Key Concepts for Informed Health Choices: A framework for enabling people to think critically about health claims

Version 2022

Andy Oxman, Iain Chalmers, and Astrid Dahlgren led the development of the original list of Informed Health Choices (IHC) Key Concepts, and they have amended it in the light feedback and suggestions since then. They were responsible for the final decisions about amendments and additions. Many other people have contributed to this work, including other members of the IHC Network and people from around the world with various kinds of expertise (see Acknowledgements).

Suggested citation: *Oxman AD, Chalmers I, Dahlgren A, and the Informed Health Choices Group. Key Concepts for assessing claims about treatment effects and making well-informed treatment choices (Version 2022). IHC Working Paper. 2022. http://doi.org/10.5281/zenodo.6611932*

Introduction

There are endless claims about treatments in the mass media, advertisements, and everyday personal communication. Some are true and some are false. Many are unsubstantiated: we do not know whether they are true or false. Unsubstantiated claims about the effects of treatments often turn out to be wrong. Consequently, people who believe and act on these claims suffer unnecessarily and waste resources by doing things that do not help and might be harmful, and by not doing things that do help.

In response to these challenges, we developed the Informed Health Choices Key Concepts as the first step in the Informed Health Choices (IHC) project, an initiative supported by the Research Council of Norway *[Informed Health Choices Group 2018]*. The aim of the IHC project and ongoing work by the IHC Network is to help people make informed health choices.

In this document, we use the term "**treatment**" to include any intervention (action) intended to improve health, including preventive, therapeutic and rehabilitative interventions, and public health or health system interventions. Although we have developed and framed the Key Concepts to address treatment claims, people in other disciplines may find them relevant; for example, for assessing claims about the effects of educational interventions or environmental measures *[Aronson 2019 , Muller 2020]*. Adaptations of the IHC Key Concepts for other disciplines can be found here.

The Informed Health Choices (IHC) Key Concepts

The concepts serve as the basis for developing learning resources to help people understand and apply the concepts when claims about the effects of treatments (and other interventions) are made, and when they make health choices *[Chalmers 2018]*. They are also the basis for an item bank of multiple-choice questions (the Claim Evaluation Tools item bank) that can be used for assessing people's ability to apply the IHC Key Concepts *[Austvoll-Dahlgren 2017 (RS)]*.

The IHC Key Concepts are principles for evaluating the trustworthiness of treatment claims, comparisons, and choices. The concepts can help people to:

1. Recognise when a **claim** about the effects of treatments has an untrustworthy basis
2. Recognise when evidence from **comparisons** of treatments is trustworthy and when it is not
3. Make well-informed **choices** about treatments

They can help anyone, not just researchers, to think critically about whether to believe a treatment claim and what to do. This is sometimes referred to as critical health literacy *[Chinn 2011 (OR), Nutbeam 2000 , Sørensen 2012 , Sykes 2013 (RS)]*. The Key Concepts are intended for people using research, not for doing research.

The concepts and the basis for the concepts tend to focus on ways in which claims and comparisons can be misleading, and choices can be misinformed. This is not because we underestimate the tremendous gains that have been made in health care based on appropriate use of research. Our aim is to promote healthy scepticism, not excessive scepticism, and to help people decide what to believe and what to do. How sure we can be about the effects of treatments varies, as does how sure we can be about the balance between the advantages and disadvantages of treatments. When teaching, learning, or using these concepts, it is important to bear this in mind. Nonetheless, many claims about the effects of treatments are not reliable.

The Covid-19 pandemic has highlighted both the importance of critical thinking about treatment claims and choices, and the value of fair comparisons of treatments and systematic reviews to inform decisions. The pandemic has been accompanied by an "infodemic" –too much information – including false or misleading information during a disease outbreak *[Pian 2021 (SR)]*. At the same time, the need for evidence to inform decisions has been addressed by thousands of randomized trials of vaccines and drug treatments that have identified both effective and ineffective treatments – albeit few randomized trials of non-pharmacological interventions (BESSI) [Glasziou 2021]. Hundreds of systematic reviews have been rapidly conducted and reported to summarise the available evidence, but often with important limitations [Abbott 2021 (SR)].

This response to the pandemic has been inspiring, sometimes disappointing, and often frightening. It has been inspiring because of the rapid response and gains made by appropriate use of research. It has been disappointing because people sometimes have not used or benefited from this research and because of inadequate efforts to address important questions about non-pharmacological treatments. It has been frightening because of the sometimes overwhelming amount of misleading information and the impacts that has had on people's lives. Broader understanding and application of the Key Concepts could help people to navigate the infodemic, to avoid being misled, and to capitalise on the benefits of fair comparisons of treatments and reliable systematic reviews.

How we developed this list of Key Concepts (Methods)

We started to develop this list of concepts in 2013. We developed the IHC Key Concepts by searching the literature and checklists written for the public, journalists, and health professionals; and by considering concepts related to assessing the certainty of evidence about the effects of treatments. The first version included 32 concepts in six groups [Austvoll-Dahlgren 2015 (RS)]. We revised the Key Concepts yearly between 2015 and 2019. The main changes were as follows:

- The 2016 version was revised to include 34 concepts organised in three groups [Austvoll-Dahlgren 2016].
- The 2017 version was revised to include 36 concepts [Austvoll-Dahlgren 2017].
- The 2018 version was revised to include 44 concepts [Oxman 2018].
- The 2019 version was revised to include 49 concepts, we reorganised the concepts within each of the three main groups, and we added three subgroups to each of the three main groups [Oxman 2019].

A more detailed description of the methods that we have used can be found here: [Oxman 2019]. The changes that we made to the IHC Key Concepts were based on yearly review of feedback and suggestions we received, learning from using the IHC Key Concepts [Chalmers 2018 , Oxman 2019 (RS)], adaptation of the IHC Key Concepts by other disciplines [Aronson 2019], and a systematic review of other relevant frameworks [Oxman 2020 (SR)]. We have used four criteria in deciding on changes to the list of concepts. New Key Concepts must:

- be within the scope of the IHC Key Concepts – standards for judgment, or principles for evaluating the trustworthiness of treatment claims and treatment comparisons (research) used to support claims, and to inform treatment choices,
- address ways in which treatment claims and comparisons are frequently misleading or ways in which poorly informed decisions are taken,
- be useful for people without a research background to use research, not just for researchers or for doing research, and
- overlap as little as possible with other Key Concepts

There is, perhaps inevitably, some overlap among the Key Concepts, although we have tried to minimise this. For example, the concepts about a lack of evidence, misinterpretation of p-values, and "statistical significance" are closely related (see Concepts 2.3d, 2.4c, and 2.4d). When relevant, we have included links from the basis for one concept to the basis for another concept to clarify these relationships.

