Research White Paper

Assessing the Predictive Validity of Strength of Evidence Grades: A Meta-Epidemiological Study

DEPARTMENT OF HEALTH & HUMAN SERVICES • USA

AHRQ

Agency for Healthcare Research and Quality

Advancing Excellence in Health Care • **www.ahrq.gov**

Research White Paper

Assessing the Predictive Validity of Strength of Evidence Grades: A Meta-Epidemiological Study

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. 290-2012-00008-I

Prepared by:
RTI International–University of North Carolina Evidence-based Practice Center
Research Triangle Park, NC

Investigators:
Gerald Gartlehner, M.D., M.P.H.[1,2]
Andreea Dobrescu, M.D.[3]
Tammeka Swinson Evans, M.O.P.[2]
Carla Bann, Ph.D.[2]
Karen A. Robinson, M.Sc., Ph.D. [4]
James Reston, Ph.D., M.P.H.[5]
Kylie Thaler, M.D., M.P.H[1]
Andrea Skelly, Ph.D., M.P.H., [6]
Anna Glechner, M.D.[1]
Kimberly Peterson, M.S.[6]
Christina Kien, M.A.[1]
Kathleen N. Lohr, Ph.D., M.Phil., M.A.[2]

[1] Cochrane Austria, Danube University, Krems, Austria
[2] RTI International, Research Triangle Park, NC
[3] Victor Babes University of Medicine and Pharmacy, Timisoara, Romania
[4] John Hopkins University Medical Center, Baltimore, MD
[5] ECRI Institute, Plymouth Meeting, PA
[6] Oregon Health & Science University, Portland, OR

AHRQ Publication No. 15-EHC028-EF
September 2015

This report is based on research conducted by the RTI International–University of North Carolina Evidence-based Practice Center (EPC) under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No290-2012-00008-I). The findings and conclusions in this document are those of the authors, who are responsible for its contents; the findings and conclusions do not necessarily represent the views of AHRQ. Therefore, no statement in this report should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services.

Drs. Gartlehner and Thaler are members of the GRADE Working Group. Drs. Gartlehner, Lohr, and Reston are co-authors of the AHRQ guidance for grading the strength of evidence. The other authors have no disclosures to report.

Suggested citation:
Gartlehner G, Dobrescu A, Swinson Evans T, Bann C, Robinson KA, Reston J, Thaler K, Skelly A, Glechner A, Peterson K, Kien C, Lohr KN. Assessing the Predictive Validity of Strength of Evidence Grades: A Meta-Epidemiological Study, Research White Paper. (Prepared by the RTI International–University of North Carolina Evidence-based Practice Center under Contract No. 290-2012-00008-I.) AHRQ Publication No. 15-EHC028-EF. Rockville, MD: Agency for Healthcare Research and Quality. September 2015. www.effectivehealthcare.ahrq.gov/reports/final.cfm.

Preface

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

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

AHRQ expects that the EPC evidence reports and technology assessments will inform individual health plans, providers, and purchasers, as well as the health care system as a whole by providing important information to help improve health care quality.

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

Richard G. Kronick, Ph.D.
Director
Agency for Healthcare Research and Quality

Arlene S. Bierman, M.D., M.S
Director
Center for Evidence and Practice Improvement
Agency for Healthcare Research and Quality

Stephanie Chang, M.D., M.P.H.
Director, Evidence-based Practice Program
Task Order Officer
Center for Evidence and Practice Improvement
Agency for Healthcare Research and Quality

Acknowledgments

The authors gratefully acknowledge the continuing support of our AHRQ Task Order Officer, Stephanie Chang, M.D., M.P.H. We would also like to thank Donna Dryden, Ph.D., University of Canada, Alberta; Marian McDonagh, PharmD., Oregon Health & Science University; Margaret Maglione, M.P.P., RAND Corporation; Barbara Nußbaumer, M.Sc., Danube University, Austria; and Jonathan Treadwell, Ph.D., ECRI Institute, for their contributions to this study. In addition, we thankfully acknowledge Evelyn Auer for administrative support and Loraine Monroe for document preparation support.

Peer Reviewers

Prior to publication of the white paper, we sought input from independent Peer Reviewers without financial conflicts of interest. However, the conclusions and synthesis of the scientific literature presented in this report does not necessarily represent the views of individual reviewers.

Peer Reviewers must disclose any financial conflicts of interest greater than $10,000 and any other relevant business or professional conflicts of interest. Because of their unique clinical or content expertise, individuals with potential nonfinancial conflicts may be retained. The Task Order Officer and the EPC work to balance, manage, or mitigate any potential nonfinancial conflicts of interest identified

The list of Peer Reviewers follows:

Mohammed T. Ansari, M.B.B.S., M.Med.Sc., M.Phil.
Clinical Meta-Research Methodologist and EBM Consultant
Adjunct Professor, University of Ottawa
Ottawa, Canada

Rongwei (Rochelle) Fu, Ph.D.
Pacific Northwest Evidence-based Practice Center
Oregon Health & Science University
Portland, OR

Jennifer Graff, PharmD.
National Pharmaceutical Council
Washington, DC

Gordon Guyatt, B.Sc., M.D. M.Sc., FRCPC, OC
Department of Clinical Epidemiology & Biostatistics
McMaster University
Hamilton, Ontario, Canada

Susan L. Norris, M.D., M.P.H.
World Health Organization
Geneva, Switzerland

Evelyn Whitlock, M.D., M.P.H
Kaiser Permanente Research Affiliates Evidence-based Practice Center
Portland, OR

Assessing the Predictive Validity of Strength of Evidence Grades: A Meta-Epidemiological Study

Structured Abstract

Objective. We sought to determine the predictive validity of the U.S. Evidence-based Practice Center (EPC) approach to GRADE (Grading of Recommendations Assessment, Development and Evaluation) by examining how reliably it can predict the likelihood that treatment effects remain stable as new studies emerge.

Study design and setting. Based on 37 Cochrane reports with outcomes graded as high strength of evidence (SOE), we prepared 160 documents using portions of these bodies of evidence in a chronological order. We randomly assigned these documents, which represented different levels of SOE, to professional systematic reviewers from seven academic centers in Austria, Canada, and the United States, who dually graded the SOE using guidance for the EPC program. For each of the 160 documents, we determined whether estimates remained stable as subsequent studies were added to the evidence base. For each grade of SOE, we compared the observed proportion of stable estimates with the expected proportion from an international survey. To determine the predictive validity, we used the Hosmer-Lemeshow test to assess calibration and the C (concordance) index to assess discrimination.

Results. Overall, the predictive validity of the EPC approach to GRADE for the stability of effect estimates was limited. Except for moderate SOE, the expected and observed proportions of stable effect estimates differed considerably. Estimates graded as high SOE were less likely to remain stable than expected by producers and users of systematic reviews. By contrast, estimates graded as low or insufficient SOE were substantially more likely to remain stable than expected. In this sample, the EPC approach to GRADE could not reliably predict the likelihood that individual bodies of evidence remain stable as new evidence becomes available. Depending on the definition used, C-indices ranged between 0.56 (95% CI, 0.47 to 0.66) and 0.58 (95% CI, 0.50 to 0.67) indicating a low discriminatory ability.

Conclusion. The limited predictive validity of the EPC approach to GRADE seems to reflect a mismatch between expected and observed changes in treatment effects as bodies of evidence advance from insufficient to high SOE. In addition, many low or insufficient grades appear to be too strict.

Contents

Introduction

Despite the enormous amount of new information that medical research generates every year, uncertainty plays a major role in health care decisionmaking. The challenging task for clinical and health policy decisionmakers is to balance considerations about evidence, values, preferences, and resources, all of which are often fraught with uncertainty and conflicting perspectives.[1]

GRADE (Grading of Recommendations Assessment, Development and Evaluation) has evolved as a widely used approach to communicate certainties and uncertainties in systematic reviews to readers and other stakeholders.[2,3] GRADE uses information about risk of bias, imprecision, inconsistency, indirectness, and reporting bias to categorize the degree of uncertainty concerning the correctness of findings into four grades of quality of evidence.

The Evidence-based Practice Center (EPC) program of the U.S. Agency for Healthcare Research and Quality (AHRQ) has made small adaptations to the GRADE system to meet its specific needs.[4,5] Guidance for EPCs refers to quality of evidence as "strength of evidence"(SOE) and defines it as the degree of confidence that estimates are close to the true effect and the likelihood that findings will remain stable over time (i.e., the likelihood that future studies will not have an important impact on the estimate of an effect).[4] In this paper, we refer generically to the GRADE approach but, as necessary for clarity, specify that specific points or findings refer to just the EPC approach. Table 1 summarizes the EPC definitions of the four levels of SOE.

Decisionmakers who rely on the GRADE approach assume that estimates of effect that are graded as high SOE are "close to the true effect" and, therefore, will remain stable as new evidence emerges. By contrast, decisionmakers can interpret effect estimates that are graded as low SOE as quite likely to change as new evidence accrues. In a recent international survey, we determined that producers and users of systematic reviews associated each grade of SOE with a distinct likelihood that estimates of effect will remain stable as new evidence emerges (see Table 1).[6]

Table 1. Definitions of grades of strength of evidence from the EPC program guidance

Grade	Definition	Expected Proportions of Stable Effect Estimates[a]
High	**We are very confident that the estimate of effect lies close to the *true effect* for this outcome.** The body of evidence has few or no deficiencies. We believe that the findings are stable, i.e., that another study would not change the conclusions.	86% to 100%
Moderate	**We are moderately confident that the estimate of effect lies close to the *true effect* for this outcome.** The body of evidence has some deficiencies. We believe that the findings are likely to be stable, but some doubt remains.	61% to 85%
Low	**We have limited confidence that the estimate of effect lies close to the *true effect* for this outcome.** The body of evidence has major or numerous deficiencies (or both). We believe that additional evidence is needed before concluding either that the findings are stable or that the estimate of effect is close to the *true effect*.	34% to 60%
Insufficient[b]	**We have no evidence, we are unable to estimate an effect,** or we have no confidence in the estimate of effect for this outcome. No evidence is available or the body of evidence has unacceptable deficiencies, precluding reaching a conclusion.	0% to 33%

[a] Expected proportions are based on an international survey of producers and users of systematic reviews.[6]
[b] The EPC category of insufficient also includes outcomes without evidence. For the purpose of this study, we did not consider situations without any evidence whatsoever.

To date, the predictive validity of the GRADE approach concerning the stability of effect estimates has not been tested. Predictive validity, in general terms, refers to the degree to which a score (such as the grades cited in Table 1) predicts an outcome on a criterion measure.[7] For this analysis, predictive validity refers to the degree to which this approach, and specifically different SOE grades, reliably predicts the stability of an estimate of effect because it is close to the true effect.

A true effect can be viewed as the effect size that we would observe if a study had an infinitely large sample size (and thus no sampling error).[8] Realistically, however, a true treatment effect can rarely be determined and used as a reference standard. For that reason, here we equate true effect with stability of effect as new studies emerge, a concept that can be measured. Given accurate predictive validity, a rating of “high SOE” would reliably predict that future studies will have a minor impact on the estimate of effect of a given outcome. Likewise, a rating of “low SOE” would reliably predict a high likelihood that future studies will have a substantial impact on the direction or magnitude of the estimate of effect of a given outcome.

The objective of our study was to determine the predictive validity of the EPC approach to GRADE based on a diverse sample of interventions. That is, we examined how reliably it can predict the likelihood that treatment effects remain stable.

Methods

We used a meta-epidemiological approach based on large, systematically appraised bodies of evidence that authors of Cochrane reports had graded as high SOE. We used effect estimates of such bodies of evidence as reference points because a grade of high SOE implies that investigators were very confident that the estimate of effect is close to the truth and that new studies are unlikely to change conclusions. Thus, we used these estimates as “gold standards” to determine the predictive validity. We did not assess the correctness of SOE grades in the Cochrane reports because we wanted to take a pragmatic perspective using real-world examples rather than an explanatory perspective using an ideal dataset. We assumed that users of systematic reviews would also take grades of SOE at face value.

