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.

What is new?

Key findings

- In this large sample of 1567 interventions studied within Cochrane reviews published since 2008, we found that the effectiveness of most (95%) interventions is not supported by high quality evidence.
- The safety of healthcare interventions is measured more rarely than benefits within Cochrane reviews.

What this adds to what was known?

- Previous estimates of the proportion of healthcare interventions that are not based on high-quality evidence are heterogeneous, out-of-date, use questionable methods to determine evidence-quality, or based on small samples.
- Our large sample is up-to-date, and based on the more widely accepted Grading of Recommendations Assessment, Development and Evaluation (GRADE) for rating evidence quality.

What is the implication and what should change now?

- The benefits and harms of healthcare interventions should be measured in higher quality evidence.
- Cochrane reviews should insist that intervention harms be measured more rigorously and consistently.
- Patients, doctors, and policy makers should consider the lack of high quality evidence to support the benefits and safety of many interventions in their decision making.

Abstract

Background

Previous estimates of the proportion of healthcare interventions that are not based on high-quality evidence are heterogeneous, out-of-date, or based on small samples. The objective of this study was to determine the proportion of healthcare interventions that are effective according to high-quality evidence.

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 healthcare interventions tested within Cochrane Reviews, whose effectiveness was based on high-quality evidence according to GRADE, had statistically significant positive (favourable) effects, and were judged to be effective by the review authors. We also calculated the proportion of healthcare interventions within Cochrane Reviews that suggested harm.

Results

A total of 1567 interventions were eligible for analysis. Eighty-seven (5.6%) of the interventions had a positive, statistically significant result and high quality first listed primary outcomes and were rated as being effective by review authors. Harms were measured for 577 (36.8%) of interventions, while evidence of statistically significant harm was substantiated

for 127 of those (127/1567; 8.1%). Our study was limited by potential unreliability of GRADE for determining the quality of evidence.

Conclusion

Most healthcare interventions studied within Cochrane Reviews do not have high quality evidence supporting their effectiveness, and often lack evidence of harms.

Key words

Evidence; systematic review; epidemiology; quality; safety

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 even 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). In another example, 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 oseltamivir for influenza were called into question by a systematic review of clinical study reports (4).

Critics of evidence-based medicine 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). Nevertheless, some scholars have argued that the increase in life span may be more due to general social, hygienic and economic progress than to the benefit of medical interventions (7).

Several meta-epidemiological studies have investigated the extent to which healthcare interventions are evidence-based, with varying results. A 2001 estimate identified about a quarter (22.5%) of healthcare interventions studied in 160 Cochrane Reviews to have good evidence of a positive effect (8). To reach their conclusion, two study authors were asked to independently rate the interventions on a 6-point scale ranging from "positive effect" to "harmful". In a study published in 2007, authors used a similar 6-point scale to rate interventions in a sample of 1016 Cochrane reviews (9). They concluded that 56% of

healthcare interventions were likely to be beneficial. Also published in 2007, a review claimed that of 2500 treatments that were supported by "good" evidence (defined as evidence from randomised trials), 22% were likely to be beneficial (10).

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

1.2.Objectives

The aim of this study is to overcome this gap in the literature and provide an up-to-date and comprehensive estimate of the proportion of healthcare intervention whose beneficial effects are supported by high quality evidence according to GRADE. We also aim 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). Also, the quality of evidence for effects was rated using GRADE. We excluded interventions within outdated versions of reviews that were superseded by newer versions, and interventions within withdrawn reviews.

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 between 1 January 2008 and 5 March 2021 (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 from the included studies for assessment of study quality and evidence synthesis. A single reviewer extracted data. We extracted information about the condition, the population (adults, children, mixed), included study designs (randomised trials or observational studies), intervention, and control (placebo/sham, usual care, no treatment). We also recorded the intervention category (pharmacological, psychological, surgical, behavioural, diet and exercise, manual therapies, alternative, other), as well as the type of outcomes (mortality; other objective outcomes

assessed with an instrument or prespecified measurable criteria; or subjective), and overall GRADE rating.

If an outcome of the intervention was rated as high quality according to GRADE, we recorded 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 positive interpretations include: “Buprenorphine should be supported as a medication to use” (18).

In cases where multiple Summary of Findings tables within the same review existed for the intervention’s primary outcome, we considered only the one listed first. In cases where no high-quality evidence was recorded for the first-listed primary outcome, we documented whether up to three other outcomes (primary or not) were rated as based on high-quality evidence. If not, we reported the highest GRADE rating for a primary outcome.

In addition, one author (JH) checked 10%, and a computer program using R (19) extracted data directly from Cochrane Reviews and checked all sampled reviews for the following domains: 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

The risk of bias within the systematic reviews was checked using the Kleijnen Systematic Review (KSR) Evidence database has ranked all Cochrane reviews dating back to

2015 using the ROBIS tool (20). We used the database to identify which of the high quality first reported primary outcomes were supported by reviews with a low risk of bias, methodologically.

2.6 Data Synthesis and Analysis

We produced descriptive statistics and proportions (n/N). Results were presented as absolute numbers and percentages, while also as effect sizes with the respective frequentist uncertainty estimators, the 95% confidence interval and p- values, where applicable. 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 being rated as high quality according to GRADE, bearing also a statistically significant effect and being effective according to review authors' interpretation.

