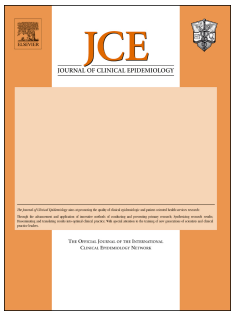


Most healthcare interventions tested in Cochrane Reviews are not effective according to high quality evidence: a systematic review and meta-analysis

Jeremy Howick, PhD, Despina Koletsi, DDS, MSc, Dr. med. dent., MSc DLSHTM, PGCHEd, John P.A. Ioannidis, MD, DSc, Professor, Dr. Claire Madigan, PhD, Nikolaos Pandis, PhD, Professor, Dr. Martin Loef, PhD, Harald Walach, PhD, Professor, Sebastian Sauer, PhD, Professor, Jos Kleijnen, PhD, Professor, Dr. Jadbinder Seehra, Ms. Tess Johnson, Stefan Schmidt, PhD, Professor

PII: S0895-4356(22)00100-7

DOI: https://doi.org/10.1016/j.jclinepi.2022.04.017

Reference: JCE 10816

To appear in: *Journal of Clinical Epidemiology*

Received Date: 24 January 2022

Revised Date: 25 March 2022

Accepted Date: 12 April 2022

Please cite this article as: Howick J, Koletsi D, Ioannidis JPA, Madigan C, Pandis N, Loef M, Walach H, Sauer S, Kleijnen J, Seehra J, Johnson T, Schmidt S, Most healthcare interventions tested in Cochrane Reviews are not effective according to high quality evidence: a systematic review and meta-analysis, *Journal of Clinical Epidemiology* (2022), doi: https://doi.org/10.1016/j.jclinepi.2022.04.017.

Author statement

JH (guarantor) conceived of the idea (together with StS), and wrote the first draft of the protocol all authors contributed to developing the protocol. JH piloted the data extraction form and all authors made suggestions for improvement. JH, DK, TJ, CM, ML, HW, SeS, JS, NP, StS, JPAI contributed to the data extraction. SeS developed a computerized quality check; HW and JH resolved discrepancies. JH, JPAI, CM, and DK developed a plan for and analyzed the data. JH drafted the final manuscript, with contributions from all authors.

Title

Most healthcare interventions tested in Cochrane Reviews are not effective according to high quality evidence: a systematic review and meta-analysis

Authors

Jeremy Howick, PhD (corresponding author)

jeremy.howick@philosophy.ox.ac.uk

Kleijnen Systematic Reviews and the Faculty of Philosophy

University of Oxford

Oxford OX2 6GG

+44 (0)7771925412

Despina Koletsi, DDS, MSc, Dr. med. dent., MSc DLSHTM, PGCHEd, Clinic of Orthodontics and Pediatric Dentistry, Center of Dental Medicine, University of Zurich, Switzerland [joint 1st author]

Professor John P.A. Ioannidis, MD, DSc, Departments of Medicine, of Epidemiology and Population Health, of Biomedical Data Science, and of Statistics and Meta-Research Innovation Center at Stanford (METRICS), Stanford University

Dr. Claire Madigan, PhD, Loughborough University

Professor Nikolaos Pandis, PhD, Department of Orthodontics and Dentofacial Orthopedics, School of Dental Medicine, University of Bern, Bern, Switzerland

Dr. Martin Loef, PhD, CHS-Institute, Berlin

Professor Harald Walach, PhD, CHS-Institute, Berlin

Professor Sebastian Sauer, PhD, Ansbach University

Professor Jos Kleijnen, PhD, Kleijnen Systematic Reviews Ltd.

Dr. Jadbinder Seehra, Centre for Craniofacial Development & Regeneration, Faculty of Dentistry, Oral & Craniofacial Sciences, King's College London

Ms. Tess Johnson, Oxford Uehiro Centre for Practical Ethics, University of Oxford

Professor Stefan Schmidt, PhD, Department of Psychosomatic Medicine and Psychotherapy, Medical Center, University of Freiburg and Institute for Frontier Areas in Psychology and Mental Health, Freiburg

Word count

3563 words

Contributions

JH (guarantor) conceived of the idea (together with StS), and wrote the first draft of the protocol all authors contributed to developing the protocol. JH piloted the data extraction form and all authors made suggestions for improvement. JH, DK, TJ, CM, ML, HW, SeS, JS, NP, StS, JPAI contributed to the data extraction. SeS developed a computerized quality check; HW and JH resolved discrepancies. JH, JPAI, CM, and DK developed a plan for and analyzed the data. JH drafted the final manuscript, with contributions from all authors.

Abstract

Objective

To estimate the proportion of healthcare interventions tested within Cochrane Reviews that are effective according to high-quality evidence.

Study design and setting

We selected a random sample of 2428 (35%) of all Cochrane Reviews published between 1 January 2008 and 5 March 2021. We extracted data about interventions within these reviews that were compared with placebo, or no treatment, and whose outcome quality was rated using Grading of Recommendations Assessment, Development and Evaluation (GRADE). We calculated the proportion of interventions whose effectiveness was based on high-quality evidence according to GRADE, had statistically significant positive effects, and were judged as beneficial by the review authors. We also calculated the proportion of interventions that suggested harm.

Results

Of 1567 eligible interventions, 87 (5.6%) had high quality evidence on first-listed primary outcomes, positive, statistically significant results and were rated by review authors as beneficial. Harms were measured for 577 (36.8%) interventions, 127 of which (8.1%) had statistically significant evidence of harm. Our dependence on the reliability of Cochrane author assessments (including their GRADE assessments) was a potential limitation of our study.

Conclusion

Most healthcare interventions studied within recent Cochrane Reviews are not supported by high quality evidence, and harms are under-reported.