Assembling Empirical Data

We searched the Cochrane Library from 2010 onward to find Cochrane reports that: (1) include an outcome with more than eight randomized controlled trials (RCTs) on therapeutic interventions that had been graded as high SOE; (2) present meta-analytic outcomes that were reported as relative risks or odds ratios for binary outcomes or as weighted mean differences or standardized mean differences (SMDs) for continuous outcomes; and (3) provide data to reproduce the meta-analyses. We chose a threshold of eight RCTs so that we had enough studies to meta-analyze portions of these bodies of evidence in a chronological order of publication.

Overall, we drew information from 37 Cochrane reports on 50 bodies of evidence that had been graded as high SOE. Table 2 presents characteristics of these bodies of evidence.

Table 2. Cochrane reports and characteristics of high-strength bodies of evidence used to prepare summary documents

Cochrane Report	Intervention and Outcome	Number of Participants	Effect Estimate (Confidence Interval)
Amato et al., 2010[9]	Benzodiazepines and adverse events	471	RR: 1.50 (0.83 to 2.70)
Amato et al., 2010[9]	Benzodiazepines and dropouts	839	RR: 1.10 (0.75 to 1.63)
Amato et al., 2011[10]	Psychosocial maintenance intervention and retention in treatment	2,582	RR: 1.02 (0.97 to 1.07)
Amato et al., 2013[11]	Tapered methadone and completion of treatment	1,381	RR: 1.08 (0.97 to 1.21)
Buchleitner et al., 2012[12]	Perioperative glycemic control and mortality	1,365	RR: 1.19 (0.89 to 1.59)
Chauhan et al., 2014[13]	Long-acting beta agonists and exacerbations	6,257	RR: 0.87 (0.76 to 0.99)
Chin et al., 2013[14]	Infraclavicular block and adequate surgical anesthesia	1,051	RR: 0.88 (0.51 to 1.52)
Chin et al., 2013[14]	Infraclavicular block and tourniquet pain	615	RR: 0.66 (0.47 to 0.92)
Chin et al., 2013[14]	Infraclavicular block and need for supplemental local anesthetic blocks or systemic analgesia	1,412	RR: 0.95 (0.62 to 1.46)
Chong et al., 2013[15]	Phosphodieserase-4-inhibitors and exacerbations	15,035	OR: 0.77 (0.71 to 0.83)
Chong et al., 2013[15]	Phosphodieserase-4-inhibitors and gastrointestinal side effects	15,241	OR: 3.07 (2.66 to 3.53)
Clifford et al., 2012[16]	Autologous adult stem cells and left-ventricular ejection fraction	879	WMD: 1.78 (0.27 to 3.28)
Feagan et al., 2012[17]	Oral 5-aminosalicylic acid and failure to maintain remission	1,298	RR: 0.69 (0.62 to 0.77)
Fernandes et al., 2013[18]	Systemic or inhaled glucocorticoids and rate of hospital admission	1,762	RR: 0.92 (0.78 to 1.08)

Table 2. Cochrane reports and characteristics of high-strength bodies of evidence used to prepare summary documents (continued)

Cochrane Report	Intervention and Outcome	Number of Participants	Effect Estimate (Confidence Interval)
Fernandes et al., 2013[18]	Systemic or inhaled glucocorticoids and length of hospital stay	633	WMD: -0.18 (-0.39 to 0.04)
Gafter et al., 2012[19]	Antibiotic prophylaxis and mortality	5,635	RR: 0.66 (0.55 to 0.79)
Gowing et al., 2009[20]	Buprenorphine and completion of withdrawal treatment	1,206	RR: 1.64 (1.31 to 2.06)
Griffiths et al., 2013[21]	Inhaled anticholinergic drugs and hospital admission	2,497	RR: 0.73 (0.63 to 0.85)
Gurion et al., 2012[22]	Colony stimulating factors and mortality	3,405	RR: 1.03 (0.99 to 1.07)
Hauser et al., 2013[23]	Serotonin and noradrenaline reuptake inhibitors and 50% pain reduction in fibromyalgia	5,994	RR: 1.49 (1.35 to 1.64)
Hauser et al., 2013[23]	Serotonin and noradrenaline reuptake inhibitors and withdrawals due to adverse events	6,179	RR: 1.83 (1.53 to 2.18)
Hemmingsen et al., 2013[24]	Intensive glycemic control and hypoglycemia	28,127	RR: 2.05 (1.39 to 3.02)
Hodson et al., 2013[25]	Antiviral prophylaxis and cytomegalovirus disease (patients with organ transplants)	1,132	RR: 0.42 (0.32 to 0.57)
Hodson et al., 2013[25]	Antiviral prophylaxis and cytomegalovirus disease (all treated patients)	1,981	RR: 0.42 (0.34 to 0.52)
Howe et al., 2011[26]	Exercise and change in bone mineral density	1,441	WMD: 0.08 (-1.08to 0.92)
Katalinic et al., 2010[27]	Stretch interventions and joint mobility	221	WMD: 1.00 (0.00 to 3.00)
Lai et al., 2013[28]	Antimicrobial impregnation, coating, or bonding and mortality	2,371	RR: 0.88 (0.75 to 1.05)
Lai et al., 2013[28]	Antimicrobial impregnation, coating, or bonding and adverse effects	3,003	RR: 1.09 (0.94 to 1.27)
Law et al., 2013[29]	Sumatriptan plus naproxen and pain free at 2 hours	3,370	RR: 2.76 (2.43 to 3.13)
Law et al., 2013[29]	Sumatriptan plus naproxen and pain free at 24 hours	3,396	RR: 3.04 (2.59 to 3.56)
Lemiengre et al., 2012[30]	Antibiotics and cure from rhinosinusitis	1,687	OR: 1.25 (1.02 to 1.53)
Lemiengre et al., 2012[30]	Antibiotics and treatment failure	2,175	OR: 0.49 (0.36 to 0.66)
Lewis et al., 2013[31]	Nonsteroidal anti-inflammatory drugs and vomiting	1,021	RR: 0.72 (0.61 to 0.85)
Liakopoulos et al., 2012[32]	Statins and atrial fibrillation	841	OR: 0.55 (0.44 to 0.69)
Liakopoulos et al., 2012[32]	Statins and length of stay in hospital	877	RR: -0.48 (-0.85 to -0.11)
Main et al., 2013[33]	Hormone therapy and stroke	33,197	RR: 1.26 (1.11 to 1.43)
Moja et al., 2012[34]	Trastuzumab and congestive heart failure	10,281	RR: 5.11 (3 to 8.72)
Musini Vijaya et al., 2009[35]	Pharmacotherapy and cardiovascular morbidity and mortality	23,094	RR: 0.72 (0.68 to 0.77)
Nannini et al., 2013[36]	Long-acting beta2-agonist+inhaled corticosteroid and mortality	7,518	OR: 0.78 (0.64 to 0.94)
Nelson et al., 2012[37]	Surgical therapy of anal fissure and healing	979	OR: 0.11 (0.06 to 0.23)
Nüesch et al., 2010[38]	Opioids and withdrawal because of adverse events	2,403	RR: 4.05 (3.06 to 5.38)
Pandian et al., 2013[39]	Double embryo transfer and live birth rate	1,564	OR: 2.07 (1.68 to 2.57)
Pandian et al., 2013[39]	Double embryo transfer and multiple pregnancy rate	1,612	OR: 8.47 (4.97 to 14.43)
Pani et al., 2011[40]	Antidepressant medication and alcohol abstinence	942	RR: 1.22 (0.99 to 1.51)
Paul et al., 2013[41]	Antibiotic therapy and death in cancer patients with neutropenia	7,186	RR: 0.87 (0.75 to 1.02)

Table 2. Cochrane reports and characteristics of high-strength bodies of evidence used to prepare summary documents (continued)

Cochrane Report	Intervention and Outcome	Number of Participants	Effect Estimate (Confidence Interval)
Paul et al., 2013[41]	Antibiotic therapy and nephrotoxicity in cancer patients with neutropenia	6,608	RR: 0.45 (0.35 to 0.57)
Perez et al., 2009[42]	Angiotensin converting enzyme inhibitors and mortality	84,311	RR: 0.93 (0.87 to 0.98)
Perez et al., 2009[42]	Beta-blockers and mortality	71,457	RR: 0.96 (0.91 to 1.02)
Rehman et al., 2011[43]	Traditional suburethral sling procedures and incontinence	693	RR: 0.97 (0.78 to 1.2)
Wilhelmus et al., 2010[44]	Antiviral therapies and healing of herpes simplex virus keratitis	401	RR: 1.96 (1.67 to 2.31)

OR = odds ratio; RR = relative risk; SMD = standardized mean differences, WMD = weighted mean differences

Preparing "Gradeable" Documents

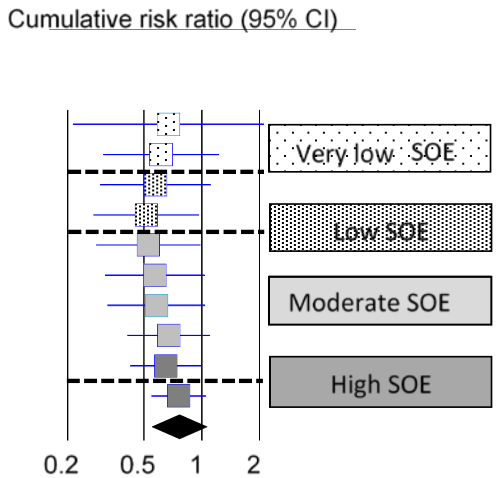
From each of the 50 included bodies of evidence, we used portions in a chronological order of publication to prepare a total of 160 documents (which we called "gradeable" documents) reflecting different SOE categories. Sample size calculations indicated that 130 documents would provide 80 percent power for a 4 x 2 chi-square test of SOE (high, medium, low, or insufficient) by stability of results (stable vs. not stable) for a medium-sized effect (Cohen's d of 0.3) as a threshold for stability.

In a first step, we reanalyzed each body of evidence using cumulative meta-analyses. In general, a cumulative meta-analysis shows how the body of evidence evolves over time as new studies accrue. Likewise, the SOE changes (or can be expected to change) over time as new studies contribute to the body of evidence. Using information from the cumulative meta-analyses and information about individual studies from the Cochrane reports (e.g., risk of bias ratings), an independent investigator (who was not involved in the subsequent grading of the SOE) meta-analyzed the portions of the high-strength bodies of evidence in a chronological order (e.g., the first four studies, the first six studies, etc.) to prepare the gradeable documents.

Figure 1 illustrates this concept. The investigator took risk of bias of individual studies, precision of estimates, consistency of studies, indirectness, and the other domains of the grading scheme into consideration to decide what portions of studies were used for the gradeable documents.

Figure 1. Illustration of the concept of using portions of studies to create bodies of evidence to grade

Study name	Cumulative statistics		
	Point	Lower limit	Upper limit
Study 1	0.67	0.21	2.13
Study 2	0.61	0.30	1.24
Study 3	0.57	0.29	1.12
Study 4	0.51	0.27	0.98
Study 5	0.53	0.28	0.99
Study 6	0.57	0.31	1.04
Study 7	0.58	0.32	1.05
Study 8	0.67	0.41	1.12
Study 9	0.65	0.42	1.01
Study 10	0.76	0.55	1.06
	0.76	0.55	1.06



CI = confidence interval; SOE = strength of evidence

The aim was to create approximately 40 documents for each category of SOE with sufficient information for the project's investigators to grade the SOE. These documents included: information on the objective of the Cochrane review; the PICO (population-intervention-control-outcome); study characteristics and risk of bias ratings of included trials as presented in the Cochrane report; a forest plot of a random effects meta-analysis; information about minimal important differences for continuous outcomes; and information about reporting bias (funnel plot, Kendell's tau, Egger's regression intercept, and Fail-Safe N). We relied on judgments of the Cochrane authors regarding risk of bias of individual trials. We pilot-tested the format and content of the gradeable documents and revised them based on feedback from investigators. Appendix A provides an example of a gradeable document.

Grading Strength of Evidence

To grade the SOE, investigators used EPC guidance for GRADE. Investigators took part in a calibration exercise and had access to a published guidance document.[4]

We randomly allocated 160 gradeable documents to 13 investigators from six U.S. and Canadian EPCs and Cochrane Austria. All are professional systematic reviewers; however, their experience with GRADE varied. Three investigators (23 percent) stated that they had used the GRADE approach for more than 20 systematic reviews; three (23 percent) used the approach for 10 to 15 systematic reviews; one (8 percent) used the approach for 6 to 10 reviews; and six investigators (46 percent) declared that they had used GRADE for up to 5 systematic reviews.