We have received only a few suggestions since the 2019 version was published and did not publish a new version of the Key Concepts in 2020 or 2021. We have decided that the version presented here will be the last revision made by us. This does not mean that this list of concepts cannot be further improved, but we will leave any further development of the IHC Key Concepts to others.

We have tried to include all concepts that are important for people to consider when they are assessing treatment claims and making health choices. We have tried to limit the number of concepts by minimising redundancy, and we have organised the concepts into three thematic groups (claims, comparisons, and choices). The concepts are not organised based on how complex or difficult they are to understand and apply, or in an order in which they should be learned.

We have written the concepts and explanations in plain language. We have also included a glossary with plain language definitions and explanations of some health research terms that we have used. Even so, some of the concepts presented may be unfamiliar and difficult to understand. The list is not designed as a teaching tool, and it is not intended to be read from beginning to end. It is a framework, or starting point, for teachers, journalists, and other intermediaries between people and health research to identify and develop resources (such as longer explanations, examples, games and interactive applications) to help people understand and apply the concepts.

The basis for each concept

In this update, we started with the explanations and implications from the 2019 version of the IHC Key Concepts *[Oxman 2019]*. For each concept, we have provided one or more examples to illustrate each explanation, and the basis for each concept, drawing on relevant research that informed the development of the IHC Key Concepts.

Whenever possible, we have referenced systematic reviews that provide a basis for a concept (see Concept 2.2a). We started with reviews with which we were familiar, including some that we had co-authored. Additional systematic reviews were identified by searching and screening the following sources:

- All Cochrane methodology reviews *[Clarke 2008]* (n = 36, on 29 June 2021)
- Epistemonikos using the terms “methodology review” OR “meta-epidemiological” in the title or abstract (n = 161, on 11 October 2021)
- PubMed using the terms “methodology review” OR “meta-epidemiological” (n = 193, on 11 October 2021)
- Google Scholar using the terms “methodology review” OR “meta-epidemiological” in the title, restricted to “Review articles” (n = 370, on 11 October 2021)

In addition, we searched Epistemonikos, PubMed, and Google Scholar for systematic reviews that support the explanation for each concept using key terms from the explanation or related terms.

The basis for each concept is presented logically and coherently rather than as a list of systematic reviews. We also have referenced some other reviews (that do not describe the methods that were used to prepare them), research studies, and other references. To clarify the nature of the evidence that we reference, we have used the codes below to categorise and label references as systematic reviews, other reviews of research studies, and research studies. Other references are not labelled. We have grouped the references in those four categories after the Implications for each concept.

(SR) **Systematic reviews**, including systematic reviews of methodological studies and overviews of reviews, such as Cochrane methodology reviews (e.g., *[Vist 2008 (SR)]*, methodological studies based on a systematic review of research (“meta-epidemiological” studies) (e.g., *[Ioannidis 2005 (SR), Nagendran 2016 (SR)]*, and systematic reviews of treatment effects (e.g., *[Gilbert 2005 (SR)]*). We categorised as systematic reviews any review of research that included a methods section that described the search strategy for finding studies and selection criteria.

(OR) **Other reviews** of research studies that address a specific question or topic and did not describe the methods that were used (e.g., *[Aronson 2020 (OR), Chinn 2011 (OR)]*).

(RS) **Research studies** (e.g., *[Austvoll-Dahlgren 2015 (RS), Frosch 2007 (RS), Sykes 2013 (RS)]*). We categorised as research studies any study that had a methods section describing how data were collected and analysed.

Other publications without a methods section do not have a code. This includes commentaries (e.g., *[Aronson 2019 , Berndt 2005]*), guides (e.g., *[Guyatt 2011a]*), analyses (e.g., *[Glasziou 2007]*), and books (e.g., *[Baron 2008]*).

Evidence and other references that provide the basis for each concept are referenced under the heading “Basis”. Only references for the examples that are used in the explanation are referenced under the heading “Explanation”.

One of us (AO) drafted the text, searched for, screened, and categorised the references. The other two (AD and IC) reviewed the text and references. All three agreed on the final version.

Organisation of the concepts

The current version includes the same concepts as the 2019 version. We have incorporated some suggestions in the explanations for a few concepts. We have also reorganised the concepts into four subgroups (high-level concepts) within each of the first two main groups (claims and comparisons) and into two subgroups within the third main group (choices) (Table 1).

We did this to make the organisation of the concepts more logical and the long list of concepts in some of the subgroups less overwhelming. The subgroups provide a more transparent logic for organising the concepts in each of the three main groups. The 10 high-level concepts also make it easier to get the gist of the concepts and makes the list for some of the subgroups less overwhelming and easier to remember. Table 2 is an overview of the 49 concepts in the three main groups and 10 subgroups.

Table 1. Ten high-level concepts within three main groups

1. Claims *Claims about effects that are not supported by evidence from fair comparisons are not necessarily wrong, but there is an insufficient basis for believing them.*	**2. Comparisons** *Studies should make fair comparisons, designed to minimise the risk of systematic errors (biases) and random errors (the play of chance).*	**3. Choices** *What to do depends on judgements about a problem, the relevance of the available evidence, and the balance of expected benefits, harms, and costs.*
1.1 Assumptions that treatments are safe or effective can be misleading. **1.2 Seemingly logical assumptions about *research* can be misleading.** **1.3 Seemingly logical assumptions about *treatments* can be misleading.** **1.4 Trust based on the source of a claim alone can be misleading.**	**2.1 Comparisons of treatments should be fair.** **2.2 Reviews of the effects of treatments should be fair.** **2.3 Descriptions of effects should clearly reflect *the size of the effects*.** **2.4 Descriptions of effects should reflect *the risk of being misled by the play of chance*.**	**3.1 Evidence should be relevant.** **3.2 Expected advantages should outweigh expected disadvantages.**