Key words

Evidence; systematic review; epidemiology; quality; safety, harm

Registration

PROSPERO: CRD42021240989

Funding Source

This study was not externally funded.

1. Introduction

1.1.Rationale

Early evidence-based medicine researchers exposed some widely used interventions as useless or harmful. For example, antiarrhythmic drugs were widely prescribed in the belief that they would reduce mortality from myocardial infarction until a placebo-controlled trial found that the drugs increased mortality (1). Putting infants to sleep on their stomachs was recommended based on experts' idea that babies would be less likely to choke on their vomit (2), until large epidemiological studies found that stomach sleeping increased the risk of sudden infant death syndrome (3). More recently, the benefits of (widely used) oseltamivir for influenza were called into question by a systematic review of clinical study reports (4). Yet, critics have questioned the extent to which these celebrated examples are representative (5, 6). Pointing to examples of modern medicine's success—including the discovery of penicillin and the great increases in life span over the last 100 years—some have claimed that the examples of harmful or useless medicine are exceptional (5). However, the increase in life span can at least partly be explained by general social, hygienic and economic progress than to the benefit of modern medicine (7).

Several studies have gone beyond examples to systematically investigate the extent to which high quality evidence supports the benefits of healthcare interventions. In a study published in 2001, authors independently rated interventions studied in 160 Cochrane Reviews on a 6-point scale ranging from "harmful" to "positive effect." They found that 22.5% of healthcare interventions report evidence of a positive effect. A study published in 2007, used a similar scale to rate interventions in a sample of 1016 Cochrane reviews and found that 56% of healthcare interventions were likely to be beneficial (8). A problem with these two studies is that the scale used to determine whether an intervention was effective has not been validated or widely accepted. Also published in 2007, a review of 2500 treatments that were supported

by “good” evidence (defined as evidence from randomised trials) found that 22% were likely to be beneficial (9). Problems with this third study is that it equates good evidence with evidence from randomised trials (10) and did not use a transparent method for determining whether an intervention was beneficial.

Since these earlier studies have been done, the Grading of Recommendations Assessment, Development and Evaluation (GRADE) system was introduced (11). Compared with previous methods used to rate evidence quality, GRADE is robust, transparent, and has been endorsed by over 100 organisations worldwide (12). Using GRADE, a recent sample of 154 Cochrane reviews published in 2020 that were updates of reviews published in 2013/14 (13) found that 10% had primary outcomes supported by high quality evidence (14). However, this review was based on a small sample and may not be representative.

1.2.Objectives

The aim of this study was to provide an up-to-date and robust estimate of the proportion of healthcare intervention whose beneficial effects are supported by high quality evidence. We also aimed to measure the proportion of interventions whose harms were measured.

2. Methods

Our meta-epidemiological study was reported in accordance with the 2009 PRISMA statement (15). Our review protocol was registered with PROSPERO in April 2021 (registration number CRD42021240989).

2.1.Information sources

To identify eligible interventions, we searched all systematic reviews of interventions in the Cochrane Database of Systematic Reviews published between 1 January 2008 (when GRADE became commonly used) and 5 March 2021. Our units of assessment were interventions within Cochrane reviews.

2.2.Eligibility criteria

To be eligible for inclusion, interventions had to be compared with placebo, no treatment, or treatment as usual (but not an active comparator). This was to avoid problems that could arise from comparison with an active intervention which itself might not have evidence to support its benefits. Only reviews with interventions rated by GRADE were included. We excluded interventions within outdated versions of reviews that were superseded by newer versions, and interventions within withdrawn reviews.

To satisfy our primary outcome, an intervention had to have at least one high quality GRADE primary outcome, a statistically significant beneficial effect, the review authors had to deem the intervention to be effective.

2.3.Search strategy

The titles, authors, year, doi, and Cochrane Review Group information were retrieved directly from the Cochrane Library for all intervention reviews published within our specified timeframe (N=6928 reviews) into Excel. One author (JH) used a random number generator to obtain a stratified random sample of 35% of reviews from each review group (16, 17).

2.4.Data items

A standardised, pre-piloted form was used to extract data. For all eligible interventions, we extracted information about the author, year, review group, condition,

population (adults, children, mixed), study designs (randomised trials or observational studies), intervention (description), intervention category (pharmacological, psychological, surgical, behavioural, diet and exercise, manual therapies, alternative, other), and control (placebo/sham, usual care, no treatment). Where available, we also extracted data about harms for all eligible studies. We recorded evidence of harms that was quantified either in the Summary of Findings table or elsewhere in the report. We classified harms into: mortality, other mortality, other objective (observer reported), subjective (participant reported), and unclear. Where more than one harm was listed, we collected data for the first reported one.

If none of the outcomes were rated as high quality, we recorded the highest GRADE rating for any outcome (primary or not). For interventions with a high-quality outcome, we extracted additional data. Following a method used in a related review, we first reported reviews in which the primary outcome was rated high quality (14). For these, we extracted the following: description of the outcome, outcome category (subjective/objective), number of trials and number of participants, effect, effect size, significance level, whether there was a statistically significant positive result, and whether the original review authors deemed the intervention to be effective. This information was obtained from the conclusions section in the review abstract and the body of the review (subsections "implications for practice" and/or "implications for research") (13). An example of a positive interpretation was: "Buprenorphine should be supported as a medication to use" (18). We accounted for the fact that a beneficial effect could be either a desirable reduction or an increase in an outcome. To avoid missing important information despite focusing on the first reported primary outcome, we also recorded whether up to three other outcomes (primary or not) were rated as high-quality and extracted the same data for these as for first reported high-quality outcomes.

