumption that each subject has an equal probability of selection to the study. Weighting of the estimates by non-response adjustment within sampling subgroups.	Minor flaw if the authors did not reduce possible sampling bias in the analysis

	Post-stratification by age, sex, race or other variables to minimize the impact of differences in non-selection and non-response at the levels of the sampling	
Subject flow	Number of screened, eligible, and enrolled subjects. Recruitment fractions are calculated (automatic calculation in Access interface) Number needed to screen	Minor flaw if enrollment fraction is less than acceptable ranges specific for the area of research
Internal Validity		
Source of measure incidence/prevalence of chronic diseases	Self-reported (collected for the study) Proxy reported (collected for the study) 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) Obtained from administrative database (mining of the data collected for health care purposes) Obtained from registries or administrative databases (collected for epidemiologic evaluation independent of health care).	Minor flaws—self reported outcomes or mining of the data collected for health care business purposes
Definition of the outcomes		
Duration of symptoms in the definition of the outcome	Relevance of the time of occurrence for the nature of the outcome should be predefined before quality evaluation. Reference period recommended by the CDC or guidelines is 12 months for chronic diseases, reference period different from recommended should be justified.	Minor flaw if reference period may be relevant but not included in definition of the outcome or reference period different from recommended and not justified
Severity in the definition of the outcome	Relevance of the degree of the symptoms of the chronic disease for the nature of the outcome should be predefined before quality evaluation	Major flaw if severity can be relevant but not assessed in the study
Frequency of symptoms of the chronic disease	Relevance of the of the symptoms for the nature of the outcome should be predefined before quality evaluation	Major flaw if frequency can be relevant but not assessed in the study
Measurements of outcomes		
Validation of the methods to measure the outcomes	Variables can be measured using known “gold standard” the method considered by the consensus of the experts to be the best available method for establishing the presence or absence of the condition of interest. The study can validate the methods to measure outcomes with “gold standard” or with other methods when the gold standard is not available.	Major flaw if nonvalid methods were obtained to measure the outcomes. Minor flaw if the study reported inter-methods validation (one method vs. another) when gold standard is available
Reliability of the estimates	Intra-observer variability or inter-observer variability can be within acceptable for the outcome standards that should be predefined before quality evaluation. The study can use the methods to measure the outcomes with reliability that was assumed acceptable according to previous published analyses	Minor flaw if intra-observer or inter-observer variability are reported with subjective judgment of reliability and not acceptable according to the nature of the outcomes
Outcomes in race, ethnic, age, or gender subpopulations	The study should use the same methods to measure the outcome in the total sample and in the subgroups.	Minor flaw if outcomes in subpopulations were measured differently. Major flaw if the study aimed to estimate incidence or prevalence in specific subpopulations but assessment of the outcomes was invalid or unreliable
Reporting of outcomes: type of outcome	Period prevalence Point prevalence Incidence rate	Minor flaw if point prevalence was reported
Precision of estimate	Mean and variance of incidence or prevalence estimates should be reported (error, 95% CI)	
Estimate in total sample	Population estimates of incidence or prevalence should be age adjusted, prevalence or incidence can be standardized by age and gender to the standard population	Minor flaw if crude estimates only were provided
Estimate in population subgroups (age, gender, race, other subgroups)	Subpopulation estimates of incidence or prevalence in gender, race, or other subgroups should be age adjusted, prevalence or incidence can be standardized by age and gender to the standard population	Minor flaw if crude estimates only were provided

The checklist identifies valid measures used to diagnose the chronic diseases as important considerations of internal validity. For example, prevalence of clinical manifestation of genital herpes was less than 10% while seroprevalence is dramatically larger at 20% in the adults in the United States.[51] The definition and prevalence of urinary incontinence varied widely, with over 20 definitions having been used. [52] We defined a minor flaw in internal validity when the duration of symptoms differed from recommended without justification; for example, a study on chronic stable atrial fibrillation should include patients with symptoms for at least 12 months. [53] Valid outcome measurement is an essential

quality component. [11] If there is a gold standard or reference standard available, methods used in studies should ideally use these standards, or validate the methods used compared to these established standards.

The checklist was designed to evaluate quality of both the prevalence of disease in the general population as well as sub-populations in the same study. These subgroups should be identified *a priori*. We defined a minor flaw when different methods are used to measure the outcomes in the total sample versus in subgroups or the method could have different validity or reliability in subpopulations. For example, prevalence estimate differed when history of tuberculosis was self-reported in the total

sample of the general population, was obtained from x-rays of legal immigrants, or was obtained from voluntarily performed x-rays in a subpopulation of illegal immigrants. [54,55,56] We defined a flaw in internal validity when crude estimates of incidence or prevalence are provided in race, gender, or other subgroups.