Table 2. Overview of the IHC Key Concepts

1. Claims *Claims about effects that are not supported by evidence from fair comparisons are not necessarily wrong, but there is an insufficient basis for believing them.*	**2. Comparisons** *To identify treatment effects, studies should make fair comparisons, designed to minimise the risk of systematic errors (biases) and random errors (the play of chance).*	**3. Choices** *What to do depends on judgements about a problem, the relevance of the available evidence, and the balance of expected benefits, harms, and costs.*
1.1 Assumptions that treatments are safe or effective can be misleading. Do not assume that a) treatments are safe, b) treatments have large, dramatic effects, c) treatment effects are certain, d) it is possible to know who will benefit and who will be harmed, or e) comparisons are not needed. **1.2 Seemingly logical assumptions about *research* can be misleading.** Do not assume that a) a plausible explanation is sufficient, b) association is the same as causation, c) more data is better data, d) a single study is sufficient, or e) fair comparisons are not applicable in practice. **1.3 Seemingly logical assumptions about *treatments* can be misleading.** Do not assume that a) treatment is needed, b) more treatment is better, c) a treatment is helpful or safe based on how widely used it is or has been, d) a treatment is better based on how new or technologically impressive it is, or e) earlier detection of 'disease' is better. **1.4 Trust based on the source of a claim alone can be misleading.** Do not assume that a) personal experiences alone are sufficient, b) your beliefs are correct, c) opinions alone are sufficient, d) peer review and publication is sufficient, or e) there are no competing interests.	**2.1 Comparisons of treatments should be fair.** Consider whether a) the people being compared were similar, b) the people being compared were cared for similarly, c) the people being compared knew which treatments they received, d) outcomes were assessed similarly in the people being compared, e) outcomes were assessed reliably, f) outcomes were assessed in all (or nearly all) the people being compared, and g) people's outcomes were analysed in the group to which they were allocated. **2.2 Reviews of the effects of treatments should be fair.** Consider whether a) systematic methods were used, b) unpublished results were considered, c) treatments were compared across studies, and d) important assumptions were tested. **2.3 Descriptions of effects should clearly reflect *the size of the effects*.** Be cautious of a) verbal descriptions alone of the size of effects, b) relative effects of treatments alone, c) average differences between treatments, and d) lack of evidence being interpreted as evidence of "no difference". **2.4 Descriptions of effects should reflect *the risk of being misled by the play of chance*.** Be cautious of a) small studies, b) results for a selected group of people within a study, c) p-values, and d) results reported as "statistically significant" or "non-significant".	**3.1 Evidence should be relevant.** a) Be clear about what the problem or goal is and what the options are. Consider the relevance of b) the outcomes measured in the research, c) fair comparisons in laboratories, animals, or highly selected people, d) the treatments that were compared, and e) the circumstances in which the treatments were compared. **3.2 Expected advantages should outweigh expected disadvantages.** a) Weigh the benefits and savings against the harms and costs of acting or not. Consider b) the baseline risk or severity of the symptoms when estimating the size of expected effects, c) how important each advantage and disadvantage is when weighing the pros and cons, d) how certain you can be about each advantage and disadvantage, and e) the need for further fair comparisons.

Competences and dispositions

In 2018, in addition to modifying the Key Concepts, we added lists of competences (required skills, knowledge, or capacity to do something) and dispositions (frequent and voluntary habits of thinking and doing) for thinking critically about treatments. The competences needed to achieve that goal (Table 3) and the dispositions (Table 4) are unchanged from the 2019 version.

Table 3. Goals, Competences and Dispositions for Informed Health Choices

Goal

To enable people to make good decisions* about which claims to believe about the effects of things they can do for their health, the health of others or for other reasons, and about what to do to achieve their goals.

Competences

To achieve this goal, people should be able to:

1) **Recognise when a claim has an untrustworthy basis by**
   a) recognising claims about the effects of treatments
   b) questioning the basis for treatment claims
   c) thinking carefully about treatment claims before believing them
   d) recognising when a treatment claim is relevant and important, and warrants reflection

2) **Recognise when evidence used to support a treatment claim is trustworthy or untrustworthy by**
   a) recognising the assumptions, evidence, and reasoning used to support treatment claims
   b) recognising fair and unfair treatment comparisons
   c) recognising reliable and unreliable summaries of treatment comparisons
   d) recognising important assumptions in summaries of treatment comparisons
   e) recognising misleading ways of presenting treatment effects
   f) understanding how systematic errors (the risk of bias), random errors (the play of chance), and the relevance (applicability) of treatment comparisons can affect the degree of confidence in estimates of treatment effects
   g) understanding the extent to which evidence does or does not support a treatment claim

3) **Make well-informed decisions about treatments by**
   a) being aware of cognitive biases when making decisions
   b) clarifying and understanding the problem, options, and goals when making a decision
   c) recognising when decisions have irreversible consequences
   d) judging the relevance of evidence used to inform decisions about treatments
   e) weighing the advantages and disadvantages of treatments, considering the size of treatment effects, how important each outcome is, the costs, and the certainty of the evidence
   f) communicating with others about the advantages and disadvantages of treatments

4) **Reflect on their competences and dispositions by**
   a) monitoring how they decide which treatment claims to believe and what to do
   b) monitoring how they adjust the processes they use to decide what to believe and do to fit the relevance, importance, and nature of different types of treatment claims and choices
   c) being aware of when they are making treatment claims themselves

* A good decision is one that makes effective use of the information available to the decision maker at the time the decision is made. A good outcome is one that the decision maker likes. The aim of thinking critically about treatments is to increase the probability of good outcomes (and true conclusions), but many other factors affect outcomes aside from critical thinking *[Baron 2008]*.

Table 4. Dispositions

People should be in the habit of thinking critically about:

1. **Claims, by**
    a) being aware of treatment claims (including those they make themselves) and choices
    b) questioning the basis for treatment claims
    c) being aware of cognitive biases and going from fast to slow thinking before forming an opinion about a treatment claim, making a claim, or taking a decision
    d) seeking evidence to reduce uncertainty when considering a relevant and important treatment claim or decision

2. **Evidence used to support claims, by**
    a) questioning the trustworthiness of evidence used to support treatment claims
    b) being alert to misleading presentations of treatment effects
    c) acknowledging and accepting uncertainty about the effects of treatments
    d) being willing to admit errors and modify judgements when warranted by evidence or a lack of evidence

3. **Choices, by**
    a) clarifying and understanding the problem, options, and goals when making decisions about treatments
    b) preferring evidence-based sources of information about treatment effects
    c) considering the relevance of the evidence used to inform decisions about treatments
    d) considering effect estimates, baseline risk, the importance of each advantage and disadvantage, the costs, and the certainty of the evidence when making decisions about treatments
    e) making informed judgements about the certainty of estimates of treatment effects
    f) making well-informed decisions
    g) Being aware of how people decide which treatment claims to believe and what to do