Data was extracted by a single reviewer (from among JH, DK, CM, NP, ML, HW, SS, JS, TJ). One author (JH) checked 10% of all extractions, with discrepancies resolved in

discussion with other reviewers. In addition, all extractions were checked by a computer program using R (19). The program extracted the following data directly from Cochrane Reviews: whether there was a GRADE rating, whether there was a high GRADE-rated outcome and if so, what the high GRADE-rated outcome was, and whether the review had been withdrawn, and whether there was a newer version of the review. Discrepancies between computer and human extractions were resolved by two reviewers (HW, JH).

2.5.Risk of bias assessment

To determine whether the reviews themselves (and therefore our conclusions extracted from the reviews) were reliable, we assessed the risk of bias of the systematic reviews using the Kleijnen Systematic Review (KSR) Evidence database. This database has rated all Cochrane reviews dating back to 2015 using the Risk of Bias Tool in Systematic Reviews (ROBIS) tool (20). We were able to identify which of the high quality first reported primary outcomes were supported by reviews with a low risk of bias.

2.6 Data Synthesis and Analysis

We produced descriptive statistics and proportions (n/N) and where applicable 95% confidence intervals and P-values were presented. Data management, processing and analysis was performed with Stata version 15.1 (21).

On an exploratory basis, we designed a mixed effects logistic regression model to predict any effect of year, intervention category and comparator category, on the odds of an intervention meeting our primary outcome. Also on an exploratory basis, we investigated the proportion of interventions that had at least one primary outcome that was either high quality or moderate quality.

2.7 Protocol Amendments

We initially intended to test for differences between the different Cochrane Review groups. However, there were too few reviews in some groups and too many groups to allow for this. Instead, we presented the results of the primary outcome by review group, allowing for informal comparisons (Table 1). Also, after the protocol was published, we decided to use a computerized algorithm to check human extraction and explore how well a text mining program could extract information. In addition, to account for problems with associating statistical significance with effectiveness, we modified the primary outcome to include interventions whose effectiveness was also supported by the review authors.

3. Results

3.1.Sample identification

From our sample of 2428 reviews, 1567 interventions from 1076 independent reviews met our inclusion criteria (see Appendix Table 1). The overlapping reasons for exclusion were: did not include a GRADE assessment (n=875). interventions were compared with an active intervention (n=684), were studied in reviews that had been withdrawn (n=112), and had been superseded by more up-to-date version (n=46).

Our sample included interventions from all 53 Cochrane Review groups (see Table 1). The interventions were tested in adults (n= 892), adults or children (n=413), children (n= 80), infants or children (n= 38), infants (n=77), and unclear populations (n=67). Most (1468, 93.7%) interventions' effects were tested in randomised trials, with 88 (5.6%) tested in a mix of randomised and non-randomised trials, and 11 (0.7%) tested in non-randomised trials. Over half (n=820, 52.3%) the interventions were pharmacological, 247 (15.8%) were behavioural or psychological, 100 (6.4%) surgical, 62 (4.0%) diet, 56 (3.6%) exercise, 46 (2.9%) alternative, 38 (2.4%) manual therapies, and 198 (12.6%) other (not included in the

abovementioned categories). The comparators were: placebo or sham in 708 (45.2%), usual or standard care in 546 (34.8%), and no treatment in 313 (20.0%) interventions.

The text-mining program revealed 1846 (19% of all relevant cells in the extraction table) discrepancies between the program and human. The human was deemed to be correct in most of the cases (97%).

3.2.Quality of Evidence Supporting Intervention Effects

One in 10 (158 of 1567 interventions had a first listed primary outcome with evidence rated as high quality. Of these, 106 (6.8%) also had a positive, statistically significant result, while 87 (5.6%) were also rated by review authors as being at least very likely to be effective (see Table 1). A breakdown of those 87 interventions across intervention category, type of comparator treatment, type of outcome, and target population is presented in Table 2. The majority of the 87 interventions were pharmacological (73.6%).

An additional 31 interventions (2.0%) had at least one other primary but not first reported) outcome with high quality evidence. Among the interventions that did not have any outcomes with high quality evidence, the highest GRADE rating was moderate in 472 (30.1%), low in 533 (34.0%), and very low in 373 (23.8%) (see Figure 1).

Figure 1. Proportion of interventions according to their highest GRADE outcome (high, moderate, low, very low)

[Figure 1 here]

3.3.Quality of Evidence Supporting Intervention Harms

Harms of 577 out of 1567 (36.8%) interventions were quantified. Very few (n=33, 5.7%) were supported by high quality evidence, while an additional 150 out of 577 (26.0%) were rated as moderate quality. Forty (6.9%) of these measured mortality, 402 (69.7%) measured other objective outcomes, 25 (4.3%) measured subjective outcomes, and 110 (19.1%) measured unclear or unspecified types of outcomes. Out of the 577 interventions that reported harms, there was evidence of a statistically significant harm in 127 (22%). The breakdown of these 127 that suggested a significant harmful effect across intervention category, type of comparator treatment, type of outcome, target population and grade (see Table 3). Few of those (18/127; 14.2%) were supported by high quality evidence. Other than for pharmacological interventions, harms were unlikely to be reported (see Table 3).