A research associate at RTI International connected each participant with a unique identification number and emailed the gradeable documents. This research associate was not involved in either the grading exercise or analysis of results. Two investigators, blinded to the results of the underlying Cochrane report (i.e., the reference standard), graded each body of evidence independently. Investigators were blinded to the second person grading the same body of evidence. When grades differed, the research associate put investigators in contact with each other; investigators resolved conflicts by consensus or by involving a third, senior researcher.

Assessing the Stability of Effect Estimates

To determine the stability of effects, we compared effect estimates of the gradeable documents with the high SOE estimates from the Cochrane reports (the gold standard). To do so, we modified an approach developed to detect signals for updating systematic reviews.[45] We used three definitions of stability (Table 3), which differed in the thresholds that determined whether the magnitude of treatment effects was similar. We deemed an estimate of effect as stable when (1) statistical significance did not change *and (*2) the magnitude of treatment effects remained similar to the high SOE estimate of the Cochrane report.

Table 3. Three definitions of stability of effect based on change in statistical significance and magnitude of effect

Stability of Effect: Definition 1 (Strict Definition)	
Change in statistical significance	Statistical significance does not change between graded effect and gold standard effect (changes within the range of p-values 0.04 to 0.06 are not counted as change).
Change in magnitude of effect	Difference in magnitude of effects is smaller than a relative risk change (increase or reduction) of 25 percentage points for dichotomous outcomes or 0.20 SMDs for continuous outcomes.
Stability of Effect: Definition 2 (Lenient Definition)	
Change in statistical significance	Same as definition 1.
Change in magnitude of effect	Difference in magnitude of effects is smaller than a relative risk change of 50 percentage points for dichotomous outcomes or 0.50 SMDs for continuous outcomes.
Stability of Effect: Definition 3 (Staggered Definition)	
Change in statistical significance	Same as definition 1.
Change in magnitude of effect	• For graded effects with small treatment effects (relative risk 0.5 to 2.00, or SMD <0.8): Same as definition 1. • For graded estimates with large treatment effects (relative risk <0.5 and >2.00, or SMD >0.8): Same as definition 2. • For outcomes that can be considered extremely patient-relevant (e.g., mortality, stroke, myocardial infarction): Difference in magnitude of effects is smaller than relative risk change of less than 10 percentage points.

SMDs = standardized mean differences.

To avoid counting trivial or 'borderline' changes in statistical significance, we required that at least one of the two results had had a p-value outside the range of 0.04 to 0.06. In other words, we did not consider cases in which a p-value changed statistical significance within this range. For example, neither a change from p=0.041 to p=0.059 nor a change from p=0.059 to p=0.041 counted as a change in statistical significance.

Conducting Statistical Analysis

To assess the inter-rater reliability of reviewers grading the SOE, we calculated intra-class correlations using a one-way random effects model. Intra-class correlations measure the consistency of agreement of reviewers when dually grading bodies of evidence.

To determine the predictive validity, we compared the expected proportion of stable effect estimates (presented in Table 1) with the observed proportion of stable effect estimates for different thresholds from our sample. Statistically, predictive validity can be determined by calculating two characteristics: (1) calibration and (2) discrimination. Calibration refers to the ability to estimate correctly the likelihood of a future event. In our study, calibration is the ability to determine the likelihood that estimates remain stable. Discrimination refers to the ability to differentiate between those that will experience a future event and those that will not. In our study, discrimination is the ability to differentiate between effect estimates that will remain stable and those that will substantially change.[46]

We determined the calibration of the EPC approach to GRADE with the Hosmer-Lemeshow test[47] and its discrimination with the concordance (C) index. Bodies of evidence that remain stable should have higher expected likelihoods than those that do not. The C index compares the expected likelihoods from pairs of observations. In this case, the term “pairs” refers to stable versus not stable effect estimates, as shown below:[48]

$$C\ index = \frac{\#\ of\ concordant\ pairs + \frac{1}{2}\ (\#\ of\ tied\ pairs)}{Total\ \#\ of\ pairs}$$

Concordant pairs are pairs for which the expected likelihood for the stable body of evidence is higher than the expected likelihood for the nonstable body of evidence. *Tied pairs* are pairs for which the stable and nonstable bodies of evidence have the same expected likelihood. Higher values for the C index indicate better discrimination. A C index of 0.50 would indicate no discrimination between stable and nonstable bodies of evidence. We conducted all statistical analyses with the rcorr.cens procedure in the Hmisc package in R[49] or Microsoft Excel.

Results

Of 160 bodies of evidence, researchers dually graded 11 percent (n=17) as high, 42 percent (n=68) as moderate, 32 percent (n=51) as low, and 15 percent (n=24) as insufficient (very low) SOE. The inter-rater reliability was 0.56 (95% CI, 0.40 to 0.68), suggesting moderate agreement of researchers assigning SOE grades.

Concordance Between Expected and Observed Proportions of Stable Effect Estimates

For each grade, we compared the expected proportions of stable effect estimates with the observed proportion from our sample, using three different definitions of stability (see Methods and Table 2). Table 1 gave the proportions of estimates that producers and users of systematic reviews expected to remain stable for each SOE grade.

Overall, except for moderate SOE, the stability differed considerably between expected and observed proportions regardless of the definition used. *Fewer* estimates graded as high SOE in our sample remained stable relative to the expectations of producers and users of systematic reviews; that is, in our survey 208 experts expected high SOE outcomes to remain stable in at least 86 percent of the cases.[6] In our sample, the observed proportions of stable estimates for definitions 1, 2, and 3 were, respectively, 71 percent, 76 percent, and 76 percent. Conversely, substantially *more* low or insufficient SOE estimates than expected remained stable. Table 4 presents expected and observed proportions of stable effect estimates by grade of SOE for each of the three definitions of stability.

Table 4. Comparison of expected with observed proportions of stable effect estimates for different definitions of stability

Grade	Number of Effect Estimates	Expected Proportions (%)[a]	Observed Proportions (%) Definition 1 (95% CI)	Observed Proportions Definition 2 (95% CI)	Observed Proportions Definition 3 (95% CI)
High	17	86-100	71 (43 to 88)	76 (48 to 92)	76 (48 to 92)
Moderate	68	61-85	71 (58 to 80)	75 (63 to 84)	72 (59 to 91)
Low	51	34-60	55 (41 to 68)	73[b] (58 to 83)	59 (44 to 72)
Insufficient (very low)	24	0-33	54[b] (33 to 74)	58[b] (37 to 77)	58[b] (37 to 77)

CI = confidence interval
[a] Expected proportions are based on an international survey of producers and users of systematic reviews.[6]
[b] Statistically significantly different from the upper bound of expected stability.

Figures 2, 3, and 4 illustrate the overlap of expected proportions of stable effects (black large boxes) and confidence intervals (CI) of observed proportions (grey columns) for different grades of SOE and different definitions of stability. The circles in the columns reflect the point estimates. The y-axis delineates the proportion of estimates that remained stable; the x-axis presents the four grades of SOE. For insufficient SOE, for example, producers and users of systematic reviews expected 0 percent to 33 percent of estimates to remain stable as new studies are added to the evidence base. For definition 1, which was the most rigorous of the three definitions of stability, more than half (54 percent) of effect estimates graded as insufficient remained stable. The CIs ranged from 33 percent to 74 percent, which barely overlaps the expected range for insufficient SOE. For the less rigorous definitions 2 and 3, CIs did not overlap at all with the range that producers and users of systematic reviews expected from insufficient SOE grades. By contrast, observed proportions of stable results for moderate SOE grades were concordant for all three definitions. Confidence intervals overlap widely with the range of expected proportions. Estimates graded as low SOE show some concordance for definitions 1 and 3 but little for definition 2.

Figure 2. Comparison of expected proportions of stable effect estimates with confidence intervals of observed proportions for different definitions of stability—Definition 1

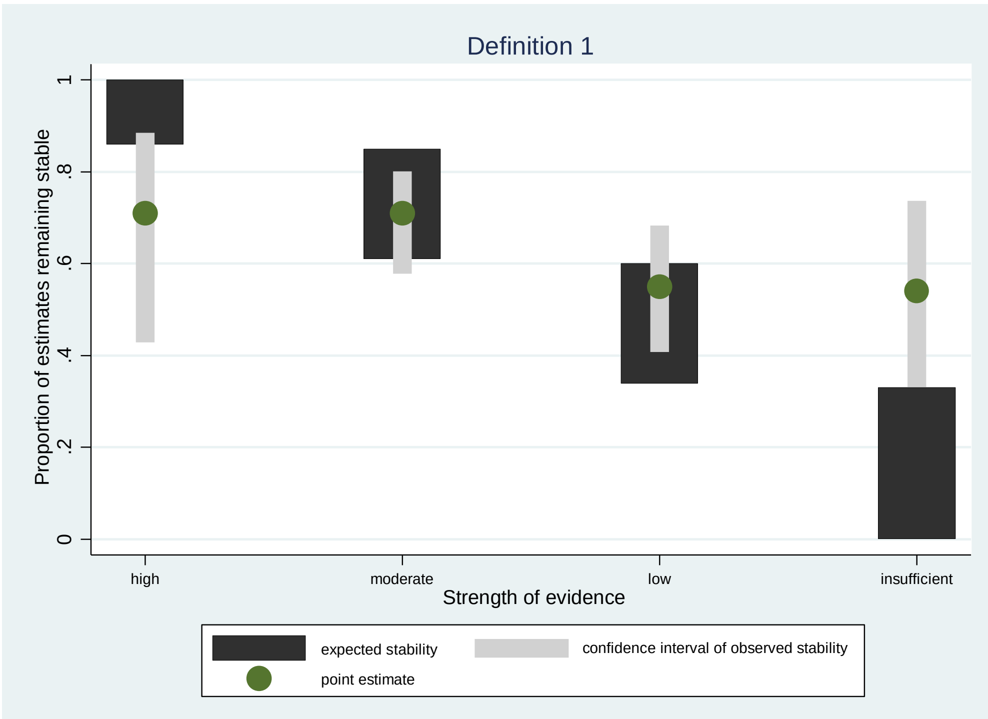


Figure 3. Comparison of expected proportions of stable effect estimates with confidence intervals of observed proportions for different definitions of stability—Definition 2