4. **People's own thinking, by**
    a) Being aware of how people decide which treatment claims to believe and what to do

Acknowledgements

We would like to thank Steven Woloshin for reviewing this update, Chui Hsia Yong for proofreading it, and Sarah Rosenbaum for formatting it. The following people have provided feedback and made helpful suggestions for improving the Key Concepts: Loai Albarqouni, Jeffrey Aronson, Douglas Badenoch, Jasmine Bampton, Eric Barends, Arild Bjørndal, Robert Boruch, Marnie Brennan, Hamish Chalmers, Yaolong Chen, Joe Chislett, David Colquhoun, Helen Cowan, Katherine Cowan, Sally Crowe, Peter Cunliffe-Jones, Mary Dixon-Woods, Klaus Eichler, Mike English, Adib Essali, Marie Gaarder, Abdul Ghaffar, Paul Glasziou, Claire Glenton, Metin Gülmezoglu, Andy Haines, Carl Heneghan, David Henry, Ann Blair Kennedy, Richard Lehman, Simon Lewin, Kamal Mahtani, Catherine Mathews, Robert Matthews, Brandy Maynard, Margaret McCartney, Angela Morelli, Paola Mosconi, Ray Moynihan, Lena Nordheim, Allen Nsangi, David Nunan, Kjetil Olsen, Newton Opiyo, Matt Oxman, Sarah Pannell, Anita Peerson, Matt Penfold, Livia Puljak, Andrew Pullin, Nicola Randall, Vanesa Ringle Hazel Roddam, Christopher Rose, Nick Ross, Gary Schwitzer, Selena Ryan-Vig, Sarah Rosenbaum, Anel Schoonees, Dan Semakula, Nelson Sewankambo, Jonathan Sharples, Paulina Stehlik, Ruth Stewart, Janet Stott, Raymond Tallis, Nerys Thomas, Carole Torgerson, Lyndal Trevena, Luke Vale, Jimmy Volmink, Linda Ware, and Irena Zalarija.

References

Systematic reviews

Abbott R, Bethel A, Rogers M, Whear R, Orr N, Shaw L, et al. Characteristics, quality and volume of the first 5 months of the COVID-19 evidence synthesis infodemic: a meta-research study. BMJ Evid Based Med. 2021. https://doi.org/10.1136/bmjebm-2021-111710

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. Int J Epidemiol. 2005;34(4):874-87. https://doi.org/10.1093/ije/dyi088

Ioannidis JPA. Contradicted and initially stronger effects in highly cited clinical research. JAMA. 2005;294(2):218-28. https://doi.org/10.1001/jama.294.2.218

Nagendran M, Pereira TV, Kiew G, Altman DG, Maruthappu M, Ioannidis JP, et al. Very large treatment effects in randomised trials as an empirical marker to indicate whether subsequent trials are necessary: meta-epidemiological assessment. BMJ. 2016;355:i5432. https://doi.org/10.1136/bmj.i5432

Oxman AD, Garcia LM. Comparison of the Informed Health Choices Key Concepts Framework to other frameworks relevant to teaching and learning how to think critically about health claims and choices: a systematic review. F1000Res. 2020;9:164. https://doi.org/10.12688/f1000research.21858.1

Pian W, Chi J, Ma F. The causes, impacts and countermeasures of COVID-19 "Infodemic": A systematic review using narrative synthesis. Inf Process Manag. 2021;58(6):102713. https://doi.org/10.1016/j.ipm.2021.102713

Sørensen K, Van den Broucke S, Fullam J, Doyle G, Pelikan J, Slonska Z, et al. Health literacy and public health: a systematic review and integration of definitions and models. BMC Public Health. 2012;12:80. https://doi.org/10.1186/1471-2458-12-80

Vist GE, Bryant D, Somerville L, Birminghem T, Oxman AD. Outcomes of patients who participate in randomized controlled trials compared to similar patients receiving similar interventions who do not participate. Cochrane Database Syst Rev. 2008(3):MR000009. https://doi.org/10.1002/14651858.mr000009.pub4

Other reviews

Aronson JK, Green AR. Me-too pharmaceutical products: History, definitions, examples, and relevance to drug shortages and essential medicines lists. Br J Clin Pharmacol. 2020;86(11):2114-22. https://doi.org/10.1111/bcp.14327

Chinn D. Critical health literacy: a review and critical analysis. Soc Sci Med. 2011;73(1):60-7. https://doi.org/10.1016/j.socscimed.2011.04.004

Research studies

Austvoll-Dahlgren A, Oxman AD, Chalmers I, Nsangi A, Glenton C, Lewin S, et al. Key concepts that people need to understand to assess claims about treatment effects. J Evid Based Med. 2015;8(3):112-25. https://doi.org/10.1111/jebm.12160

Austvoll-Dahlgren A, Semakula D, Nsangi A, Oxman AD, Chalmers I, Rosenbaum S, et al. Measuring ability to assess claims about treatment effects: the development of the 'Claim Evaluation Tools'. BMJ Open. 2017;7(5):e013184. https://doi.org/10.1136/bmjopen-2016-013184

Frosch DL, Krueger PM, Hornik RC, Cronholm PF, Barg FK. Creating demand for prescription drugs: a content analysis of television direct-to-consumer advertising. Ann Fam Med. 2007;5(1):6-13. https://doi.org/10.1370/afm.611

Oxman AD, Chalmers I, Austvoll-Dahlgren A, Informed Health Choices Group. Key Concepts for assessing claims about treatment effects and making well-informed treatment choices. F1000Res. 2019;7:1784. https://doi.org/10.12688/f1000research.16771.2

Sykes S, Wills J, Rowlands G, Popple K. Understanding critical health literacy: a concept analysis. BMC Public Health. 2013;13:150. https://doi.org/10.1186/1471-2458-13-150

Other references

Aronson JK, Barends E, Boruch R, Brennan M, Chalmers I, Chislett J, et al. Key concepts for making informed choices. Nature. 2019;572(7769):303-6. https://doi.org/10.1038/d41586-019-02407-9

Austvoll-Dahlgren A, Chalmers I, Oxman AD, Informed Health Choices Group. Assessing claims about treatments effects: key concepts that people need to understand (Version 2016). 2016. http://doi.org/10.5281/zenodo.4746689

Austvoll-Dahlgren A, Chalmers I, Oxman AD, Informed Health Choices Group. Assessing claims about treatment effects: key concepts that people need to understand (Version 2017). 2017. http://doi.org/10.5281/zenodo.4746689

Baron J. Thinking and Deciding. 4th ed. Cambridge, UK: Cambridge University Press; 2008.