Table 1. Characteristics of included interventions

	Intervention benefits				Harms	
	No. interventions that met inclusion criteria	**% with first listed primary outcome high quality according to GRADE**	**% first listed primary outcome high quality according to GRADE + % Effective**	**% first listed primary outcome high quality according to GRADE +% Effective + review authors state effective**	**n (%) quantify harms**	**n (%) evidence of harm**
Cochrane Review Group						
Overall	*1567*	*158 (10.1)*	*106 (6.8%)*	*87 (5.6%)*	*577 (36.8%)*	*127 (8.1%)*
Acute Respiratory Infections	30	6 (20.0%)	5 (16.7%)	5 (116.7%)	12 (40.0%)	3 (10.0%)
Airways	52	6 (11.5%)	6 (11.5%)	5 (9.6%)	25 (48.1%)	3 (5.8%)
Anaesthesia	22	2 (9.1%)	2 (9.1%)	2 (9.1%)	16 (72.7%)	1 (4.5%)
Back and Neck	10	0 (0.0%)	0 (0.0%)	0 (0.0%)	2 (20.0%)	1 (10.0%)
Bone, Joint and Muscle Trauma	21	3 (14.3%)	3 (14.3%)	3 (14.3%)	4 (19.1%)	1 (4.8%)
Breast Cancer	11	5 (45.5%)	2 (18.2%)	1 (9.1%)	2 (18.2%)	2 (18.2%)
Childhood Cancer	5	0 (0.0%)	0 (0.0%)	0 (0.0%)	2 (40.0%)	0 (0.0%)
Colorectal	14	3 (21.4%)	3 (21.4%)	3 (21.4%)	8 (57.1%)	2 (14.3%)
Common Mental Disorders	48	3 (6.3%)	2 (4.2%)	2 (4.2%)	19 (39.6%)	3 (6.3%)
Consumers and Communication	12	1 (8.3%)	1 (8.3%)	1 (8.3%)	1 (8.3%)	0 (0.0%)
Cystic Fibrosis and Genetic Disorders	36	1 (2.8%)	0 (0.0%)	0 (0.0%)	14 (38.9%)	1 (2.8%)
Dementia and Cognitive Improvement	29	6 (20.7%)	2 (6.9%)	1 (3.4%)	14 (48.3%)	5 (17.2%)

	Intervention benefits				Harms	
	No. interventions that met inclusion criteria	**% with first listed primary outcome high quality according to GRADE**	**% first listed primary outcome high quality according to GRADE + % Effective**	**% first listed primary outcome high quality according to GRADE +% Effective + review authors state effective**	**n (%) quantify harms**	**n (%) evidence of harm**
Developmental, Psychosocial and Learning Problems	42	4 (9.5%)	3 (7.1%)	3 (7.1%)	14 (33.3%)	3 (7.1%)
Drugs and Alcohol	14	3 (21.4%)	3 (21.4%)	2 (14.3%)	2 (14.3%)	1 (7.1%)
Effective Practice and Organisation of Care	37	8 (21.6 %)	6(16.2%)	5 (13.5%)	4 (10.8%)	0 (0.0%)
Emergency and Critical Care	15	0 (0.0%)	0 (0.0%)	0 (0.0%)	9 (60.0%)	0 (0.0%)
ENT	18	1 (5.6%)	1 (5.6%)	1 (5.6%)	11 (61.1%)	2 (11.1%)
Epilepsy	12	2 (16.7%)	1 (8.3%)	1 (8.3%)	6 (50.0%)	4 (33.3%)
Eyes and Vision	44	8 (18.2%)	4 (9.1%)	4 (9.1%)	17 (38.6%)	6 (13.6%)
Fertility Regulation	9	3 (33.3%)	0 (0.0%)	0 (0.0%)	1 (11.1%)	1 (11.1%)
Gut	49	5 (10.2%)	5 (10.2%)	5 (10.2%)	29 (59.2%)	2 (4.1%)
Gynaecological, Neuro-oncology and Orphan Cancer	36	5 (13.9%)	4 (11.1%)	4 (11.1%)	14 (38.9%)	4 (11.1%)
Gynaecology and Fertility	62	0 (0.0%)	0 (0.0%)	0 (0.0%)	32 (51.6%)	8 (12.9%)
Haematology	15	1 (6.7%)	1 (6.7%)	1 (6.7%)	8 (53.3%)	2 (13.3%)
Heart	44	2 (4.5%)	1 (2.3%)	0 (0.0%)	21 (47.7%)	9 (20.5%)
Hepato-Biliary	41	1 (2.4%)	0 (0.0%)	0 (0.0%)	21 (51.2%	1 (2.4%)
HIV/AIDS	7	3 (42.9%)	3 (42.9%)	2 (28.6%)	2 (28.6%)	0 (0%)
Hypertension	18	1 (5.6%)	1 (5.6%)	1 (5.6%)	10 (55.6%)	3 (16.7%)
Incontinence	9	2 (22.2%)	1 (11.1%)	0 (0.0%)	3 (33.3%)	0 (0.0%)
Infectious Diseases	22	6 (27.3%)	4 (18.2%)	3 (13.6%)	7 (31.8%)	3 (13.6%)
Injuries	10	1 (10%)	0 (0.0%)	0 (0.0%)	1 (10.0%)	0 (0.0%)

	Intervention benefits				Harms	
	No. interventions that met inclusion criteria	**% with first listed primary outcome high quality according to GRADE**	**% first listed primary outcome high quality according to GRADE + % Effective**	**% first listed primary outcome high quality according to GRADE +% Effective + review authors state effective**	**n (%) quantify harms**	**n (%) evidence of harm**
Kidney and Transplant	55	4 (7.3%)	3 (5.5%)	3 (5.5%)	12 (21.8%)	3 (5.5%)
Lung Cancer	7	2 (28.6%)	1 (14.3%)	1 (14.3%)	3 (42.9%)	2 (28.6%)
Metabolic and Endocrine Disorders	21	0 (0.0%)	0 (0.0%)	0 (0.0%)	7 (33.3%)	3 (14.3%)
Movement Disorders	7	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Multiple Sclerosis and Rare Diseases of the CNS	14	1 (7.1%)	0 (0.0%)	0 (0.0%)	4 (28.6%)	1 (7.1%)
Musculoskeletal	50	12 (24%)	11 (22.0%)	10 (20.0%)	27 (54.0%)	8 (16.0%)
Neonatal	63	4 (6.3%)	3 (4.8%)	3 (4.8%)	25 (39.7%)	3 (4.8%)
Neuromuscular	54	5 (9.3%)	2 (3.7%)	1 (1.9%)	21 (38.9%)	5 (9.3%)
Oral Health	38	1 (2.6%)	0 (0.0%)	0 (0.0%)	3 (7.9%)	0 (0.0%)
Pain, Palliative and Supportive Care	53	3 (5.7%)	3 (5.7%)	2 (3.8%)	18 (34.0%)	7 (13.2%)
Pregnancy and Childbirth	76	10 (13.2%)	5 (6.6%)	4 (5.3%)	14 (18.4%)	3 (3.9%)
Public Health	23	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Schizophrenia	48	3 (6.3%)	3 (6.3%)	3 (6.3%)	26 (54.2%)	5 (10.4%)
Sexually Transmitted Infections	4	1 (25%)	0 (0.0%)	0 (0.0%)	1 (25.0%)	1 (25%)
Skin	65	7 (10.8%)	4 (6.2%)	3 (4.6%)	43 (66.2%)	5 (7.7%)
Stroke	35	1 (2.9%)	0 (0.0%)	0 (0.0%)	10 (28.6%)	1 (2.9%)
Tobacco Addiction	25	1 (4.0%)	1 (4.0%)	1 (4.0%)	1 (4.0%)	0 (0.0%)