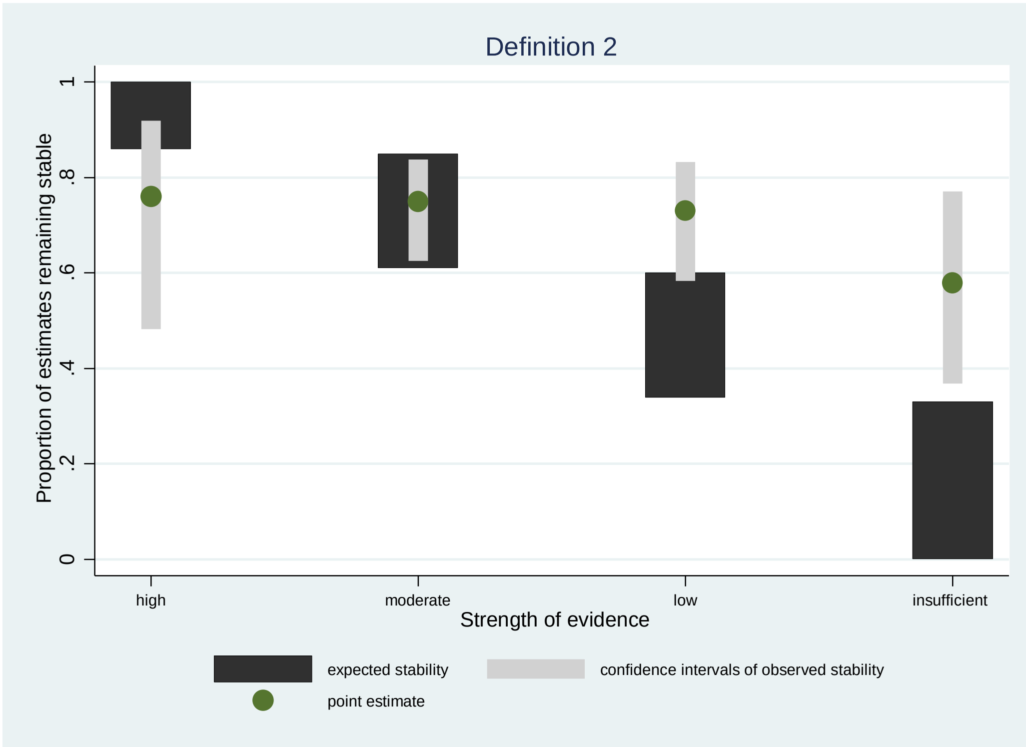
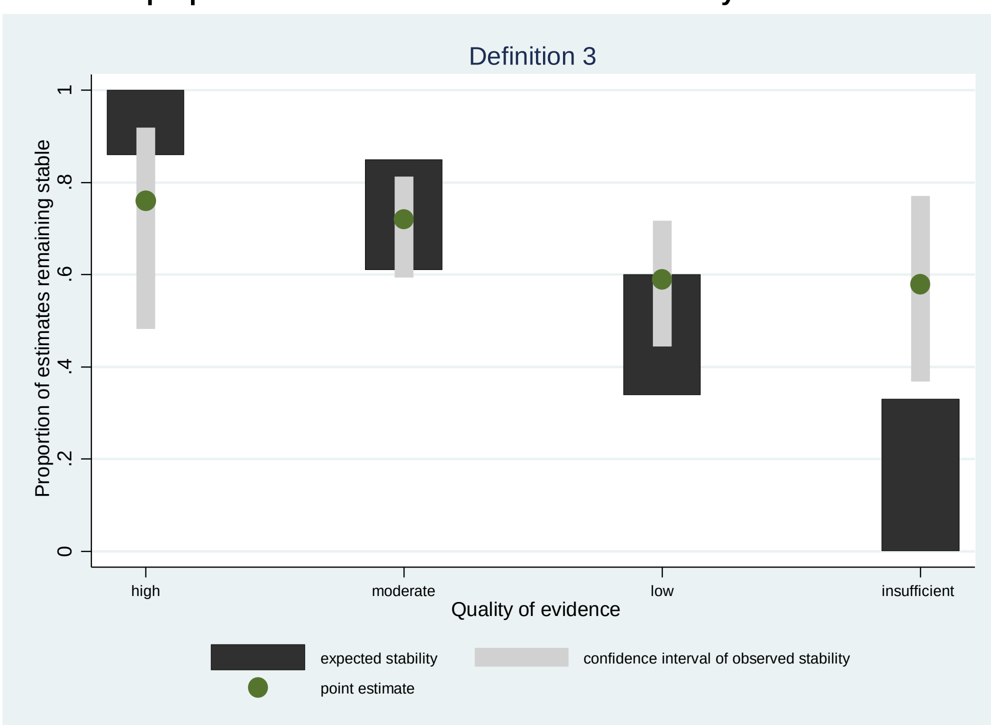


Figure 4. Comparison of expected proportions of stable effect estimates with confidence intervals of observed proportions for different definitions of stability—Definition 3



Predictive Validity of the EPC Approach to GRADE

To determine the predictive validity of the EPC approach to GRADE, we assessed the calibration (how accurately it can predict the likelihood that effect estimates will remain stable as new evidence evolves) and the discrimination (how accurately it can differentiate between effect estimates that will remain stable and those that will substantially change). In theory, an ideal predictive tool would reliably identify estimates with a high likelihood of remaining stable and always grade them as high SOE. Conversely, effect estimates with a very low likelihood of remaining stable would always be graded as insufficient. Such an ideal tool would have high calibration and a C index of 1.

Overall, regardless of the definition used, the calibration of the EPC approach to GRADE was suboptimal. When we compared observed proportions of stable effect estimates with lower, middle, and upper values of the ranges of expected proportions, eight of nine comparisons were statistically significantly different based on the Hosmer-Lemeshow test (Table 5), indicating a lack of calibration.

Table 5. Results of Hosmer-Lemeshow tests for different expected and observed proportions of stability

Levels of Expected Proportions	Observed Stability: Definition 1 (p-value)	Observed Stability: Definition 2 (p-value)	Observed Stability: Definition 3 (p-value)
Expected proportions upper-ranges (high 99%, moderate 68%, low 60%, very low 33%)	0,0661[a]	0,0305	0,0473
Expected proportions mid-ranges (high 93%, moderate 73%, low 47%, very low 17%)	<0.001	<0.001	<0.001
Expected proportions lower-ranges (high 86%, moderate 61%, low 31%, very low 1%)	<0.001	<0.001	<0.001
Expected proportions using best fitting values (high 86%, moderate 71%, low 60%, very low 33%)	0,1448[a]	0,0421	0,0925[a]

* Lack of statistical significance indicates satisfactory calibration.

Likewise, the C indices for the EPC approach to GRADE were low, with values close to that expected by chance (i.e., C index=0.50). For definitions 1, 2, and 3, the C indices were 0.57 (95% CI, 0.50 to 0.67), 0.56 (95% CI, 0.47 to 0.66), and 0.58 (95% CI, 0.50 to 0.67), respectively. C indices for definitions 1 and 3 reached statistical significance (CIs did not cross 0.5). Taking the uncertainty of the confidence intervals into consideration, results mean that in the worst case (lower limit of CIs), the EPC approach to GRADE has no discriminatory ability for distinguishing between effect estimates with a low or high likelihood of remaining stable. In the best case (upper confidence limits), it can accurately distinguish between effect estimates with a low or high likelihood of remaining stable in 67 percent of cases.

The low overall predictive validity, however, is caused primarily by the discordance of expected and observed proportions of stable effect estimates for high and insufficient SOE. In a post-hoc sensitivity analysis, we chose proportions within the expected ranges (Table 1) that were closest to the observed proportions of stable effect estimates. Using expected proportions of 86 percent for high (lower end of expected range), 71 percent for moderate, 60 percent for low, and 33 percent for insufficient SOE (both upper end of expected range), we found that the EPC approach to GRADE achieved satisfactory calibration for definitions 1 and 3 (Table 5).

Discussion

To our knowledge, our study was the first attempt to determine the predictive validity of the GRADE approach. To be considered useful in practice, any tool that conveys certainties and uncertainties of estimates of effect should have a high ability to discriminate between estimates that will remain stable in the future and those that will substantially change; it should also be able to associate respective likelihoods of stability with an expected outcome. Our research indicates that the EPC approach to GRADE only partly fulfilled these qualities of predictive validity: Only moderate SOE had satisfactory predictive validity. In the following sections, we discuss possible reasons for these findings and potential starting points for improving the predictive validity.

A predictive model, in general, is a mathematical equation describing the relationship between a prognostic marker (here, a grade of SOE) and a given outcome (stability of effect estimates).[48] In our study, three main factors determined the predictive validity of the EPC approach to GRADE:

1. The definition of stability,
2. The likelihood of expected stability associated with each grade of SOE (the prognostic marker), and
3. The operationalization of the prognostic tool (the EPC approach to GRADE) to achieve the most appropriate prognostic marker (i.e., the grade of SOE).

With respect to the first factor, the definition of stability, our study showed that strict or lenient definitions of stability had minimal impact on the predictive validity of the EPC approach to GRADE. Therefore, the other two factors appear to be the reasons for the low predictive validity and could serve as starting points for future improvements.

To determine the proportion of stable estimates that users and producers of systematic reviews associate with each grades of SOE, we recently conducted an international survey that we used as the basis of the comparison between expected and observed proportions of stable results.[6] The rationale for applying our survey results was that users of systematic reviews make decisions based on their individual interpretations of definitions of grades of SOE. If individual interpretations substantially over- or under-estimate the actual stability of effect estimates, decisions based on systematic reviews could be misguided.

Our findings indicate that, except for moderate SOE, the expectations of survey participants did not match results from our sample. Expectations were too optimistic for high SOE and too pessimistic for low and insufficient SOE. Current definitions of different grades of SOE, however, employ vague terminology to forecast certainty—such as “likely,” “very likely,” or “may be substantially different”—and this practice might contribute to the low predictive validity. Psychological research has demonstrated that perceptions of certainty can vary substantially among individuals, and that interpretation of qualitative certainty expressions also differ depending on the context in which they are used and on baseline event rates. Adding numerical predictions such as likelihoods to the definitions of the individual grades of SOE seems to be one solution that could reduce unwarranted variation in interpretations.

Finally, the EPC approach to GRADE, or the way systematic reviewers operationalize it, appears to be too strict. More than half of estimates graded as insufficient (defined as “we have no confidence in the estimate of effect for this outcome”) remained stable; this indicates that the approach too often leads to low or insufficient grades of SOE. Possible reasons could be: (a) systematic reviewers use GRADE too mechanistically, (b) recommended thresholds for downgrading in guidance documents are too strict, or (c) a tool with four levels of SOE is not granular enough to categorize uncertainty. Adding a fifth category—e.g., by using GRADE *very low* for bodies of evidence in which systematic reviewers still have some (albeit little) confidence and EPC's *insufficient* for bodies of evidence that have truly unacceptable deficiencies that preclude reaching a conclusion—would allow for more granularity.

Our study has several limitations. First, we relied on risk of bias assessments and SOE grades of Cochrane authors. Because author groups differed across these systematic reviews, heterogeneity in approaches and varying adherence to guidance documents regarding SOE grades is likely. For example, in about 20 percent of our sample (presented in Table 1), confidence intervals cross both the line of no effect and thresholds of appreciable benefits and harms. According to GRADE and EPC guidance, such a situation would require reviewers to grade down for imprecision. Nevertheless, we deliberately did not reassess SOE grades because we wanted to take a real-world, pragmatic perspective with our assessment of predictive validity. We assumed that most guideline developers or other decisionmakers who use Cochrane reports to support decisions also would not reassess SOE and would take respective grades at face value. In addition, Cochrane reports go through rigorous international peer review, and the methodological quality is usually high.

Second, how representative our sample is remains unclear. Because we wanted to use a reference standard for which researchers had high confidence that effect estimates are correct (close to the true effect), we focused on high SOE evidence. A remaining question is whether our findings are generalizable to bodies of evidence that will never progress to high SOE. In addition, our sample was limited to RCTs, so findings are likely not generalizable to research based on nonrandomized studies.

Third, systematic reviewers grading the SOE had access to guidance documents but they did not use a formal instrument such as the GRADEpro Guideline Development Tool (www.guidelinedevelopment.org) to navigate the grading exercises in a standardized manner. Using such a tool could increase inter-rater reliability and might reduce the number of grades of SOE that are too strict.[50] In situations with conflicting grades, strong personalities (maybe with a tendency to strict grades) often dominate the consensus process. Increasing inter-rater reliability would reduce the number of situations that require systematic reviewers to reach a consensus.

Over the past decade, GRADE has evolved as a widely used approach to convey the certainties and uncertainties inherent in research. Its conceptual framework uses information about factors that most researchers would intuitively consider when assessing the confidence in findings based on a body of evidence. Compared with other approaches, GRADE has clear advantages because it makes decisions about the SOE transparent and explicit.[51]

The lack of predictive validity, therefore, is probably not grounded in the concept of GRADE but rather in the way the instrument is operationalized, which, overall, appears to be too strict. The GRADE Working Group, as well as organizations such as AHRQ EPCs, need to reflect on how to reduce unwarranted variation in the interpretation of the definitions of individual grades of SOE and how to avoid overly strict grades.

Future research needs to confirm or refute our findings and explore which domains may lead to excessively strict operationalization and influence the predictive validity of the GRADE approach. Future research also needs to test whether its predictive validity is satisfactory under ideal circumstances using bodies of evidence as reference standards that are undisputed in their high SOE grade and also have a low statistical likelihood of changing as future studies accrue. Such research also needs to examine whether more (or fewer) than four grades would better capture and communicate uncertainty of research findings.

Given the moderate inter-rater reliability in our study, the EPC guidance document on grading the strength of evidence might need to be revised to provide clearer advice on how to grade the individual domains. Qualitative research can explore areas that reviewers struggle with the most when applying the GRADE approach.