In the absence of a gold standard, a formal test for criterion validity was not feasible. Testers noted that complete quality assessment was time-consuming. Poor quality studies with major flaws required more time to assess quality than well designed studies. However, while time-consuming, comprehensive assessment of the risk of bias of a study is an essential element to evidence-based research. Identification of pre-defined stopping rules improved the efficiency of the quality review process by identifying major flaws for which a study may be triaged due to low quality. We proposed using the developed checklist in systematic reviews of non-therapeutic studies with predefined in the protocols topic-specific methodological standards and essential reliability testing [36].

4. Discussion

As a result of our collaborative effort we develop and validated a checklist for comprehensive quality evaluation of observational non-therapeutic studies. In contrast with previously available scales [20] or checklists [24] our tool discriminates reporting vs. methodological quality and external vs. internal validity. Previously published systematic review of non-therapeutic observational studies used different tools for quality appraisal since the authors found no single tool applicable for their research questions. [20,21,22,25,27] Our tool was already utilized in several published systematic reviews [57-64].

Recent publications of the systematic reviews using the developed checklist demonstrated the importance of predefined research specific quality standards [57-65]. We believe that with predefined research specific quality standards in review protocols our tool is applicable for all topics concerned with incidence or prevalence of chronic conditions.

The researchers continue developing new checklists for observational studies that examined prevalence of specific diseases because quality standards differ for various diseases and chronic conditions.[66,67] We have argued that our proposed generic measures can be adapted to various diseases [36,38] We proposed first to achieve consensus around universal flaws and then a priori defined disease-specific flaws with regard to external and internal validity [36].

Our work has policy implications. Evidence based decisions in public health and clinical settings should be made based on comprehensive literature reviews [1]. The Institute of Medicine developed standards for comprehensive evidence reviews [1]. Through quality appraisal of the primary studies contributed to the reviews is critical part when providing valid evidence for decision makers [1]. We propose using the developed checklist to appraise quality of the studies of incidence or prevalence of chronic diseases in systematic reviews of such studies. 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 the evaluations. All protocols of systematic reviews should be registered in the international prospective register of systematic review protocols in health and social care [69,70]. Systematic reviews should incorporate quality of the studies into the synthesis of evidence to estimate to what extent quality was associated with the results of the primary studies and conclusions of the review [33].

The evaluation of the level of evidence from several observational nontherapeutic studies was beyond our present goals and should be conducted in the future. Future research should also establish the best practices incorporating quality of the primary studies into the synthesis of evidence and actionable guideline recommendations [71].

Acknowledgements

We would like to thank our reviewers David Atkins, MD, John Hoey, MD, and Christine Laine, MD, for reviewing and commenting on the draft. We also want to thank the librarian, Judith Stanke, for her contributions to the literature search; research assistants Stacy Dickinson, MPH, Emily Zabor candidate for MS in biostatistics, and Akweley Ablorh, candidate for MS in biostatistics, for the data abstraction, quality control, and synthesis of evidence; Zhihua Bian candidate for MS in biostatistics, for her statistical help; Zhiyuan Xu, candidate for MS 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, PhD, for her help estimating sample size; Jeannine Ouellette for her help in writing the manuscript, 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.

Statement of Interest

The authors have no competing interests.

References

- [1] Institute of Medicine (U.S.). *Finding What Works in Health Care: Standards for Systematic Reviews*. Heidelberg, Neckar: National Academies Press; 2011.
- [2] Bero LA, Jadad AR. How consumers and policymakers can use systematic reviews for decision making. *Annals of internal medicine*. Jul 1 1997;127(1):37-42.
- [3] Briss PA, Brownson RC, Fielding JE, Zaza S. Developing and using the Guide to Community Preventive Services: lessons learned about evidence-based public health. *Annual review of public health*. 2004;25:281-302.
- [4] 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. *American journal of preventive medicine*. Jan 2000;18(1 Suppl):35-43.
- [5] Chan KS, Morton SC, Shekelle PG. Systematic reviews for evidence-based management: how to find them and what to do with them. *The American journal of managed care*. Nov 2004;10(11 Pt 1):806-812.