	Intervention benefits				Harms	
	No. interventions that met inclusion criteria	**% with first listed primary outcome high quality according to GRADE**	**% first listed primary outcome high quality according to GRADE + % Effective**	**% first listed primary outcome high quality according to GRADE +% Effective + review authors state effective**	**n (%) quantify harms**	**n (%) evidence of harm**
Urology	20	3 (15.0%)	0 (0.0%)	0 (0.0%)	12 (60.0%)	3 (15.0%)
Vascular	39	7 (17.9%)	5 (12.8%)	1 (2.6%)	8 (20.5%)	4 (10.3%)
Work	56	1 (1.8%)	1 (1.8%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Wounds	20	0 (0.0%)	0 (0.0%)	0 (0.0%)	11 (55.0%)	2 (10.0%)

Table 2. Characteristics of interventions with high quality evidence, statistically significant effects, and authors interpreting the intervention as beneficial)

	All sample of included interventions	High quality GRADE	High quality GRADE+ statistically significance	High quality GRADE+ statistically significance + authors' interpretation as beneficial
Intervention Category	**n (%)**	**n (%)**	**n (%)**	**n (%)**
Total	**1567 (100.0%)**	**158 (100.0%)**	**106 (100.0%)**	**87 (100.0%)**
Pharmacological	820 (52.3%)	108 (68.4%)	76 (71.7%)	64 (73.6%)
Behavioural/ Psychological	247 (15.8%)	20 (12.6%)	12 (11.3%)	9 (10.3%)
Exercise	56 (3.6%)	7 (4.4%)	6 (5.7%)	5 (5.8%)
Diet	62 (4.0%)	1 (0.6%)	1 (0.9%)	1 (1.2%)
Surgical	100 (6.4)	5 (3.2%)	4 (3.8%)	4 (4.6%)
Alternative	46 (2.9%)	2 (1.3%)	1 (0.9%)	0 (0.0%)
Manual therapies	38 (2.4%)	1 (0.6%)	0 (0.0%)	0 (0.0%)
Other	198 (12.6%)	14 (8.9%)	6 (5.7%)	4 (4.6%)
Comparator				
Placebo/ Sham	708 (45.2%)	87 (55.1%)	61 (57.5)	51 (58.6%)
Usual/ Standard Care	546 (34.8%)	46 (29.1%)	27 (25.5%)	22(25.3%)
No treatment	313 (20.0%)	25 (15.8%)	18 (17.0%)	14 (16.1%)
Type of Outcome				
Mortality	NR	15 (9.5%)	6 (5.7%)	5 (5.8%)
Other Objective	NR	119 (75.3%)	85 (80.2%)	71 (81.6%)
Subjective	NR	24 (15.2%)	15 (14.2%)	11 (12.6%)
Target Population				
Adults	892 (56.9%)	96 (60.8%)	63 (59.4%)	48 (55.2%)
Mixed (adults or children/ infants)	413 (26.4%)	43 (27.2%)	29 (27.4%)	25 (28.7%)
Children	80 (5.1%)	4 (2.5%)	2 (1.9%)	2 (2.3%)
Mixed (children or infants)	38 (2.4%)	3 (1.9%)	2 (1.9%)	2 (2.3%)
Infants	77 (4.9%)	5 (3.2%)	4 (3.8%)	4 (4.6%)
Not stated/ unclear	67 (4.3%)	7 (4.4%)	6 (5.6%)	6 (6.9%)

NR=not recorded (this data was only recorded for high quality (++) outcomes

Table 3. Harms outcomes and statistically significant harmful effect of interventions, divided by categories

	All sample of included interventions	Harms outcomes	Harms outcome+ statistically significant
Intervention Category	**n (%)**	**n (%)**	**n (%)**
Pharmacological	820 (52.3%)	431 (74.7%)	107 (84.3%)
Behavioural/ Psychological	247 (15.8%)	28 (4.9%)	3 (2.3%)
Exercise	56 (3.6%)	4 (0.7%)	0 (0.0%)
Diet	62 (4.0%)	14 (2.4%)	0 (0.0%)
Surgical	100 (6.4)	45 (7.8%)	10 (7.9%)
Alternative	46 (2.9%)	3 (0.5%)	0 (0.0%)