Berndt ER. To inform or persuade? Direct-to-consumer advertising of prescription drugs. N Engl J Med. 2005;352(4):325-8. https://doi.org/10.1056/nejmp048357

Chalmers I, Oxman AD, Austvoll-Dahlgren A, Ryan-Vig S, Pannell S, Sewankambo N, et al. Key Concepts for Informed Health Choices: a framework for helping people learn how to assess treatment claims and make informed choices. BMJ Evid Based Med. 2018;23(1):29-33. https://doi.org/10.1136/ebmed-2017-110829

Clarke M, Oxman AD, Paulsen E, Higgins JPT, Green S. Guide to the contents of a Cochrane methodology protocol and review. Cochrane Handbook for Systematic Reviews of Interventions version 5,0. 2008. https://training.cochrane.org/handbook/archive/v5.0.0/

Glasziou P, Chalmers I, Rawlins M, McCulloch P. When are randomised trials unnecessary? Picking signal from noise. BMJ. 2007;334(7589):349-51. https://doi.org/10.1136/bmj.39070.527986.68

Glasziou PP, Michie S, Fretheim A. Public health measures for covid-19. BMJ. 2021;375:n2729. https://doi.org/10.1136/bmj.n2729

Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol. 2011a;64(4):383-94. https://doi.org/10.1016/j.jclinepi.2010.04.026

Informed Health Choices Group. The Informed Healthcare Choices Group. Supporting informed healthcare choices in low-income countries – final report. 2018. http://doi.org/10.5281/zenodo.4748333

Muller L-M, Morris A, Sharples JM, Chislett J, Rose N, Chalmers H. How to assess claims about cognition and learning: The ACE Concepts. Impact J R Coll Teach. 2020;18:19. https://impact.chartered.college/article/how-to-assess-claims-cognition-learning-ace-concepts/

Nutbeam D. Health literacy as a public health goal: a challenge for contemporary health education and communication strategies into the 21st century. Health Promot Int. 2000;15(3):259-67. https://doi.org/10.1093/heapro/15.3.259

Oxman AD, Chalmers I, Austvoll-Dahlgren A, Informed Health Choices Group. Key Concepts for assessing claims about treatment effects and making well-informed treatment choices (Version 2018). 2018. http://doi.org/10.5281/zenodo.4746689

Oxman AD, Chalmers I, Dahlgren A, Informed Health Choices Group. Key Concepts for assessing claims about treatment effects and making well-informed treatment choices. Version 2019. IHC Working Paper. 2019. https://doi.org/10.5281/zenodo.4746689

Explanations, illustrative examples, bases, implications, and references

1. Claims

Claims about treatment effects that are not supported by evidence from fair comparisons are not necessarily wrong, but there is an insufficient basis for believing them.

1.1 It should not be assumed that treatments are safe or effective – or that they are not.

1.1a Do not assume that treatments are safe.

Explanation

People often exaggerate the benefits of treatments and ignore or downplay potential harms. However, few effective treatments are 100% safe. Similarly, people in need or desperation hope that treatments will work, and they may ignore potential harms – especially when reliable evidence of treatment effects is lacking. As a result, they may waste time, money, or both on treatments that have never been shown to be useful and may cause harm. The harm that is caused may be minor, but treatments also sometimes cause serious, irreversible harms, including death.

Even simple advice can sometimes cause serious harm. For example, in many countries, parents and health professionals were led to believe that babies should be put to sleep on their tummies, so that they would not choke if they puked. However, researchers looking into possible causes of unexplained deaths during infancy found that babies who had died were more likely to have been put to sleep on their tummies than babies who had survived infancy. Three times as many babies died suddenly and unexpectedly if they had been put to sleep on their tummies. Earlier recognition of the risks of putting babies to sleep on their tummies might have prevented over 10,000 infant deaths in the UK and at least 50,000 in Europe, the USA, and Australasia *[Gilbert 2005 (SR)]*.

Basis for this concept

Both patients and health professionals tend to overestimate the benefits and underestimate harms of treatments *[Hoffmann 2015 (SR), Hoffmann 2017 (SR), Rejas Bueno 2022 (RS)]*.

Most people are aware that surgery and medicines can have unwanted (adverse) effects as well as beneficial effects. Adverse (side) effects include everything from mild symptoms, like nausea, to serious effects, like heart attacks. Even new “me too” medicines that are very similar to other medicines known to be effective and acceptably safe can turn out to have unexpected, serious side effects *[Aronson 2020 (OR)]*. Herbal remedies, too, can have adverse side effects *[Lee 2016 (SR)]*.

Studies that show benefits, especially large benefits, are more likely to be noticed than studies that do not [Duyx 2017 (SR), Ioannidis 2005 (SR)]. Subsequent studies, which often contradict those studies or show smaller benefits, [Ioannidis 2005 (SR), Serra-Garcia 2021 (SR)], get less attention [Serra-Garcia 2021 (SR)]. Research reports commonly emphasise findings that suggest benefits, while ignoring other findings [Chiu 2017 (SR)]. Press releases are often designed to attract favourable media attention and news reports of those studies do the same [Yavchitz 2012 (RS)]. Most news reports about treatments mention at least one benefit, but less than half mention or adequately discuss harms [Oxman 2022 (SR)].

Harms are often poorly reported in treatment evaluations [Eidet 2020 (SR), Hodkinson 2013 (SR)], as well as in news reports. Advertisements are used to promote purchase of treatments. Even when advertisements are regulated, they emphasise benefits while information about potential harms is provided in fine print [Berndt 2005 , Frosch 2007 (RS), Woloshin 2001 (RS)]. Public health authorities, health services, and governments, whilst acting with good intentions, also sometimes emphasise potential benefits while ignoring or downplaying potential harms of behaviours that they believe to be beneficial, such as participating in screening programmes [Jørgensen 2004 (RS), McCartney 2010].

Implications

Always consider the possibility that a treatment may have harmful effects.