- [6] National Center for Chronic Disease Prevention and Health Promotion. Chronic Disease Overview. <http://www.cdc.gov/NCCDphp/overview.htm>. 2013; Accessed August 2013.
- [7] Fox DM. Evidence of evidence-based health policy: the politics of systematic reviews in coverage decisions. *Health affairs*. Jan-Feb 2005;24(1):114-122.
- [8] Lavis J, Davies H, Oxman A, Denis JL, Golden-Biddle K, Ferlie E. Towards systematic reviews that inform health care management and policy-making. *Journal of health services research & policy*. Jul 2005;10 Suppl 1:35-48.
- [9] Chapman DP, Williams SM, Strine TW, Anda RF, Moore MJ. Dementia and its implications for public health. *Preventing chronic disease*. Apr 2006;3(2):A34.
- [10] Launer LJ, Andersen K, Dewey ME, et al. Rates and risk factors for dementia and Alzheimer's disease: results from EURODEM pooled analyses. EURODEM Incidence Research Group and Work Groups. European Studies of Dementia. *Neurology*. Jan 1 1999;52(1):78-84.
- [11] Aschengrau A SG. Essentials of Epidemiology in Public Health. Sudbury, Mass. 2003; Jones and Bartlett.
- [12] West S, King V, Carey TS, et al. Systems to rate the strength of scientific evidence. *Evidence report/technology assessment*. Mar 2002(47):1-11.
- [13] Higgins JP, Altman DG, Gotzsche PC, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *Bmj*. 2011;343:d5928.
- [14] Viswanathan M, Ansari MT, Berkman ND, et al. Assessing the Risk of Bias of Individual Studies in Systematic Reviews of Health Care Interventions. *Methods Guide for Effectiveness and Comparative Effectiveness Reviews*. Rockville (MD)2008.
- [15] Viswanathan M, Berkman ND. Development of the RTI item bank on risk of bias and precision of observational studies. *J Clin Epidemiol*. Feb 2012;65(2):163-178.
- [16] Shamliyan T, Kane RL, Dickinson S. A systematic review of tools used to assess the quality of observational studies that examine incidence or prevalence and risk factors for diseases. *J Clin Epidemiol*. Oct 2010;63(10):1061-1070.
- [17] Shamliyan T, Kane RL, Jansen S. Quality of systematic reviews of observational nontherapeutic studies. *Preventing chronic disease*. Nov 2010;7(6):A133.
- [18] Shamliyan T, Kane RL, Jansen S. Systematic reviews synthesized evidence without consistent quality assessment of primary studies examining epidemiology of chronic diseases. *J Clin Epidemiol*. Jun 2012;65(6):610-618.
- [19] Centers for Disease Control and Prevention (U.S.). Principles of epidemiology in public health practice : an introduction to applied epidemiology and biostatistics. Atlanta, GA: U.S. Dept. of Health and Human Services, Centers for Disease Control and Prevention (CDC), Office of Workforce and Career Development. 2006;3rd ed.
- [20] Loney PL, Chambers LW, Bennett KJ, Roberts JG, Stratford PW. Critical appraisal of the health research literature: prevalence or incidence of a health problem. *Chronic diseases in Canada*. 1998;19(4):170-176.
- [21] Woodbury MG, Houghton PE. Prevalence of pressure ulcers in Canadian healthcare settings. *Ostomy/wound management*. Oct 2004;50(10):22-24, 26, 28, 30, 32, 34, 36-28.
- [22] Macfarlane TV, Glenny AM, Worthington HV. Systematic review of population-based epidemiological studies of oro-facial pain. *Journal of dentistry*. Sep 2001;29(7):451-467.
- [23] Lundh A, Gotzsche PC. Recommendations by Cochrane Review Groups for assessment of the risk of bias in studies. *BMC medical research methodology*. 2008;8:22.
- [24] DuRant RH. Checklist for the evaluation of research articles. *The Journal of adolescent health : official publication of the Society for Adolescent Medicine*. Jan 1994;15(1):4-8.
- [25] 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. *Journal of epidemiology and community health*. Jun 1998;52(6):377-384.
- [26] Kaplin AL, Williams M. How common are the "common" neurologic disorders? *Neurology*. Jul 24 2007;69(4):410; author reply 410-411.