Manual therapies	38 (2.4%)	14 (2.4%)	3 (2.3%)
Other	198 (12.6%)	38 (6.6%)	4 (3.2%)
Comparator			
Placebo/ Sham	708 (45.2%)	358 (62.1%)	87 (68.5%)
Usual/ Standard Care	546 (34.8%)	145 (25.1%)	29 (22.8%)
No treatment	313 (20.0%)	74 (12.8%)	11 (8.7%)
Type of Outcome			
Mortality	NR	40 (6.9%)	2 (1.6%)
Other Objective	NR	402 (69.7%)	95 (74.8%)
Subjective	NR	25 (4.3%)	7 (5.5%)
Unclear/ unspecified	NR	110 (19.1%)	23 (18.1%)
Target Population			
Adults	892 (56.9%)	319 (55.3%)	83 (65.4%)
Mixed (adults or children/ infants)	413 (26.4%)	178 (30.8%)	28 (22.1%)
Children	80 (5.1%)	23 (4.0%)	7 (5.5%)
Mixed (children or infants)	38 (2.4%)	9 (1.6%)	3 (2.4%)
Infants	77 (4.9%)	29 (5.0%)	4 (3.2%)
Not stated/ unclear	67 (4.3%)	19 (3.3%)	2 (1.6%)
Grade Rating			
High	NR	33 (5.7%)	18 (14.2%)
Moderate	NR	150 (26.0%)	58 (45.7)
Low	NR	222 (38.5%)	38 (29.9%)
Very Low	NR	156 (27.0%)	13 (10.2%)
Unspecified	NR	16 (2.8%)	0 (0.0%)
Total	**1567 (100.0%)**	**577 (100.0%)**	**127 (100.0%)**

NR, not recorded for the entire sample of interventions; this is a harm-specific category.

3.4.Risk of Bias in Individual Studies

Of the 87 interventions that met our inclusion criteria, we were able to access ROBIS ratings for 35 (40.2%) of the reviews. Of those, most (n=32, 91.4%) were rated as having a low risk of bias. The remaining four (8.6%) had a high risk of bias.

3.5.Results of Syntheses

The results of the exploratory regression model to identify any predictors of our primary outcome based on year, intervention category, and comparator category are displayed in the Appendix (Table 1).

4. Discussion

4.1. Summary of findings

Our large, recent, random sample of interventions used a transparent method for judging evidence quality, revealed that very few interventions evaluated within recent Cochrane Reviews are effective according to high quality evidence, and harms are rarely quantified. These results are consistent with previous estimates that less than half of medical interventions are supported by high quality evidence (8, 9, 22), and a recent study suggesting that only 10% of interventions have outcomes supported by high quality outcomes (14). The relative lack of evidence for harms in our study echoes a recent review suggesting that measurements of harms within systematic reviews is incomplete (23), despite Cochrane authors being encouraged to report harms as primary outcomes of reviews (24).

4.2. Limitations

Our review had several limitations. First, we relied on the work of the review authors, including their GRADE assessments, and views regarding whether the intervention was likely to be beneficial. To account for this potential limitation, we checked whether the Cochrane Reviews were at a high risk of bias using ROBIS, and found that most Cochrane Reviews are reliable (although our sample, 35 out of 87, was small). We are also aware that GRADE is reliable among trained reviewers (25), and given Cochrane's endorsement of GRADE (24) it may be reasonable to assume that Cochrane authors do use GRADE reliably.

Second, while GRADE has the advantage of being more transparent and widely accepted than earlier methods for rating the quality of evidence, the GRADE standard could be too demanding. Closely related to this point, the risk of bias assessments which have a strong influence on GRADE ratings, could be too stringent (26). Healthcare practitioners and patients may be happy to use interventions whose effects are supported by moderate quality evidence. Resolving the issue of whether GRADE is too strict would require a more in-depth study (involving patients, practitioners, and policymakers) of the usefulness of GRADE for

clinical decision-making. The results of such a study could potentially result in a radical overhaul of the GRADE system. We also note that only 26% of interventions had effects that were supported by moderate quality evidence according to GRADE. So, even if GRADE is too strict, only a minority of healthcare interventions have high or moderate quality evidence to support their benefits.

Third, healthcare interventions evaluated in recent Cochrane reviews may not be representative. Older interventions could be more likely to be effective and dramatically effective treatments could be less likely to be tested in systematic reviews (10). That being said, dramatically effective interventions are very rare (27). Moreover, our exploratory analysis did not detect statistically significant differences between the proportion of high-quality primary outcomes in older reviews (2008—2014) compared with more recent reviews (2015—2021) (see Appendix). Also mitigating this point, Cochrane reviews may address narrower questions than are typically asked in routine clinical practice. If so, then what clinicians do in clinical practice could be even less supported by high-certainty evidence then suggested by this study.

Fourth, we limited our analysis to the first listed primary outcome and up to three additional high-quality outcomes, and the first listed harms. Had we expanded our analysis to include all primary outcomes (or even all outcomes), we might have found some additional high-quality outcomes. However, only an additional 2% of interventions had high quality outcomes that were not listed first. Therefore, it is unlikely that our results would change by much.

Fifth, it was sometimes difficult to determine whether a control intervention described as standard or usual care was, in fact, an active intervention based on the report. Control interventions described as standard care (comparisons in roughly a third of included interventions) sometimes included an active comparator, while in other cases was akin to no

or minimal treatment, and in many cases incomplete reporting made it difficult to distinguish. Reviewers therefore had to use their judgment. Given the proportion of studies with ambiguous descriptions of comparators, this limitation is unlikely to have influenced our main findings.

Sixth, the data in our review was not all extracted by two independent reviewers. We used two strategies to mitigate this. First, one reviewer (JH) checked a sample of 10% of the interventions extracted by all reviewers. Second, a machine algorithm was used to check several key domains for all included reviews, and discrepancies were resolved by a third independent reviewer. In short, this potential limitation is unlikely to have led to a substantive difference in our results.

Seventh, systematic reviewers may not necessarily explicitly endorse an intervention as effective, but it may still be effective. Systematic reviewers at Cochrane often avoid making recommendations. Then, statistical significance does not equate with clinical significance. To account for this, we also reported the number of interventions with a high quality outcome, and it was still a small minority of the interventions (10%).