References

Systematic reviews

Chiu K, Grundy Q, Bero L. 'Spin' in published biomedical literature: a methodological systematic review. PLoS Biol. 2017;15(9):e2002173. https://doi.org/10.1371/journal.pbio.2002173

Duyx B, Urlings MJE, Swaen GMH, Bouter LM, Zeegers MP. Scientific citations favor positive results: a systematic review and meta-analysis. J Clin Epidemiol. 2017;88:92-101. https://doi.org/10.1016/j.jclinepi.2017.06.002

Eidet LM, Dahlgren A, Elvsåshagen M. Unwanted effects of treatments for depression in children and adolescents: a mapping of systematic reviews. BMJ Open. 2020;10(3):e034532. https://doi.org/10.1136/bmjopen-2019-034532

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. Int J Epidemiol. 2005;34(4):874-87. https://doi.org/10.1093/ije/dyi088

Hodkinson A, Kirkham JJ, Tudur-Smith C, Gamble C. Reporting of harms data in RCTs: a systematic review of empirical assessments against the CONSORT harms extension. BMJ Open. 2013;3(9):e003436. http://dx.doi.org/10.1136/bmjopen-2013-003436

Hoffmann TC, Del Mar C. Patients' expectations of the benefits and harms of treatments, screening, and tests: a systematic review. JAMA Intern Med. 2015;175(2):274-86. https://doi.org/10.1001/jamainternmed.2014.6016

Hoffmann TC, Del Mar C. Clinicians' Expectations of the Benefits and Harms of Treatments, Screening, and Tests: A Systematic Review. JAMA Intern Med. 2017;177(3):407-19. https://doi.org/10.1001/jamainternmed.2016.8254

Ioannidis JPA. Contradicted and initially stronger effects in highly cited clinical research. JAMA. 2005;294(2):218-28. https://doi.org/10.1001/jama.294.2.218

Lee JY, Jun SA, Hong SS, Ahn YC, Lee DS, Son CG. Systematic review of adverse effects from herbal drugs reported in randomized controlled trials. Phytother Res. 2016;30(9):1412-9. https://doi.org/10.1002/ptr.5647

Oxman M, Larun L, Gaxiola GP, Alsaid D, Qasim A, Rose CJ, et al. Quality of information in news media reports about the effects of health interventions: systematic review and meta-analyses. F1000Res. 2022;10:433. https://doi.org/10.12688/f1000research.52894.2

Serra-Garcia M, Gneezy U. Nonreplicable publications are cited more than replicable ones. Sci Adv. 2021;7(21):eabd1705. https://advances.sciencemag.org/content/advances/7/21/eabd1705.full.pdf

Other reviews

Aronson JK, Green AR. Me-too pharmaceutical products: History, definitions, examples, and relevance to drug shortages and essential medicines lists. Br J Clin Pharmacol. 2020;86(11):2114-22. https://doi.org/10.1111/bcp.14327

Research studies

Frosch DL, Krueger PM, Hornik RC, Cronholm PF, Barg FK. Creating demand for prescription drugs: a content analysis of television direct-to-consumer advertising. Ann Fam Med. 2007;5(1):6-13. https://doi.org/10.1370/afm.611

Jørgensen KJ, Gøtzsche PC. Presentation on websites of possible benefits and harms from screening for breast cancer: cross sectional study. BMJ. 2004;328(7432):148. https://doi.org/10.1136/bmj.328.7432.148

Rejas Bueno M, Bacaicoa López de Sabando A, Sánchez Robles GA. [Health professionals expectations' about the benefit of regular primary care interventions]. Aten Primaria. 2022;54(4):102235. https://doi.org/10.1016/j.aprim.2021.102235

Woloshin S, Schwartz LM, Tremmel J, Welch HG. Direct-to-consumer advertisements for prescription drugs: what are Americans being sold? Lancet. 2001;358(9288):1141-6. https://doi.org/10.1016/s0140-6736(01)06254-7

Yavchitz A, Boutron I, Bafeta A, Marroun I, Charles P, Mantz J, et al. Misrepresentation of randomized controlled trials in press releases and news coverage: a cohort study. PLoS Med. 2012;9(9):e1001308. https://doi.org/10.1371/journal.pmed.1001308

Other references

Berndt ER. To inform or persuade? Direct-to-consumer advertising of prescription drugs. N Engl J Med. 2005;352(4):325-8. https://doi.org/10.1056/nejmp048357

McCartney M. Selling health to the public. BMJ. 2010;341:c6639. https://doi.org/10.1136/bmj.c6639

1.1b Do not assume that treatments have large, dramatic effects.

Explanation

Large effects (where everyone or nearly everyone who is treated experiences a benefit or a harm) are easy to detect without fair comparisons. However, few treatments have effects that are so large that fair comparisons (designed to minimise the risk of being misled by systematic errors (biases) or the play of chance) are not needed. Treatments that do not have large, dramatic effects may be helpful, but fair comparisons are needed to determine how safe and helpful they are.

Some treatments have obvious effects. For example, if someone is bleeding and losing lots of blood, it is obvious that it is a good idea to stop the bleeding. However, most effective treatments do not have such obvious effects. For example, any effects of exercise or changes in diet on heart disease or cancer may occur only after many years. Some medicines and medical procedures have immediate and obvious effects, such as giving adrenaline to someone with a severe allergic reaction; transfusing blood to someone who has lost a lot of blood; or draining pus from a painful abscess. However, like changes in exercise or diet, any effects of most medicines and medical procedures do not have such easily observed or experienced effects by everyone who receives the medicine or procedure. This includes common medications used to prevent heart disease or strokes, such as medicines for high blood pressure or high cholesterol, which help some people but not everyone who takes them *[Leucht 2015 (SR)]*. It also includes treatments for cancer and pain, and complementary and alternative medicines, such as herbal remedies, public health measures (such as closing schools to reduce the spread of Covid-19), and changes in the ways healthcare is delivered or financed.

Basis for this concept

It has been suggested that carefully designed evaluations are not needed when the size of the treatment effect (the signal) is more than 10 times larger than the noise (what happens to people without treatment) *[Glasziou 2007]*. However, an analysis of drugs licensed despite a lack of evidence from randomized trials has suggested that it is not possible to identify a threshold above which beneficial effects are "dramatic", and that carefully designed evaluations are therefore not needed *[Hozo 2022 (RS)]*. Other factors need to be considered when deciding whether carefully designed evaluations are needed. Nonetheless, very large effects (more than ten-fold improvement or a 90% reduction in a bad outcome) are very uncommon. Even effects that are large, but not that large (a two-fold improvement or a 50% reduction in a bad outcome) are uncommon, and most of the time are found to be much smaller when assessed in subsequent evaluations *[Nagendran 2016 (SR), Oxman 2012a , Pereira 2012 (SR)]*.