- [27] Hirtz D, Thurman DJ, Gwinn-Hardy K, Mohamed M, Chaudhuri AR, Zalutsky R. How common are the "common" neurologic disorders? *Neurology*. Jan 30 2007;68(5):326-337.
- [28] Feinstein AR. Clinometrics. *New Haven: Yale University Press*. 1987.
- [29] Sanderson S, Tatt ID, Higgins JP. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology: a systematic review and annotated bibliography. *International journal of epidemiology*. Jun 2007;36(3):666-676.
- [30] Hulley SB. Designing clinical research: an epidemiologic approach. Philadelphia: Lippincott Williams & Wilkins. 2001;2nd ed.
- [31] BMJ Publishing Group Limited. Clinical evidence. *clinicalevidence.bmj.com*. 2013; Accessed in August 2013.
- [32] National Library of Medicine (U.S.). NIOHUS. PubMed Central. Bethesda, MD. 2013.
- [33] Higgins J GS. Cochrane handbook for systematic reviews of interventions. *Cochrane Collaboration*. 2011; Chichester, West Sussex (Hoboken NJ: John Wiley & Sons).
- [34] 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 services research*. Mar 23 2005;5(1):25.
- [35] 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 services research*. Dec 22 2004;4(1):38.
- [36] Shamliyan TA, Kane RL, Ansari MT, et al. Development quality criteria to evaluate nontherapeutic studies of incidence, prevalence, or risk factors of chronic diseases: pilot study of new checklists. *J Clin Epidemiol*. Jun 2011;64(6):637-657.
- [37] Efron B TR. An introduction to the bootstrap. *New York: Chapman & Hall*. 1993.
- [38] R: A language and environment for statistical computing. Foundation for Statistical Computing. Vienna, Austria. 2013.
- [39] 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.
- [40] Gwet K. Inter-rater reliability: dependency on trait prevalence and marginal homogeneity. *Statistical Methods for Inter-Rater Reliability Assessment Series*. 2002;2:1-9.
- [41] Gwet K. Computing inter-rater reliability with the SAS system. *Stat Methods Inter-rater Reliability Assess*. 2002;3:1-16.
- [42] Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. *The British journal of mathematical and statistical psychology*. May 2008;61(Pt 1):29-48.
- [43] Denny CH, Holtzman D, Goins RT, Croft JB. Disparities in chronic disease risk factors and health status between American Indian/Alaska Native and White elders: findings from a telephone survey, 2001 and 2002. *American journal of public health*. May 2005;95(5):825-827.
- [44] Collins JG. Prevalence of selected chronic conditions: United States, 1990-1992. *Vital and health statistics. Series 10, Data from the National Health Survey*. Jan 1997(194):1-89.
- [45] Blumberg SJ, Welch EM, Chowdhury SR, Upchurch HL, Parker EK, Skalland BJ. Design and operation of the National Survey of Children with Special Health Care Needs, 2005-2006. *Vital and health statistics. Ser. 1, Programs and collection procedures*. Dec 2008(45):1-188.
- [46] Mack KA, Ahluwalia IB. Observations from the CDC: Monitoring women's health in the United States: selected chronic disease indicators, 1991-2001 BRFSS. *Journal of women's health*. May 2003;12(4):309-314.
- [47] Wilt TJ, Shamliyan T, Taylor B, et al. Comparative Effectiveness of Therapies for Clinically Localized Prostate Cancer. *AHRQ Comparative Effectiveness Reviews*. 2008.
- [48] Nelson R, Norton N, Cautley E, Furner S. Community-based prevalence of anal incontinence. *JAMA : the journal of the American Medical Association*. Aug 16 1995;274(7):559-561.
- [49] Teunissen TA, van den Bosch WJ, van den Hoogen HJ, Lagro-Janssen AL. Prevalence of urinary, fecal and double incontinence in the elderly living at home. *International urogynecology journal and pelvic floor dysfunction*. Jan-Feb 2004;15(1):10-13; discussion 13.
- [50] Johanson JF, Lafferty J. Epidemiology of fecal incontinence: the silent affliction. *The American journal of gastroenterology*. Jan 1996;91(1):33-36.
- [51] Fleming DT, McQuillan GM, Johnson RE, et al. Herpes simplex virus type 2 in the United States, 1976 to 1994. *The New England journal of medicine*. Oct 16 1997;337(16):1105-1111.