2.7 Protocol Amendments

We initially intended to test for differences between the different Cochrane Review groups. However, there were too few reviews in certain groups and too large an amount of groups to allow for statistical comparisons. Instead, we presented the results of the primary outcome by review group, allowing for informal comparisons (Table 1). After the protocol was published, we decided to also use a computerized algorithm to explore how well a text mining program can be used to extract information and check human extraction. 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: n=46 had been superseded by more up-to-date version, n=112 were studied in reviews that had been withdrawn, n=684 interventions were compared with an active comparator, and n=875 did not include a GRADE assessment.

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, 56 (3.6%) exercise, 100 (6.4%) surgical, 62 (4.0%) diet, 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: n=708 (45.2%); usual or standard care: n=546 (34.8%); and no treatment: n= 313 (20.0%).

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 (1783, 97%) of the cases.

3.2.Quality of Evidence Supporting Intervention Effects

One in 10 (n=158 of N=1567) interventions had a first listed primary outcome rated as high quality according to GRADE. Of these, 106 (6.8% of the whole sample) also had a positive, statistically significant result, while 87 (5.6%) had a high-quality outcome,

statistically significant result, and were rated by review authors as being at least very likely to be effective (see Table 1). Breakdown of those 87 interventions across intervention category, type of comparator treatment, type of outcome, and target population is presented in Table 2. An additional 31 interventions (2.0%) had at least one (primary but not first reported) high quality outcome. Among the interventions that did not have a first listed or other high quality primary outcome, the highest GRADE rating for any outcome (first listed primary or not) was moderate in 472 (30.1%), low in 533 (34.0%), and very low in 373 (23.8%). Our primary outcome was met more often by pharmacological interventions (73%) than other intervention categories. However, the majority (52.3%) of the interventions were pharmacological anyhow, although in lower proportion (see Table 2).

3.3.Quality of Evidence Supporting Intervention Harms

Harms of 577 out of 1567 (36.8%) interventions were quantified in some way. 40 (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 adverse outcomes. Out of the 577 interventions that reported harms, there was evidence about a statistically significant effect of harm in 127 (8.1% of the whole sample). 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 rating is shown in Table 3. One fifth of those (18/127; 14.2%) were supported by high quality evidence as per GRADE. Outside pharmacological interventions, harms were not likely 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 that met the primary outcome (high quality, statistically significant effect, and authors interpret as effective)

	All sample of included interventions	High quality GRADE	High quality GRADE+ statistically significant effect	High quality GRADE+ statistically significant effect+ authors' interpretation as effective
Intervention Category	**n (%)**	**n (%)**	**n (%)**	**n (%)**
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%)
Total	**1567 (100.0%)**	**158 (100.0%)**	**106 (100.0%)**	**87 (100.0%)**

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 exploratory regression model to identify any predictors of our primary outcome based on year, intervention category, and comparator category, did not identify evidence for important predictors, as no statistically significant associations were detected (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 and found that very few interventions within are effective according to high quality evidence, and harms are rarely reported rigorously. Our results are consistent with previous estimates that less than half of medical interventions are supported by high quality evidence (8-10), and a recent study suggesting that only 10% of interventions produce outcomes supported by high quality outcomes (14). The relative paucity of evidence for harms that we identified is echoes a recent review suggesting that measurements of harms within systematic reviews is incomplete (22).

4.2.Limitations

Our review had a number of limitations. First, our sample was limited to interventions that were recently studied in Cochrane reviews. Healthcare interventions evaluated in recent Cochrane reviews may not be representative of all healthcare interventions. It could be that older interventions are more likely to be effective, and that dramatically effective treatments do not need to be tested in systematic reviews (23). If so, our results may understate the proportion of healthcare interventions that are effective. That being said, dramatically effective interventions are very rare (24). Moreover, we are not aware of any evidence that GRADE ratings for older reviews would be higher. In fact, 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).

Second, we reported the first listed primary outcome and up to 3 additional high-quality outcomes, and the first listed harms. Interventions may have additional benefits and additional harms. However, few of these were supported by high-quality evidence and adding

more outcomes would not have changed our main conclusion. Moreover, measuring more outcomes would have increased the risk of false positive results.

Third, it was sometimes difficult to determine whether a control intervention described as standard or usual care was, in fact, an active intervention. In the included reviews, standard care (comparisons in roughly a third of included interventions) sometimes included a competing intervention, 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. This limitation, while it may have led to a different sample size, is unlikely to have influenced our main findings.

Fourth, while GRADE has the advantage of being more transparent and widely accepted than earlier methods for rating the quality of evidence, it is not unproblematic. The inter-rater reliability of GRADE is only high among trained users (25), and all users of GRADE in Cochrane Reviews may not be well trained.

Fifth, 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.

Sixth, 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%).

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. 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.
9. 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.
10. Garrow JS. What to do about CAM: How much of orthodox medicine is evidence based? BMJ. 2007;335(7627):951.
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. 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.

23. Howick J**.** The Philosophy of Evidence-Based Medicine. Oxford: Wiley-Blackwell; 2011.

24. Glasziou P, Chalmers I, Rawlins M, McCulloch P**.** When are randomised trials unnecessary? Picking signal from noise. BMJ. 2007;334(7589):349-51.

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.