References

1. Atkins D, Siegel J, Slutsky J. Making policy when the evidence is in dispute. Health Affairs (Millwood). 2005 Jan-Feb;24(1):102-13. Epub: 2005/01/14. PMID: 15647220.
2. Schunemann HJ, Hill SR, Kakad M, et al. Transparent development of the WHO rapid advice guidelines. PLoS Med. 2007 May;4(5):e119. PMID: 17535099.
3. Qaseem A, Forland F, Macbeth F, et al. Guidelines international network: toward international standards for clinical practice guidelines. Ann Intern Med. 2012 Apr 3;156(7):525-31. Epub: 2012/04/05. PMID: 22473437.
4. Berkman ND, Lohr KN, Ansari M, et al. Grading the Strength of a Body of Evidence When Assessing Health Care Interventions for the Effective Health Care Program of the Agency for Healthcare Research and Quality: An Update. Methods Guide for Comparative Effectiveness Reviews (Prepared by the RTI-UNC Evidence-based Practice Center under Contract No. 290-2007-10056-I). AHRQ Publication No. 13(14)-EHC130-EF. Rockville, MD: Agency for Healthcare Research and Quality. November 2013. www.effectivehealthcare.ahrq.gov/reports/final.cfm.
5. Owens DK, Lohr KN, Atkins D, et al. AHRQ series paper 5: grading the strength of a body of evidence when comparing medical interventions—Agency for Healthcare Research and Quality and the Effective Health-Care Program. J Clin Epidemiol. 2010 May;63(5):513-23. PMID: 19595577.
6. Gartlehner G, Sommer I, Evans TS, et al. Grades for quality of evidence were associated with distinct likelihoods that treatment effects will remain stable. J Clin Epidemiol. 2015 May;68(5):489-97. PMID: 25484316.
7. Cronbach LJ, Meehl PE. Construct validity in psychological tests. Psychological Bulletin. 1955 Jul;52(4):281-302. PMID: 13245896.
8. Borenstein M, Hedges LV, Higgins JPT, et al. Introduction to Meta-Analysis. John Wiley & Sons Ltd; 2009.
9. Amato L, Minozzi S, Vecchi S, et al. Benzodiazepines for alcohol withdrawal. Cochrane Database Syst Rev. 2010(3):CD005063. PMID: 20238336.
10. Amato L, Minozzi S, Davoli M, et al. Psychosocial combined with agonist maintenance treatments versus agonist maintenance treatments alone for treatment of opioid dependence. Cochrane Database Syst Rev. 2011(10):CD004147. Epub: 2011/10/07. PMID: 21975742.
11. Amato L, Davoli M, Minozzi S, et al. Methadone at tapered doses for the management of opioid withdrawal. Cochrane Database Syst Rev. 2013;2:CD003409. Epub: 2013/03/02. PMID: 23450540.
12. Buchleitner AM, Martinez-Alonso M, Hernandez M, et al. Perioperative glycaemic control for diabetic patients undergoing surgery. Cochrane Database Syst Rev. 2012;9(9):CD007315. PMID: 22972106.
13. Chauhan BF, Ducharme FM. Addition to inhaled corticosteroids of long-acting beta2-agonists versus anti-leukotrienes for chronic asthma. Cochrane Database Syst Rev. 2014;1:CD003137. Epub: 2014/01/25. PMID: 24459050.
14. Chin KJ, Alakkad H, Adhikary SD, et al. Infraclavicular brachial plexus block for regional anaesthesia of the lower arm. Cochrane Database Syst Rev. 2013;8:CD005487. Epub: 2013/08/30. PMID: 23986434.
15. Chong J, Leung B, Poole P. Phosphodiesterase 4 inhibitors for chronic obstructive pulmonary disease. Cochrane Database Syst Rev. 2013;11(11):CD002309. PMID: 24190161.
16. Clifford DM, Fisher SA, Brunskill SJ, et al. Stem cell treatment for acute myocardial infarction. Cochrane Database Syst Rev. 2012;2(2):CD006536. PMID: 22336818.

17. Feagan BG, Macdonald JK. Oral 5-aminosalicylic acid for maintenance of remission in ulcerative colitis. Cochrane Database Syst Rev. 2012;10(10): CD000544. PMID: 23076890.

18. Fernandes RM, Bialy LM, Vandermeer B, et al. Glucocorticoids for acute viral bronchiolitis in infants and young children. Cochrane Database Syst Rev. 2013;6:CD004878. Epub: 2013/06/05. PMID: 23733383.

19. Gafter-Gvili A, Fraser A, Paul M, et al. Antibiotic prophylaxis for bacterial infections in afebrile neutropenic patients following chemotherapy. Cochrane Database Syst Rev. 2012;1:CD004386. Epub: 2012/01/20. PMID: 22258955.

20. Gowing L, Ali R, White JM. Buprenorphine for the management of opioid withdrawal. Cochrane Database Syst Rev. 2009(3):CD002025. Epub: 2009/07/10. PMID: 19588330.

21. Griffiths B, Ducharme FM. Combined inhaled anticholinergics and short-acting beta2-agonists for initial treatment of acute asthma in children. Cochrane Database Syst Rev. 2013;8(8):CD000060. PMID: 23966133.

22. Gurion R, Belnik-Plitman Y, Gafter-Gvili A, et al. Colony-stimulating factors for prevention and treatment of infectious complications in patients with acute myelogenous leukemia. Cochrane Database Syst Rev. 2012;6(6):CD008238. PMID: 22696376.

23. Hauser W, Urrutia G, Tort S, et al. Serotonin and noradrenaline reuptake inhibitors (SNRIs) for fibromyalgia syndrome. Cochrane Database Syst Rev. 2013;1:CD010292. Epub: 2013/02/27. PMID: 23440848.

24. Hemmingsen B, Lund SS, Gluud C, et al. Targeting intensive glycaemic control versus targeting conventional glycaemic control for type 2 diabetes mellitus. Cochrane Database Syst Rev. 2013;11(11):CD008143. PMID: 24214280.

25. Hodson EM, Ladhani M, Webster AC, et al. Antiviral medications for preventing cytomegalovirus disease in solid organ transplant recipients. Cochrane Database Syst Rev. 2013;2(2):CD003774. PMID: 23450543.

26. Howe TE, Shea B, Dawson LJ, et al. Exercise for preventing and treating osteoporosis in postmenopausal women. Cochrane Database Syst Rev. 2011(7):CD000333. PMID: 21735380.

27. Katalinic OM, Harvey LA, Herbert RD, et al. Stretch for the treatment and prevention of contractures. Cochrane Database Syst Rev. 2010(9):CD007455. PMID: 20824861.

28. Lai NM, Chaiyakunapruk N, Lai NA, et al. Catheter impregnation, coating or bonding for reducing central venous catheter-related infections in adults. Cochrane Database Syst Rev. 2013;6(6):CD007878. PMID: 23740696.

29. Law S, Derry S, Moore RA. Sumatriptan plus naproxen for acute migraine attacks in adults. Cochrane Database Syst Rev. 2013;10(10):CD008541. PMID: 24142431.

30. Lemiengre MB, van Driel ML, Merenstein D, et al. Antibiotics for clinically diagnosed acute rhinosinusitis in adults. Cochrane Database Syst Rev. 2012;10:CD006089. Epub: 2012/10/19. PMID: 23076918.

31. Lewis SR, Nicholson A, Cardwell ME, et al. Nonsteroidal anti-inflammatory drugs and perioperative bleeding in paediatric tonsillectomy. Cochrane Database Syst Rev. 2013;7(7):CD003591. PMID: 23881651.

32. Liakopoulos OJ, Kuhn EW, Slottosch I, et al. Preoperative statin therapy for patients undergoing cardiac surgery. Cochrane Database Syst Rev. 2012;4(4):CD008493. PMID: 22513959.

33. Main C, Knight B, Moxham T, et al. Hormone therapy for preventing cardiovascular disease in post-menopausal women. Cochrane Database Syst Rev. 2013;4(4):CD002229. PMID: 23633307.

34. Moja L, Tagliabue L, Balduzzi S, et al. Trastuzumab containing regimens for early breast cancer. Cochrane Database of Systematic Reviews. 2012(4)PMID: CD006243.

35. Musini VM, Tejani AM, Bassett K, et al. Pharmacotherapy for hypertension in the elderly. Cochrane Database Syst Rev. 2009(4):CD000028. PMID: 19821263.

36. Nannini LJ, Poole P, Milan SJ, et al. Combined corticosteroid and long-acting beta2-agonist in one inhaler versus inhaled corticosteroids alone for chronic obstructive pulmonary disease. Cochrane Database Syst Rev. 2013;8:CD006826. Epub: 2013/08/31. PMID: 23990350.

37. Nelson RL, Thomas K, Morgan J, et al. Non surgical therapy for anal fissure. Cochrane Database Syst Rev. 2012;2(2):CD003431. PMID: 22336789.

38. Nuesch E, Rutjes AW, Husni E, et al. Oral or transdermal opioids for osteoarthritis of the knee or hip. Cochrane Database Syst Rev. 2009(4):CD003115. Epub: 2009/10/13. PMID: 19821302.

39. Pandian Z, Marjoribanks J, Ozturk O, et al. Number of embryos for transfer following in vitro fertilisation or intra-cytoplasmic sperm injection. Cochrane Database Syst Rev. 2013;7(7):CD003416. PMID: 23897513.

40. Pani PP, Trogu E, Vecchi S, et al. Antidepressants for cocaine dependence and problematic cocaine use. Cochrane Database Syst Rev. 2011(12):CD002950. PMID: 22161371.

41. Paul M, Dickstein Y, Schlesinger A, et al. Beta-lactam versus beta-lactam-aminoglycoside combination therapy in cancer patients with neutropenia. Cochrane Database Syst Rev. 2013;6(6):CD003038. PMID: 23813455.

42. Perez MI, Musini VM, Wright JM. Effect of early treatment with anti-hypertensive drugs on short and long-term mortality in patients with an acute cardiovascular event. Cochrane Database Syst Rev. 2009(4):CD006743. PMID: 19821384.

43. Rehman H, Bezerra CC, Bruschini H, et al. Traditional suburethral sling operations for urinary incontinence in women. Cochrane Database Syst Rev. 2011(1):CD001754. Epub: 2011/01/21. PMID: 21249648.

44. Wilhelmus KR. Antiviral treatment and other therapeutic interventions for herpes simplex virus epithelial keratitis. Cochrane Database Syst Rev. 2010(12):CD002898. PMID: 21154352.

45. Shojania KG, Sampson M, Ansari MT, et al. Updating Systematic Reviews. Technical Review No. 16. AHRQ Publication No. 07-0087. Rockville, MD: Agency for Healthcare Research and Quality; Sep 2007. http://www.ncbi.nlm.nih.gov/pubmed/20734512.

46. McGeechan K, Macaskill P, Irwig L, et al. Assessing new biomarkers and predictive models for use in clinical practice: a clinician's guide. Arch Intern Med. 2008 Nov 24;168(21):2304-10. Epub: 2008/11/26. PMID: 19029492.

47. Hosmer DW, Lemeshow S. Goodness of fit tests for the multiple logistic regression model. Communications in Statistics - Theory and Methods. 1980;9(10):1043-69.

48. Tripepi G, Jager KJ, Dekker FW, et al. Statistical methods for the assessment of prognostic biomarkers (Part I): discrimination. Nephrol Dialysis Transplant. 2010 May;25(5):1399-401. Epub: 2010/02/09. PMID: 20139066.

49. Harrell J, F.E. Package ‘Hmisc’. 2014 http://cran.r-project.org/web/packages/Hmisc/Hmisc.pdf. Accessed December 19, 2014.

50. Mustafa RA, Santesso N, Brozek J, et al. The GRADE approach is reproducible in assessing the quality of evidence of quantitative evidence syntheses. J Clin Epidemiol. 2013 Jul;66(7):736-42 e5. PMID: 23623694.

51. Schunemann HJ, Best D, Vist G, et al. Letters, numbers, symbols and words: how to communicate grades of evidence and recommendations. CMAJ. 2003 Sep 30;169(7):677-80. PMID: 14517128.