- [52] Shamliyan T, Wyman J, Bliss DZ, Kane RL, Wilt TJ. Prevention of urinary and fecal incontinence in adults. *Evidence report/technology assessment*. Dec 2007(161):1-379.
- [53] Jacob K, Talwar S, Copplestone A, Gilbert TJ, Haywood GA. Activation of coagulation occurs after electrical cardioversion in patients with chronic atrial fibrillation despite optimal anticoagulation with warfarin. *International journal of cardiology*. May 2004;95(1):83-88.
- [54] U.S. Congress OoTA. The Continuing Challenge of Tuberculosis. Washington, DC: U.S. Government Printing Office. 1993;OTA-H-574(<http://ota-cdn.fas.org/reports/9347.pdf>):Accessed in August 2013.
- [55] Nagelkerke NJ, Borgdorff MW, Kalisvaart NA, Broekmans JF. The design of multi-stage tuberculin surveys: some suggestions for sampling. *The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease*. Apr 2000;4(4):314-320.
- [56] Davidow AL, Katz D, Reves R, Bethel J, Ngong L, Tuberculosis Epidemiologic Studies C. The challenge of multisite epidemiologic studies in diverse populations: design and implementation of a 22-site study of tuberculosis in foreign-born people. *Public health reports*. May-Jun 2009;124(3):391-399.
- [57] Shaghagh A, Matlabi H. Reporting of Health Promotion Research: Addressing the Quality Gaps in Iran. *Health Promotion*. 2012;2(1):48-52.
- [58] Kleijn SA, Aly MF, Knol DL, et al. A meta-analysis of left ventricular dyssynchrony assessment and prediction of response to cardiac resynchronization therapy by three-dimensional echocardiography. *European Heart Journal—Cardiovascular Imaging*. 2012;13(9):763-775.
- [59] Malboosbaf R, Hosseinpanah F, Mojarrad M, Jambarsang S, Azizi F. Relationship between goiter and gender: a systematic review and meta-analysis. *Endocrine*. 2012:1-9.
- [60] Slattery J, Morgan A, Douglas J. Early sucking and swallowing problems as predictors of neurodevelopmental outcome in children with neonatal brain injury: a systematic review. *Developmental Medicine & Child Neurology*. 2012;54(9):796-806.
- [61] Robroek SJ, Reeuwijk KG, Hillier FC, Bamba CL, van Rijn RM, Burdorf A. The contribution of overweight, obesity, and lack of physical activity to exit from paid employment: a meta-analysis. *Scandinavian journal of work, environment & health*. 2013;39(3):233-240.
- [62] Manfredini D, Restrepo C, Diaz-Serrano K, Winocur E, Lobbezoo F. Prevalence of sleep bruxism in children: a systematic review of the literature. *Journal of oral rehabilitation*. 2013.
- [63] Khalesi M, Whiteman DC, Doi SA, Clark J, Kimlin MG, Neale RE. Cutaneous markers of photo-damage and risk of basal cell carcinoma of the skin: A meta-analysis. *Cancer Epidemiology Biomarkers & Prevention*. 2013.
- [64] Edmondson D, Richardson S, Fausett JK, Falzon L, Howard VJ, Kronish IM. Prevalence of PTSD in Survivors of Stroke and Transient Ischemic Attack: A Meta-Analytic Review. *PLOS ONE*. 2013;8(6):e66435.
- [65] Sale JE, Beaton D, Posen J, Bogoch E. Medication initiation rates are not directly comparable across secondary fracture prevention programs: reporting standards based on a systematic review. *Journal of clinical epidemiology*. 2013;66(4):379-385. e374.
- [66] Hoy D, Brooks P, Woolf A, et al. Assessing risk of bias in prevalence studies: modification of an existing tool and evidence of interrater agreement. *J Clin Epidemiol*. Sep 2012;65(9):934-939.
- [67] Groenwold RH, Rovers MM. The Catch-22 of appraisals on the quality of observational studies. *J Clin Epidemiol*. Oct 2010;63(10):1059-1060.
- [68] Kane RL, Shamliyan T. Be specific and dare to generalize: do we need a rating form for every disease? *J Clin Epidemiol*. Sep 2012;65(9):921-923.
- [69] Booth A, Clarke M, Dooley G, et al. PROSPERO at one year: an evaluation of its utility. *Systematic reviews*. 2013;2:4.
- [70] Van der Wees P, Qaseem A, Kaila M, Ollenschlaeger G, Rosenfeld R, Board of Trustees of the Guidelines International N. Prospective systematic review registration: perspective from the Guidelines International Network (G-I-N). *Systematic reviews*. 2012;1:3.
- [71] Qaseem A, Forland F, Macbeth F, et al. Guidelines International Network: toward international standards for clinical practice guidelines. *Annals of internal medicine*. Apr 3 2012;156(7):525-531.