Finally, clinicians at the coalface may be happy with moderate quality evidence. Our exploratory analysis found that 42.2% of interventions had at least one high *or* moderate quality outcome. This is far greater than the proportion of interventions that satisfy our primary outcome, but still half / two thirds of interventions only have low quality support for their effectiveness. Relatedly, due to the pre-defined and sufficiently broad scope of this review, we did not report the reasons for the GRADE ratings of lower quality evidence (inconsistency, risk of bias, etc.) Future studies may investigate the main reasons for GRADE downgrading.

4.3. Interpretation of findings

The dramatic nature of our findings implies that despite any limitations of our study, there is not good evidence regarding the benefits or harms of a great number of currently used interventions. Future studies should investigate the extent to which interventions tested in Cochrane reviews are representative and reliable, as well as whether the GRADE standard is unreasonably demanding. For the relative lack of evidence about harms, future studies should explore whether this is due primarily to absence of adequate data in primary studies, or whether Cochrane Review authors could be more inclusive in the reporting of harms. Some evidence suggests that both problems may exist (28-32).

5. Conclusion

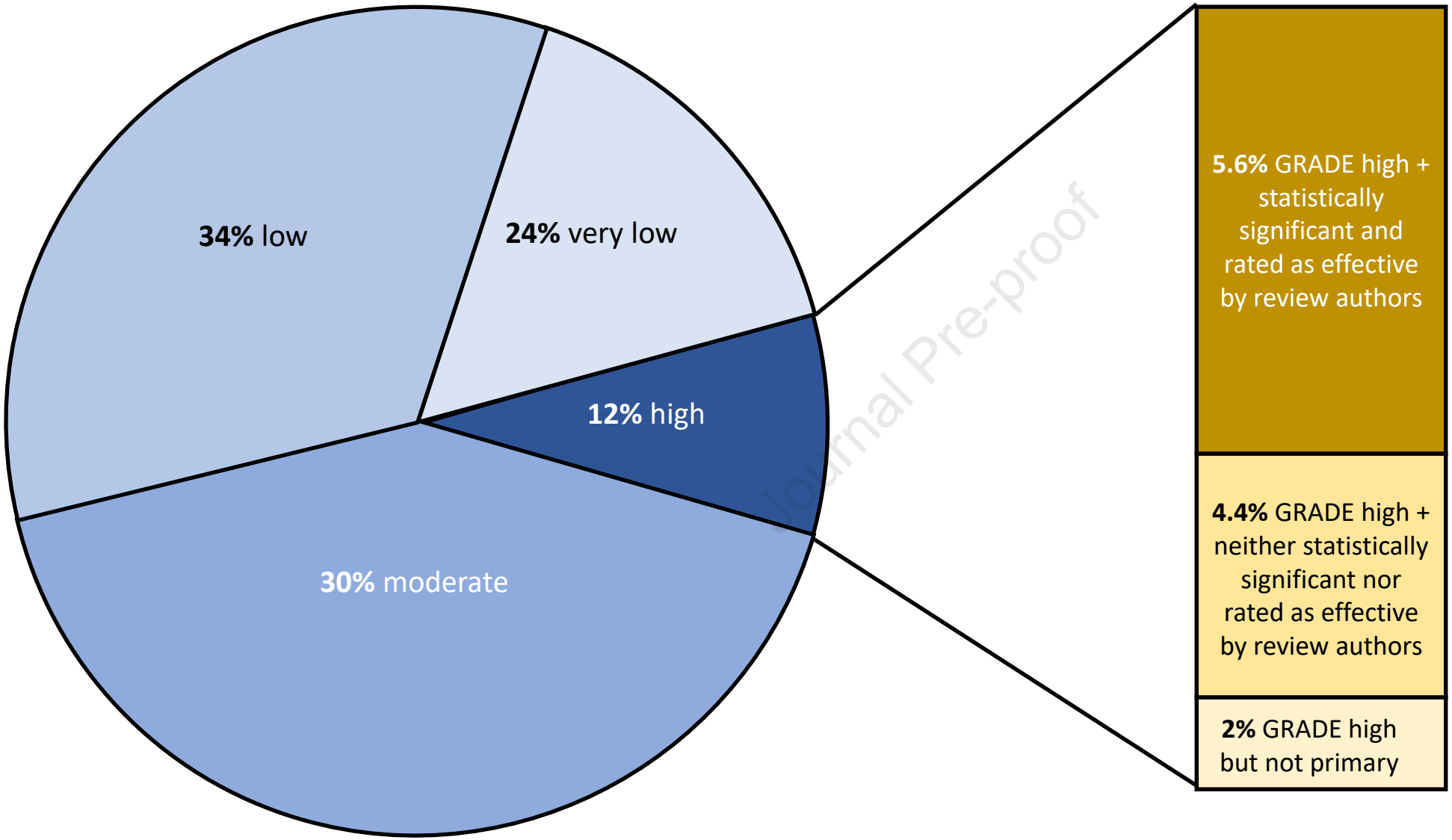
While many healthcare interventions may be beneficial, very few have high quality evidence to support their effectiveness and safety. This problem can be remedied by high quality studies in priority areas. These studies should measure harms more frequently and more rigorously. Practitioners and the public should be aware that many frequently used interventions are not supported by high quality evidence.

References

1. Moore TJ**.** Deadly Medicine: Why Tens of Thousands of Heart Patients Died in America's Worst Drug Disaster. New York ; London: Simon & Schuster; 1995.
2. Spock B, Fox D**.** Baby and child care. Illustrations by Dorothea Fox. (Enlarged, revised and updated ed. 159th printing.): New York: Pocket Books; London: New English Library; 1966.
3. Gilbert R, Salanti G, Harden M, See S**.** Infant sleeping position and the sudden infant death syndrome: systematic review of observational studies and historical review of recommendations from 1940 to 2002. International Journal of Epidemiology. 2005;34(4):874-87.
4. Jefferson T, Jones M, Doshi P, Spencer EA, Onakpoya I, Heneghan CJ**.** Oseltamivir for influenza in adults and children: systematic review of clinical study reports and summary of regulatory comments. BMJ. 2014;348:g2545.
5. Worrall J**.** *What* Evidence in Evidence-Based Medicine? Philosophy of Science. 2002;69(Supplement):S316-S30.