Implications

Claims of large treatment effects are likely to be wrong. Expect treatments to have moderate, small, or trivial effects (wanted or unwanted), rather than dramatic effects. If estimates of treatment effects are not based on systematic reviews of fair comparisons of treatments, be sceptical about claims of small or moderate effects of treatments.

References

Systematic reviews

Leucht S, Helfer B, Gartlehner G, Davis JM. How effective are common medications: a perspective based on meta-analyses of major drugs. BMC Med. 2015;13:253. https://doi.org/10.1186/s12916-015-0494-1

Nagendran M, Pereira TV, Kiew G, Altman DG, Maruthappu M, Ioannidis JP, et al. Very large treatment effects in randomised trials as an empirical marker to indicate whether subsequent trials are necessary: meta-epidemiological assessment. BMJ. 2016;355:i5432. https://doi.org/10.1136/bmj.i5432

Pereira TV, Horwitz RI, Ioannidis JP. Empirical evaluation of very large treatment effects of medical interventions. JAMA. 2012;308(16):1676-84. https://doi.org/10.1001/jama.2012.13444

Research studies

Hozo I, Djulbegovic B, Parish AJ, Ioannidis JP. Identification of threshold for large (dramatic) effects that would obviate randomized trials is not possible. J Clin Epidemiol. 2022 (In Press, Journal Pre-proof, Available online 25 January). https://doi.org/10.1016/j.jclinepi.2022.01.016

Other references

Glasziou P, Chalmers I, Rawlins M, McCulloch P. When are randomised trials unnecessary? Picking signal from noise. BMJ. 2007;334(7589):349-51. https://doi.org/10.1136/bmj.39070.527986.68

Oxman AD. Improving the health of patients and populations requires humility, uncertainty, and collaboration. JAMA. 2012a;308(16):1691-2. https://doi.org/10.1001/jama.2012.14477

1.1c Do not assume that treatment effects are certain.

Explanation

Fair comparisons of treatments can provide a basis for confidence about the probability of beneficial and harmful effects of treatments. However, it is rarely, if ever, possible to be 100% certain about the size of treatment effects, or to predict exactly what will happen if a treatment is used. This is especially true for treatments that are intended to prevent adverse effects happening a long time in the future. Fair comparisons of such treatments are difficult because they entail following people up for a very long time and it is rarely possible to ensure that people adhere to the advice they are given. Consequently, claims about the effects of such treatments are often based on associations and belief in explanations of how the treatments work. Some people argue that there should be different standards for judgements about the trustworthiness of claims when fair comparisons are difficult. However, it can be lethal not to acknowledge and reduce important uncertainties, even when there is limited potential for doing so. It is also important to recognise that certainty about treatment effects can change as new information becomes available. This is especially true for new problems and treatments, such as treatments for Covid-19.

For example, at the start of the Covid-19 pandemic, little was known about the effects of measures to control it. However, in less than a year, over 2000 randomized trials were registered *[Dillman 2020 (SR)]*. Dexamethasone – an inexpensive and widely used medicine – was shown to reduce mortality among patients with severe Covid-19 disease *[Sterne 2020 (SR)]*. On the other hand, no evidence was found to justify the use of another inexpensive and widely used medicine – hydroxychloroquine – and it was found to have harmful effects *[Singh 2021 (SR)]*. At the same time, there have been very few reports of fair comparisons of measures to reduce the spread of Covid-19 (such as closing schools), and major uncertainties exist about the effects – wanted and unwanted – of these measures *[Haber 2021 (SR)]*.

Basis for this concept

Grading of Recommendations, Assessment, Development and Evaluations (GRADE) is a widely-used approach to making systematic and transparent judgements about the certainty of evidence and the strength of recommendations *[Alonso-Coello 2016 , Guyatt 2011a , Guyatt 2008b]*. *UpToDate*, a widely used electronic medical textbook, contains over 9,400 recommendations made using the GRADE approach *[Agoritsas 2017 (RS)]*. Half (50%) of those recommendations were based on low-certainty evidence, 40% on moderate-certainty evidence, and only 10% on high-certainty evidence *[Agoritsas 2017 (RS)]*. At least 16 other studies have assessed the availability of reliable evidence for decisions made by doctors in general practice and various specialties. The approach used to assess the certainty of the evidence in those 16 studies was less rigorous. It was found that there is “high quality” evidence for between 11% and 80% of common decisions made by doctors and patients in different specialties (median 48%) and “no substantial evidence” for between 2% and 53% of common decisions (median 19%) *[Jamtvedt 2015 (OR)]*.

Although the overall quality of evidence for complementary and alternative medicine is improving *[Bloom 2000 (SR)]*, the certainty of the evidence for most complementary and alternative treatments is low *[Cao 2015 (SR), Haller 2019 (SR), Houzé 2017 (SR), Hunt 2011 (SR), Meyer 2013 (SR), Millstine 2017 (OR)]*. The certainty of the evidence for most health system decisions is also low *[Ciapponi 2017 (SR), Herrera 2017 (SR), Pantoja 2017 (SR), Wiysonge 2017 (SR)]*.

Implications

It is important to recognise that there is some uncertainty about the effects of all treatments, and that there is likely to be more uncertainty about some types of treatments. Choices are still required but it is preferable to acknowledge, accept, and take account of uncertainty than to deny it and make misinformed and potentially dangerous decisions.