Appendix A. Example Gradeable Document

Gradeable Document B-3-1.5-1

General Information and Instructions

The information in this document is based on a published Cochrane review. Risk of bias ratings and decisions to include specific studies in the meta-analysis were those of the Cochrane review authors. The current document summarizes:

1. the objectives of the Cochrane review,
2. the PICOs (population, intervention, comparator, outcome),
3. the risk of bias ratings for each included study,
4. the forest plot of a random effects meta-analysis (for dichotomous outcomes the effect measure is the risk ratio, for continuous outcomes the standardized mean difference), and
5. the funnel plot of the meta-analysis.
6. The appendix provides more detail on the individual studies as presented in the Cochrane report.

Please use this information and the EPC guidance to grade the strength of evidence for the presented outcome.

Objective of Review

To examine the effectiveness of exercise in preventing bone loss in postmenopausal women by determining whether or not exercise slows bone loss and has a beneficial effect on the axial (the skull, spine and rib cage) and appendicular (the bones of the limbs and pelvis) bone density in postmenopausal women.

PICOs

Population of Interest

Healthy postmenopausal women (including those with previous fractures) aged between 45 and 70 years.

Intervention

Exercise program (e.g. walking, calisthenics and resisted strengthening)

Comparator

Standard therapy (e.g. usual activity or placebo with or without pharmacological consumption).

Outcome

Bone mineral density % change: hip
Minimal important difference (MID): Authors provide no information on MID

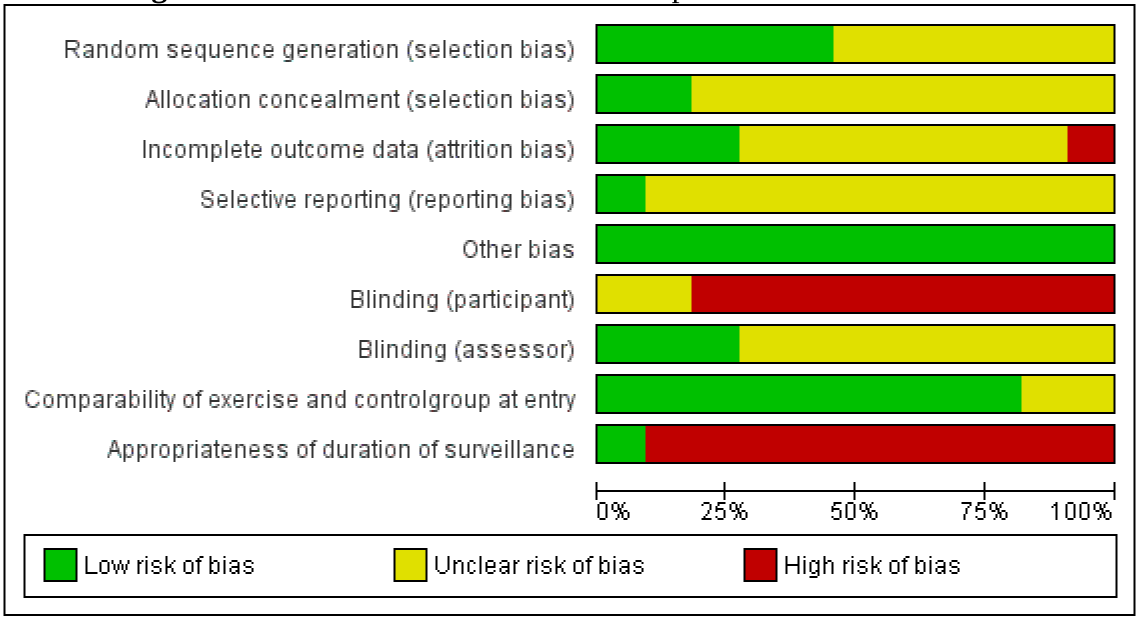
Risk of Bias of Included Studies (as assessed by authors of review)

All included studies were RCTs

	Random sequence generation (selection bias)	Allocation concealment (selection bias)	Incomplete outcome data (attrition bias)	Selective reporting (reporting bias)	Other bias	Blinding (participant)	Blinding (assessor)	Comparability of exercise and controlgroup at entry	Appropriateness of duration of surveillance
Bemben 2000	?	?	−	?	+	−	?	+	−
Bergstrom 2008	+	?	+	?	+	−	?	?	−
Cheng 2002	+	?	?	?	+	?	?	+	−
Chilibeck 2002	?	+	?	?	+	?	+	+	−
Kerr 2001	+	?	?	?	+	−	?	+	−
Korpelainen 2006	+	+	+	?	+	−	+	+	+
Maddalozzo 2007	?	?	?	?	+	−	?	+	−
Newstead 2004	?	?	?	?	+	−	?	?	−
Pruitt 1996	?	?	?	?	+	−	?	+	−
Verschueren 2004	+	?	?	?	+	−	+	+	−
Von Stengel 2009	?	?	+	+	+	−	?	+	−

low risk of bias   unclear risk of bias   high risk of bias

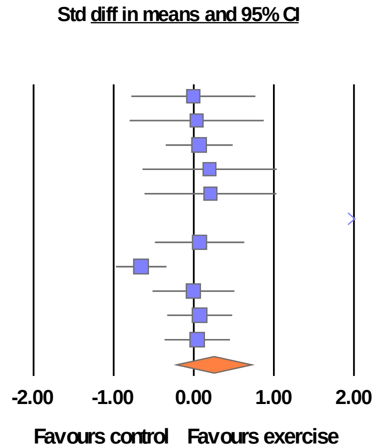
Funding of included studies: no information reported



Summary Effect of the Intervention

Random effects meta-analysis: standardized mean difference of bone mineral density % change: hip

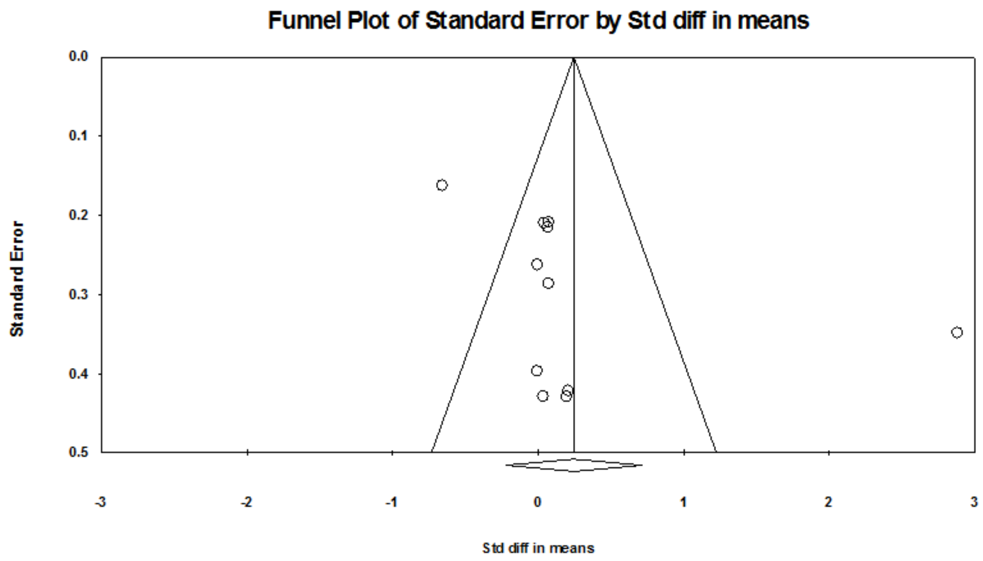
Study name	Std diff in means	Lower limit	Upper limit	Sample size: Exercise	Sample size: Control
Pruitt 1996	-0.01	-0.78	0.77	15	11
Bemben 2000	0.04	-0.80	0.88	17	8
Kerr 2001	0.07	-0.35	0.49	54	36
Chilibeck 2002	0.20	-0.64	1.04	10	12
Cheng 2002	0.21	-0.62	1.03	10	13
Verschueren 2004	2.88	2.20	3.57	47	24
Newstead 2004	0.07	-0.49	0.63	23	26
Korpelainen 2006	-0.66	-0.98	-0.34	84	76
Maddalozzo 2007	-0.00	-0.52	0.51	29	29
Bergstrom 2008	0.07	-0.34	0.48	48	44
Von Stengel 2009	0.04	-0.37	0.45	44	47
	0.24	-0.22	0.71		



I-squared 88%

Publication Bias

<u>Selective reporting:</u> Insufficient information was available to permit judgement of 'low risk' or 'high risk of bias' for selective reporting for any of the 43 studies.



Kendell's tau (with continuity correction, 2-tailed)	P-value: 0.35
Egger's regression intercept	P-value: 0.12
Fail-Safe N	Number of missing studies that would bring P-value to >0.05: 0

Details of Included Studies

Bemben 2000

Methods	Type of study: RCT
Participants	Number of participants randomised = 35 Losses: 10 (4 high repetitions, 3 high load, 3 control) Age: 41-60 years Setting: USA Inclusion:1-7 yr postmenopausal and had not performed any resistance training in the previous 6 months Exclusion: 1) diagnosed osteoporosis or a BMD site ≥ 2.5 SD below the mean for the young-adult reference population; 2) a history of cardiovascular disease; 3) physical or orthopaedic disabilities; 4) a history or current diagnosis of renal disease, chronic digestive or eating disorders, rheumatoid arthritis, or thyroid disease; 5) a history of prolonged bed rest; and 6) current or recent use of medications that affect bone density (i.e. oestrogen, steroid hormones, calcitonin or corticosteroids)
Interventions	Exercise group high load (HL) **(NWBHF)** (n = 10): 10-min warm-up, approximately 45 min of weight lifting, and ended with a 5-min cool-down. Quadriceps extension, hamstring flexion, leg press, shoulder press, biceps curl, triceps extension, seated row and latissimus pull. High load low reps (8 reps 80% 1RM) Exercise group high repetition (HR) **(NWBLF)** (n = 7): 10-min warm-up, approximately 45 min of weight lifting, and ended with a 5-min cool-down. Quadriceps extension, hamstring flexion, leg press, shoulder press, biceps curl, triceps extension, seated row and latissimus pull. Low load high reps (16 reps 40% 1RM) Control Group (n = 8): usual activity Duration and intensity: 3 sessions per week for 6 months Supervisor: Research assistants Supervision: Group Setting: Gym
Outcomes	% Change BMD spine, hip (total hip, neck of femur, trochanter, Wards triangle), total body
Notes	Compliance/adherence: average attendance for the 6-month intervention was 93% for HR and 87% for HL Adverse events: none reported Converted absolute data to % change

Risk of bias

Bias	Authors' judgement	Support for judgement
Random sequence generation (selection bias)	Unclear risk	Subjects were matched according to the BMD of the spine after baseline testing, then they were randomly assigned, method not described

Bemben 2000 *(Continued)*

Allocation concealment (selection bias)	Unclear risk	Insufficient information to permit judgement of 'high risk' or 'low risk'
Incomplete outcome data (attrition bias) All outcomes	High risk	'As-treated' analysis done, drop-outs mentioned but not accounted for in analysis
Selective reporting (reporting bias)	Unclear risk	Insufficient information to permit judgement of 'high risk' or 'low risk'
Other bias	Low risk	The study appears to be free of other sources of bias
Blinding (participant)	High risk	Not possible
Blinding (assessor)	Unclear risk	Not reported
Comparability of exercise and control group at entry	Low risk	No significant group differences existed in number of years postmenopausal or in body composition variables
Appropriateness of duration of surveillance	High risk	Only immediately postintervention data at 6 months, no follow-up data reported

Bergstrom 2008

Methods	Type of study: RCT
Participants	Number of participants randomised = 112 Losses: 20 (Exercise: 1 failed to attend DXA, 11 did not start training or trained less than 6 months, Control: 8 undertook other exercise) Age: 59.6 Exercise, 58.9 control Setting: Sweden Inclusion: postmenopausal women 45 to 65 years with forearm fractures and T-scores from −1.0 to −3.0 (total hip or spine) Exclusion: T-score lower than −3 at any site, had any disease known to interfere with bone metabolism, were on cortisone therapy or anti-resorptive medication, including hormone replacement therapy, had a BMI lower than 19.9 or higher than 30.9, or were already training at the level of or above that of the intervention
Interventions	Exercise group (**COMB**) (n = 48): 3 fast 30-minute, walks and two sessions of one-hour training per week. 5-minute warm-up, 25 minutes of strengthening exercises for the arms, legs, back and stomach, 25 minutes of aerobic exercise, and 5 minutes of stretching. Individuals chose own level and intensity and encouraged to increase level if possible Control Group (n = 44): usual activity Duration and intensity: 5 sessions per week for 12 months Supervisor: nurses Supervision: group Setting: clinic