6. Tonelli MR**.** The limits of evidence-based medicine. Respir Care. 2001;46(12):1435-40; discussion 40-1.
7. McKeown T**.** The role of medicine : dream, mirage or nemesis? London: Nuffield Provincial Hospitals Trust; 1976.
8. El Dib RP, Atallah AN, Andriolo RB**.** Mapping the Cochrane evidence for decision making in health care. J Eval Clin Pract. 2007;13(4):689-92.
9. Garrow JS**.** What to do about CAM: How much of orthodox medicine is evidence based? BMJ. 2007;335(7627):951.
10. Howick J**.** The Philosophy of Evidence-Based Medicine. Oxford: Wiley-Blackwell; 2011.
11. Schünemann H, Brozek J, Oxman A, eds. GRADE handbook for grading quality of evidence and strength of recommendation; 2008.
12. Practice BB 2021;Pages. Accessed at BMJ Publishing Group Limited at https://bestpractice.bmj.com/info/toolkit/learn-ebm/what-is-grade/ on 27 August 2021.
13. Fleming PS, Koletsi D, Ioannidis JP, Pandis N**.** High quality of the evidence for medical and other health-related interventions was uncommon in Cochrane systematic reviews. J Clin Epidemiol. 2016;78:34-42.
14. Howick J, Koletsi D, Pandis N, Fleming PS, Loef M, Walach H, et al. The quality of evidence for medical interventions does not improve or worsen: a metaepidemiological study of Cochrane reviews. J Clin Epidemiol. 2020;126:154-9.
15. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JPA, et al. The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration. Annals of Internal Medicine. 2009;151(4):W65-W94.
16. Furey E 2021;Pages. Accessed at Calculator Soup at https://www.calculatorsoup.com
17. Systems CR 2012;Pages. Accessed at Creative Research Systems at https://www.surveysystem.com/sscalc.htm on 12 May 2021.
18. Mattick RP, Breen C, Kimber J, Davoli M**.** Buprenorphine maintenance versus placebo or methadone maintenance for opioid dependence. Cochrane Database Syst Rev. 2014(2):CD002207.
19. Sauer S 2021;Pages. Accessed at Sebastian Sauer at https://github.com/sebastiansauer/Cochrane-Parsing on 3 October 2021.
20. 2021;Pages. Accessed at Kleijnen Systematic Reviews at https://ksrevidence.com/ on November 4 2021.
21. Corporation S**.** Stata Statistical Software: Release 14. College Station, TX: StataCorp LP; 2015.
22. Ezzo J, Bausell B, Moerman DE, Berman B, Hadhazy V**.** Reviewing the reviews. How strong is the evidence? How clear are the conclusions? Int J Technol Assess Health Care. 2001;17(4):457-66.
23. Qureshi R, Mayo-Wilson E, Li T**.** Summaries of harms in systematic reviews are unreliable Paper 1: An introduction to research on harms. J Clin Epidemiol. 2021.
24. Higgins JJ, Thomas JC, Chandler J, M. C, Li T, Page MJ, et al. Cochrane Handbook for Systematic Reviews of Interventions version 6.0. Version 6.0 ed. Chichester: The Cochrane Collaboration; 2019.
25. Mustafa RA, Santesso N, Brozek J, Akl EA, Walter SD, Norman G, et al. The GRADE approach is reproducible in assessing the quality of evidence of quantitative evidence syntheses. J Clin Epidemiol. 2013;66(7):736-42; quiz 42 e1-5.
26. Langendam M, Carrasco-Labra A, Santesso N, Mustafa RA, Brignardello-Petersen R, Ventresca M, et al. Improving GRADE evidence tables part 2: a systematic survey of explanatory notes shows more guidance is needed. J Clin Epidemiol. 2016;74:19-27.
27. Glasziou P, Chalmers I, Rawlins M, McCulloch P**.** When are randomised trials unnecessary? Picking signal from noise. BMJ. 2007;334(7589):349-51.

28. Papanikolaou PN, Ioannidis JP**.** Availability of large-scale evidence on specific harms from systematic reviews of randomized trials. Am J Med. 2004;117(8):582-9.
29. Saini P, Loke YK, Gamble C, Altman DG, Williamson PR, Kirkham JJ**.** Selective reporting bias of harm outcomes within studies: findings from a cohort of systematic reviews. BMJ. 2014;349:g6501.
30. Hopewell S, Wolfenden L, Clarke M**.** Reporting of adverse events in systematic reviews can be improved: survey results. J Clin Epidemiol. 2008;61(6):597-602.
31. Pitrou I, Boutron I, Ahmad N, Ravaud P**.** Reporting of safety results in published reports of randomized controlled trials. Arch Intern Med. 2009;169(19):1756-61.
32. Zorzela L, Golder S, Liu Y, Pilkington K, Hartling L, Joffe A, et al. Quality of reporting in systematic reviews of adverse events: systematic review. BMJ. 2014;348:f7668.



Highlights

- **In this large sample** of 1567 interventions studied within Cochrane reviews, effects of most interventions (94%) interventions were not supported by high quality evidence.
- Potential harms of healthcare interventions were measured more rarely than benefits.
- Patients, doctors, and policy makers should consider the lack of high-quality evidence supporting the benefits and harms of many interventions in their decision-making.

Declaration of interest

None of the authors have any personal financial interests or professional relationships to disclose.