Bergstrom 2008 *(Continued)*

Outcomes	% change BMD DEXA spine, total hip
Notes	Compliance/adherence: controlled by study nurse (compliance was 95%) Adverse events: none reported 80% power difference, 3% with 64 in each group Converted absolute data to % change

Risk of bias

Bias	Authors' judgement	Support for judgement
Random sequence generation (selection bias)	Low risk	Predefined random number table
Allocation concealment (selection bias)	Unclear risk	Insufficient information to permit judgement of 'high risk' or 'low risk'
Incomplete outcome data (attrition bias) All outcomes	Low risk	Per protocol and intention-to-treat analysis
Selective reporting (reporting bias)	Unclear risk	Insufficient information to permit judgement of 'high risk' or 'low risk'
Other bias	Low risk	The study appears to be free of other sources of bias
Blinding (participant)	High risk	Not possible
Blinding (assessor)	Unclear risk	Not reported
Comparability of exercise and control group at entry	Unclear risk	Insufficient information to permit judgement of 'high risk' or 'low risk'
Appropriateness of duration of surveillance	High risk	Only immediately postintervention data at 1 year, no follow-up data reported

Cheng 2002

Methods	Type of study: RCT
Participants	Number of participants randomised = 80 Losses: 28 (non HRT 8 exercise, 5 control; HRT 10 exercise, 5 control) Age: 50-57 years Setting: Finland Inclusion: 50 -55-year-old women, no serious cardiovascular or locomotor system problems, a body mass index of 33 kg/m^2, and not currently or previously (no longer than 6 months and at least 2 years prior to screening) using medications including oestrogen, fluoride, calcitonin, bisphosphonate's, and steroids, last menstruation at least 0.5 years but not more than 5 years ago Exclusion: not reported
Interventions	No HRT Exercise group (**DWBHF**) (n = 20): 5 circuit-training periods, each lasting 8 -10 weeks. These periods were interrupted by three high-impact aerobic dance periods, each of 2 week duration, and a summer pause for 5 weeks. Each session commenced with a 10 min warm-up period and concluded with stretching activities. During the first two circuit training periods, three rotations were performed of skipping (30 sec) , bounding over soft hurdles (13-16 cm), drop jumping (10-15 cm), and hopping (on one leg 10 times, added during the second training period). The following three periods comprised four rotations of bounding (19-25 cm), drop jumping (20-25 cm), hopping (10 times per leg) and leaping (10 times). In addition, all circuit training sessions included 3 or 4 of the following resistance exercises for the upper body: chest fly, latissimus pull down, military press, seated row and biceps curl. The home exercise programme was also designed as a circuit training routine comprising three rotations of skipping (30 sec), hopping (10 times per leg) and drop jumping (15 cm). In addition, exercises to strengthen the abdominal and lower back region were included. Average GRF was 4.3 times body weight (BW) for drop-landing from a 10 cm height, and 5.2 times BW from 20 and 25 cm heights; bounding over the hurdles 4.9-5.1 BW, skipping, hopping, and leaping 3.8, 3.4, and 4.8 BW, respectively No HRT Control Group (n = 20): usual activity HRT Exercise group (**DWBHF**) (n = 20): as exercise group above HRT Control Group (n = 20): usual activity Duration and intensity: 2 x supervised and 4 non supervised sessions per week 12 months Supervisor: not stated Supervision: group/individual Setting: gym/home
Outcomes	BMD DXA proximal femur, tibial shaft Cortical tibia

Cheng 2002 *(Continued)*

Notes	Compliance/adherence: average attendance 1 x per week Adverse events: none reported Converted absolute data to % change

Risk of bias

Bias	Authors' judgement	Support for judgement
Random sequence generation (selection bias)	Low risk	Randomisation by drawing lots
Allocation concealment (selection bias)	Unclear risk	Insufficient information to permit judgement of 'high risk' or 'low risk'
Incomplete outcome data (attrition bias) All outcomes	Unclear risk	'As-treated' analysis done drop-outs mentioned but not controlled for
Selective reporting (reporting bias)	Unclear risk	Insufficient information to permit judgement of 'high risk' or 'low risk'
Other bias	Low risk	The study appears to be free of other sources of bias
Blinding (participant)	Unclear risk	Stated double-blind (may be related to the HRT component) but insufficient information to permit judgement of 'high risk' or 'low risk'
Blinding (assessor)	Unclear risk	Stated double-blind (may be related to the HRT component) but insufficient information to permit judgement of 'high risk' or 'low risk'
Comparability of exercise and control group at entry	Low risk	No significant differences in physical characteristics at baseline
Appropriateness of duration of surveillance	High risk	Only immediately postintervention data at 12 months, no follow-up data reported

Chilibeck 2002

Methods	Type of study: RCT
Participants	Number of participants randomised = 57 Losses: 9 (4 non bisphosphonate exercise, 3 bisphosphonate exercise, 2 non bisphosphonate control) Age: mean age of groups ranged from 55.9 to 58.8 years Setting: Canada Inclusion: postmenopausal status (cessation of bleeding status for one year) Exclusion: skeletal; disorders, kidney disease or bone related disorders, chronic disease or chronic medication likely to affect metabolism or calcium imbalance. BMD z-score < -2.

	0, HRT, bisphosphonate therapy in last year, recent participation in exercise programmes, history of cardiac disease or high blood pressure
Interventions	All received 10 μg vitamin D/d and those in non bisphosphonate received 500 mg calcium carbonate/d Non bisphosphonate exercise group **(NWBHF)** (n = 10): warm up cycling and stretching, 2 sets 8-10 reps of: bench press, latissimus dorsi pull down, shoulder press, biceps curl, back extension, hip extension, flexion, adduction and abduction, knee flexion, knee extension and leg press. initially 70% 1RM then progressed Non bisphosphonate control Group (n = 12): usual activity Bisphosphonate exercise group **(NWBHF)** (n = 12): as above Bisphosphonate control group (n = 14): usual activity Duration and intensity: 3 days per week for 12 months Supervisor: not stated Supervision: individual Setting: gym
Outcomes	% change BMD spine, total hip, femoral neck, trochanter, Ward's triangle, whole body % change whole body BMC
Notes	Compliance/adherence: Non bisphosphonate exercise group 77.6%, bisphosphonate exercise group 74.8% of training sessions Adverse events: none reported 9 subjects per group would demonstrate change α of 0.05 with 80% power Converted SE to SD

Risk of bias

Bias	Authors' judgement	Support for judgement
Random sequence generation (selection bias)	Unclear risk	Randomisation mentioned but insufficient information to permit judgement of 'high risk' or 'low risk'
Allocation concealment (selection bias)	Low risk	Insufficient information although mentions double-blind
Incomplete outcome data (attrition bias) All outcomes	Unclear risk	'As-treated' analysis done drop-outs mentioned but not controlled for
Selective reporting (reporting bias)	Unclear risk	Insufficient information to permit judgement of 'high risk' or 'low risk'
Other bias	Low risk	The study appears to be free of other sources of bias
Blinding (participant)	Unclear risk	Mentions double-blind but probably relates to medication status
Blinding (assessor)	Low risk	States double-blind

Chilibeck 2002 *(Continued)*

Comparability of exercise and control group at entry	Low risk	No significant differences in characteristics at baseline
Appropriateness of duration of surveillance	High risk	Only immediately postintervention data at 12 months, no follow-up data reported

Kerr 2001

Methods	Type of study: RCT
Participants	Number of participants randomised = 126 Losses: Retention at 2 years was 71% (59% in the S group, 69% in the F group, and 83% in the C group), Age: mean 60 (6.5) years Setting: Australia Inclusion: more than 4 years past menopause and physically capable of entering exercise groups but who were not already exercising at a moderate intensity more than 2 h/week Exclusion: hormone replacement or other medications or who had diseases known to affect bone density and those who had cardiovascular, physical, or orthopedic disabilities
Interventions	All subjects given 600 mg calcium per day Exercise group **(NWBHF)** (n = 24): warm-up consisting of brisk walking and stretching. This was followed by 30 minutes of resistance weight training exercises and progressively increased the loading, wrist curl, reverse curl, biceps curl, triceps pushdown, hip flexion, hip extension, latissimus dorsi pull down, and calf raise Exercise Group **(NWBLF)** (n = 30): as above but additional stationary bicycle riding

	with minimal increase in loading Control Group (n = 36): usual activity Duration and intensity: 1 hr sessions 3 x per week 2 years Supervisor: exercise physiologists Supervision: group Setting: gym
Outcomes	BMD hip (total hip, femoral neck, trochanter, Wards triangle) , lumbar spine, and radial forearm
Notes	Compliance/adherence: Exercise compliance was very high in the first 6 months for both groups (S group, 90±12%; F group, 92± 8%) but declined from this point on. In the last 6 months of compliance was 61±23% for the S group and 67±20% for the F group. The average exercise compliance over 2 years was 74±13% in the S group and 77±14% in the F group Adverse events: none reported

Risk of bias

Bias	Authors' judgement	Support for judgement
Random sequence generation (selection bias)	Low risk	Block randomisation to one of three groups
Allocation concealment (selection bias)	Unclear risk	Insufficient information to permit judgement of 'high risk' or 'low risk'
Incomplete outcome data (attrition bias) All outcomes	Unclear risk	'As-treated' analysis done drop-outs mentioned but unclear as to which groups
Selective reporting (reporting bias)	Unclear risk	Insufficient information to permit judgement of 'high risk' or 'low risk'
Other bias	Low risk	The study appears to be free of other sources of bias
Blinding (participant)	High risk	Not possible
Blinding (assessor)	Unclear risk	Not reported
Comparability of exercise and control group at entry	Low risk	No difference between the groups at baseline
Appropriateness of duration of surveillance	High risk	Only immediately postintervention data 2 years, no follow-up data reported

Korpelainen 2006

Methods	Type of study: RCT
Participants	Number of participants randomised = 160 Losses: 68 women (81.0%) in the exercise group and 65 (85.5%) women in the control group completed the study Age: mean age 73 years Setting: Finland Inclusion: hip BMD value of more than 2 SD below the reference value Exclusion: use of a walking aid device other than a stick, bilateral hip joint replacement, unstable chronic illness, malignancy, medication known to affect bone density, severe cognitive impairment and involvement in other interventions
Interventions	Exercise group **(COMB)** (n = 84): jumping and balance exercises, including walking, knee bends, leg lifts, heel rises and drops, dancing, stamping, stair climbing and stepping up and down from benches Control Group (n = 76): usual activity Duration and intensity: 1hr sessions, 30 months Supervisor: physiotherapist Supervision:group and individual Setting: clinic and home
Outcomes	BMD Radius and hip (total hip, neck of femur, trochanter) During the 30-month follow-up, there were 88 falls in the exercise group and 101 falls in the control group (P = 0.10). The incidence of fall-related fractures was higher in the control group (n = 16) than in the exercise group (n = 6; P = 0.019). One woman in the control group had two fractures, and all other 20 women had one fracture
Notes	Compliance/adherence: Attendance at the exercise sessions averaged 78% during the first supervised 6-month period, 74% during the second supervised period and 73% during the last supervised 6 months. The average frequency of performing the home exercise programme was three times per week Adverse events: Three women in the exercise group experienced musculoskeletal problems that required minor modifications in the training regimen 5% level would require 64 women in each group to give an 80% power to detect a 0.02 g/cm^2 difference in the primary outcome (femoral neck, trochanter and total hip BMD with an SD of 0.04 g/cm^2) between the groups

Risk of bias

Bias	Authors' judgement	Support for judgement
Random sequence generation (selection bias)	Low risk	Computer-generated random numbers
Allocation concealment (selection bias)	Low risk	Randomisation provided by a technical assistant not involved in the conduction of the trial

Korpelainen 2006 *(Continued)*

Incomplete outcome data (attrition bias) All outcomes	Low risk	Data were analysed on an intention-to-treat basis, and any missing follow-up data was replaced with the last known value even if this was the baseline value
Selective reporting (reporting bias)	Unclear risk	Insufficient information to permit judgement of 'high risk' or 'low risk'
Other bias	Low risk	The study appears to be free of other sources of bias
Blinding (participant)	High risk	Not possible
Blinding (assessor)	Low risk	Operators were unaware of the women's trial status
Comparability of exercise and control group at entry	Low risk	No significant difference between the groups at baseline
Appropriateness of duration of surveillance	Low risk	Immediately postintervention data 30 months, with follow-up data reported mean 7.1 years

Maddalozzo 2007

Methods	Type of study: RCT
Participants	Number of participants randomised = 141 Losses: retention rates 83% NHRT plus exercise, 89% HRT plus exercise; 91% HRT no exercise; and 82% control group Age: 52.1± 3.0 years Setting: USA Inclusion: women who had experienced the menopause within the previous 0-36 months from the time of baseline testing as determined retrospectively from questionnaire reports; (2) no menstrual cycles within the previous 12 months without being pregnant, but not longer than 36 months (based on questionnaire recall phone screening interview); (3) follicle-stimulating hormone levels ≥ 40 mIU/mL (obtained from the subjects physician) ; (4) body mass index (19-30 kg m−2), (5) 36 months or less of being diagnosed as being postmenopausal by their general physician; and (6) either taking HRT 0.625 mg conjugated equine oestrogen, (Premarin®) or non HRT use Exclusion: non-HRT users who had taken HRT for 12 consecutive months prior to applying to the study; (2) hypertension; (3) metabolic disease that may affect bone or muscle metabolism (including diabetes and thyroid disease); (4) statin medications for hypercholesterolaemia), multiple sclerosis; and (4) osteoarthritis or other musculoskeletal disorders that prevented participation
Interventions	Non HRT Exercise group (**DWBHF**) (n = 35): free weight back squat and free weight dead lift exercises repetitions at a speed of 1-2 sets for the concentric (lifting) and 2-3 sets for the eccentric (lowering) phases. Two warm-up sets of 10-12 repetitions at 50% of 1RM then 3 working sets at 60-75% of 1 RM (set 1 = 8 reps; set 2 = 10 reps; and set 3 = 12 reps) Non HRT Control Group (n = 34) HRT exercise group (**DWBHF**) (n = 37): as Non HRT Exercise group HRT Control Group (n = 35) Duration and intensity: 50 mins 2 x week for 52 weeks Supervisor: personal trainer Supervision: individual Setting: gym
Outcomes	BMD DXA lumbar spine (L1-L4), proximal femur (total hip, femoral neck, and greater trochanter) and whole body composition
Notes	Compliance/adherence: non-HRT plus exercise (84.7±12.8%) and HRT plus exercise group (86.2±11.4%) Adverse events: none reported Desired power ≥ 0.8, alpha = 0.05, and an expected difference between groups of 4% increase in muscle mass and a 1% increase in spine BMD, 25 subjects per group were needed

Risk of bias

Bias	**Authors' judgement**	**Support for judgement**

Maddalozzo 2007 *(Continued)*

Random sequence generation (selection bias)	Unclear risk	Self selected as either HRT or non-HRT replaced then randomised. Randomisation mentioned but insufficient information to permit judgement of 'high risk' or 'low risk'
Allocation concealment (selection bias)	Unclear risk	Insufficient information to permit judgement of 'high risk' or 'low risk'
Incomplete outcome data (attrition bias) All outcomes	Unclear risk	'As-treated' analysis done drop-outs mentioned but unclear as to which groups
Selective reporting (reporting bias)	Unclear risk	Insufficient information to permit judgement of 'high risk' or 'low risk'
Other bias	Low risk	The study appears to be free of other sources of bias
Blinding (participant)	High risk	Not possible
Blinding (assessor)	Unclear risk	Not reported
Comparability of exercise and control group at entry	Low risk	No significant differences were observed at baseline on any variable except for spine BMD between HRT and non-HRT groups
Appropriateness of duration of surveillance	High risk	Only immediately postintervention data at 52 weeks, no follow-up data reported

Newstead 2004

Methods	Type of study: RCT
Participants	Number of participants randomised = 53 Losses: 7 (2 exercise, 5 control) Age: 50-65 years Setting: USA Inclusion: no co-morbidity e.g. diabetes, CHD, PVD, pulmonary or orthopaedic dysfunctions; not taking alendronate medication etc.; no current exercise programme; no history of osteoporotic fractures; BMI 21-31; on HRT if postmenopausal for >5 years. BMD T-score > -1.5 SD at hip and lumbar spine Exclusion: not reported
Interventions	Exercise group jumping (**DWBHF**)(n = 25): progressive multidirectional jumping, increasing jump heights and repetitions (max 200) Control Group (n = 28): usual activity Duration and intensity: 3 sessions per week for 12 months Supervisor: physical therapist Supervision: group 2x week, individual 1 x week Setting: gym
Outcomes	BMD femoral neck, total hip, lumbar spine

Notes	Compliance/adherence: average 82% at month 6 and 75% month 12 Adverse events: none reported

Risk of bias

Bias	**Authors' judgement**	**Support for judgement**
Random sequence generation (selection bias)	Unclear risk	Randomisation mentioned but insufficient information to permit judgement of 'high risk' or 'low risk'
Allocation concealment (selection bias)	Unclear risk	Insufficient information to permit judgement of 'high risk' or 'low risk'
Incomplete outcome data (attrition bias) All outcomes	Unclear risk	'As-treated' analysis done drop-outs mentioned but unclear as to which groups
Selective reporting (reporting bias)	Unclear risk	Insufficient information to permit judgement of 'high risk' or 'low risk'
Other bias	Low risk	The study appears to be free of other sources of bias
Blinding (participant)	High risk	Not possible
Blinding (assessor)	Unclear risk	Not reported
Comparability of exercise and control group at entry	Unclear risk	Insufficient information to permit judgement of 'high risk' or 'low risk'
Appropriateness of duration of surveillance	High risk	Only immediately postintervention data 12 months, no follow-up data reported

Pruitt 1996

Methods	RCT
Participants	Number of participants randomised - 40 Losses: 14 Age: 65-82 years Setting: America Inclusion: Healthy caucasian women not currently taking HRT, or those on HRT for 1 year or more Exclusion: evidence of acute or uncontrolled chronic illness or conditions that would prevent participation in exercise class, vertebral compression fractures, disorders affecting bone metabolism

Interventions	Exercise group 1 (**NWBHF**) (n = 15): supervised exercise session comprising bench press, lateral pull down, military press, biceps curl, knee extension, knee flexion, hip abduction and adduction, leg press, back extension. 1 set 14 reps at 40% 1RM, 2 sets 7 reps at 80% 1RM Exercise group 2 (**NWBLF**) (n = 13): a/a 3 sets 12 reps at 40% 1RM Control Group (n = 12): no exercises Duration and intensity: 3 times per week for 12 months, lifting time 50 -55mins. 1RM tests administered every 2 weeks for first 3 months then every 3 weeks to adjust workload Supervisor: not recorded Supervision: every session Setting: Gym
Outcomes	BMD lumbar spine, hip (total hip, neck of femur, Wards triangle) at baseline and 12 months
Notes	Compliance/adherence: 65% Adverse events: aggravation of pre-existing back or knee condition (n = 2)

Risk of bias

Bias	**Authors' judgement**	**Support for judgement**
Random sequence generation (selection bias)	Unclear risk	Randomisation reported but insufficient information about the sequence generation process to permit judgement of 'high risk' or 'low risk'
Allocation concealment (selection bias)	Unclear risk	Insufficient information to permit judgement of 'high risk' or 'low risk'
Incomplete outcome data (attrition bias) All outcomes	Unclear risk	'As-treated' analysis done drop-outs mentioned but different across the groups
Selective reporting (reporting bias)	Unclear risk	Insufficient information to permit judgement of 'high risk' or 'low risk'. One outlier whose spinal BMD was more than 4SD from group mean was not included in analysis
Other bias	Low risk	The study appears to be free of other sources of bias
Blinding (participant)	High risk	Not possible
Blinding (assessor)	Unclear risk	Insufficient information to permit judgement of 'high risk' or 'low risk'
Comparability of exercise and control group at entry	Low risk	No significant differences observed in baseline characteristics
Appropriateness of duration of surveillance	High risk	Only immediately postintervention data at 12 months, no follow-up data reported

Verschueren 2004

Methods	Type of study: RCT
Participants	Number of participants randomised = 70 Losses: not reported Age: 58-74 years Setting: Belgium Inclusion: 60 and 70 years of age, non-institutionalised, and free from diseases or medications known to affect bone metabolism or muscle strength Exclusion: total body BMD T-score of less than -2.5
Interventions	Exercise group vibrating platform **(DWBHF)**(n = 25): static and dynamic knee-extensor exercises on the vibration platform, progressive exercise Exercise group resistance training **(NWBHF)**(n = 22): warm-up, resistance training programme for knee extensors on a leg extension and a leg press machine. Training programme was designed (ASCM) for individuals older than 60 years of age. Progressive resistance Control Group (n = 23): usual activity Duration and intensity: 72 training sessions within a 24-week period. Training frequency was three times a week Supervisor: not stated Supervision: individual and group for resistance training Setting: gym
Outcomes	BMD DEXA total hip, total body
Notes	Compliance/adherence: not reported Adverse events: none reported

Risk of bias

Bias	Authors' judgement	Support for judgement
Random sequence generation (selection bias)	Low risk	Randomisation by computer-generated random numbers age-matched women
Allocation concealment (selection bias)	Unclear risk	Insufficient information to permit judgement of 'high risk' or 'low risk'
Incomplete outcome data (attrition bias) All outcomes	Unclear risk	'As-treated' analysis done, insufficient information to permit judgement of 'high risk' or 'low risk'
Selective reporting (reporting bias)	Unclear risk	Insufficient information to permit judgement of 'high risk' or 'low risk'
Other bias	Low risk	The study appears to be free of other sources of bias
Blinding (participant)	High risk	Not possible
Blinding (assessor)	Low risk	Technician unaware of intervention type

Verschueren 2004 *(Continued)*

Comparability of exercise and control group at entry	Low risk	No significant differences were observed at baseline between the experimental and the control groups in terms of age, weight, body mass, years since menopause, BMD, serum levels of osteocalcin and CTX, isometric and dynamic muscle strength, fat mass or lean body mass
Appropriateness of duration of surveillance	High risk	Only immediately postintervention data, no follow-up data reported

Von Stengel 2009

Methods	Type of study:RCT
Participants	Number of participants randomised = 151 Losses: 16, group 1 n = 5, group 2 n = 7, control n = 4. All invited for final measurements, 11 did not attend; group 1 n = 1, group 2 n = 6, control n = 4 Age: 65-72 years Setting: Germany Inclusion: Over 65, postmenopausal Exclusion: relevant co-morbidity or drug treatment which could influence bone metabolism
Interventions	Exercise group (**COMB**) (n = 50): Low impact aerobics, strengthening exercises and balance Exercise group (**COMB**) (n = 50): Low impact aerobics, strengthening and balance exercise as above and vibration plate. Vibration between 25-35 Hz, intensity increased at 3 and 6 months Control Group (n = 51): gentle exercise and relaxation class x 1 per week Duration and intensity: 60 minutes 2x per week for 12 months Supervisor: not reported Supervision: not reported Setting: hospital
Outcomes	BMD total hip and spine, rate of falls
Notes	Compliance/adherence: not reported Adverse events: none recorded Selected exercise group with vibration plate for analysis. Data converted to % change

Risk of bias

Bias	**Authors' judgement**	**Support for judgement**
Random sequence generation (selection bias)	Unclear risk	Randomisation mentioned but insufficient information to permit judgement of 'Yes' or 'No'
Allocation concealment (selection bias)	Unclear risk	Insufficient information to permit judgement of 'Yes' or 'No'

Von Stengel 2009 *(Continued)*

Incomplete outcome data (attrition bias) All outcomes	Low risk	Losses explained and data analysed on intention-to-treat
Selective reporting (reporting bias)	Low risk	Reporting as per protocol
Other bias	Low risk	The study appears to free of other sources of bias
Blinding (participant)	High risk	Not possible
Blinding (assessor)	Unclear risk	Not reported
Comparability of exercise and control group at entry	Low risk	No significant differences between groups at entry
Appropriateness of duration of surveillance	High risk	Only immediately postintervention data at 12 months, no follow-up data reported