References

Systematic reviews

Bloom BS, Retbi A, Dahan S, Jonsson E. Evaluation of randomized controlled trials on complementary and alternative medicine. Int J Technol Assess Health Care. 2000;16(1):13-21. https://doi.org/10.1017/s0266462300016123

Cao H, Yang G, Wang Y, Liu JP, Smith CA, Luo H, et al. Complementary therapies for acne vulgaris. Cochrane Database Syst Rev. 2015;1:Cd009436. https://doi.org/10.1002/14651858.CD009436.pub2

Ciapponi A, Lewin S, Herrera CA, Opiyo N, Pantoja T, Paulsen E, et al. Delivery arrangements for health systems in low-income countries: an overview of systematic reviews. Cochrane Database Syst Rev. 2017;9:CD011083. https://doi.org/10.1002/14651858.cd011083.pub2

Dillman A, Zoratti MJ, Park JJH, Hsu G, Dron L, Smith G, et al. The Landscape of Emerging Randomized Clinical Trial Evidence for COVID-19 Disease Stages: A Systematic Review of Global Trial Registries. Infect Drug Resist. 2020;13:4577-87. https://doi.org/10.2147/IDR.S288399

Haber NA, Clarke-Deelder E, Feller A, Smith ER, Salomon J, MacCormack-Gelles B, et al. Problems with Evidence Assessment in Covid-19 Health Policy Impact Evaluation (PEACHPIE): a systematic strength of methods review. medRxiv. 2021:2021.01.21.21250243. https://doi.org/10.1101/2021.01.21.21250243

Haller H, Anheyer D, Cramer H, Dobos G. Complementary therapies for clinical depression: an overview of systematic reviews. BMJ Open. 2019;9(8):e028527. https://doi.org/10.1136/bmjopen-2018-028527

Herrera CA, Lewin S, Paulsen E, Ciapponi A, Opiyo N, Pantoja T, et al. Governance arrangements for health systems in low-income countries: an overview of systematic reviews. Cochrane Database Syst Rev. 2017;9:CD011085. https://doi.org/10.1002/14651858.cd011085.pub2

Houzé B, El-Khatib H, Arbour C. Efficacy, tolerability, and safety of non-pharmacological therapies for chronic pain: An umbrella review on various CAM approaches. Prog Neuropsychopharmacol Biol Psychiatry. 2017;79(Pt B):192-205. https://doi.org/10.1016/j.pnpbp.2017.06.035

Hunt K, Ernst E. The evidence-base for complementary medicine in children: a critical overview of systematic reviews. Arch Dis Child. 2011;96(8):769-76. https://doi.org/10.1136/adc.2009.179036

Meyer S, Gortner L, Larsen A, Kutschke G, Gottschling S, Gräber S, et al. Complementary and alternative medicine in paediatrics: a systematic overview/synthesis of Cochrane Collaboration reviews. Swiss Med Wkly. 2013;143:w13794. https://doi.org/10.4414/smw.2013.13794

Pantoja T, Opiyo N, Lewin S, Paulsen E, Ciapponi A, Wiysonge CS, et al. Implementation strategies for health systems in low-income countries: an overview of systematic reviews. Cochrane Database Syst Rev. 2017;9:CD011086. https://doi.org/10.1002/14651858.cd011086.pub2

Singh B, Ryan H, Kredo T, Chaplin M, Fletcher T. Chloroquine or hydroxychloroquine for prevention and treatment of Covid-19. Cochrane Database Syst Rev. 2021(2). https://doi.org//10.1002/14651858.CD013587.pub2

Sterne JAC, Murthy S, Diaz JV, Slutsky AS, Villar J, Angus DC, et al. Association between administration of systemic corticosteroids and mortality among critically ill patients with Covid-19: a meta-analysis. JAMA. 2020;324(13):1330-41. https://doi.org/10.1001/jama.2020.17023

Wiysonge CS, Paulsen E, Lewin S, Ciapponi A, Herrera CA, Opiyo N, et al. Financial arrangements for health systems in low-income countries: an overview of systematic reviews. Cochrane Database Syst Rev. 2017;9:CD011084. https://doi.org/10.1002/14651858.cd011084.pub2

Other reviews

Jamtvedt G, Klemp M, Mørland B, Nylenna M. Responsibility and accountability for well informed health-care decisions: a global challenge. Lancet. 2015;386(9995):826-8. https://doi.org/10.1016/s0140-6736(15)60855-8

Millstine D, Chen CY, Bauer B. Complementary and integrative medicine in the management of headache. BMJ. 2017;357:j1805. https://doi.org/10.1136/bmj.j1805

Research studies

Agoritsas T, Merglen A, Heen AF, Kristiansen A, Neumann I, Brito JP, et al. UpToDate adherence to GRADE criteria for strong recommendations: an analytical survey. BMJ Open. 2017;7(11):e018593. https://doi.org/10.1136/bmjopen-2017-018593

Other references

Alonso-Coello P, Schunemann HJ, Moberg J, Brignardello-Petersen R, Akl EA, Davoli M, et al. GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 1: Introduction. BMJ. 2016;353:i2016. https://doi.org/10.1136/bmj.i2016

Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol. 2011a;64(4):383-94. https://doi.org/10.1016/j.jclinepi.2010.04.026

Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ. 2008b;336(7650):924-6. https://doi.org/10.1136/bmj.39489.470347.ad

1.1d Do not assume that it is possible to know who will benefit and who will be harmed.

Explanation

For some kinds of health problems, fair treatment comparisons can be made by giving different treatments to a patient at different times, and then comparing the outcomes associated with each of the different treatment periods. These are called n-of-1 trials because they compare the effects of alternative treatments in one patient *[Guyatt 1990 (SR)]*. For example, n-of-1 trials have compared paracetamol to non-steroidal anti-inflammatory drugs (NSAIDs) in patients with osteoarthritis of the hip or knee *[Wegman 2003 (RS)]*. The results varied across patients. Most uncertainties about the effects of treatments cannot be compared in this way, however. For example, this person-specific approach cannot usually be used to compare a surgical treatment with a drug treatment.

Most treatment comparisons involve comparing similar groups of patients assigned to one among alternative treatments. Fair comparisons of treatments usually tell us what happened, on average, in groups of similar people. Usually, in a group of people who have used a treatment, some people benefit, some do not, and some may even be harmed. For example, the proportion of people who benefit from common pharmacological treatments varies from 1.5% – for aspirin to prevent serious vascular events (myocardial infarction, stroke, or vascular death) in people at high risk – to 58% – for proton pump inhibitors for relief of reflux oesophagitis *[Leucht 2015 (SR)]*. It is rarely possible to know in advance who will benefit from which treatment among alternatives, who will not benefit, or who will be harmed. Paradoxically, the only way to know whether “personalised medicine” – customising treatment for individuals – works is to test it in fair comparisons. Unless the customisation is 100% effective and 100% safe, it is still not possible to know in advance who will benefit from “personalised care” and who will not. Beyond n-of-1 trials, “personalised medicine” is not really personalised; it is simply an effort to identify subgroups of people who are most likely to benefit from specific treatments.

For example, HER2-positive breast cancer is when breast cancer cells have a protein receptor called HER2 (human epidermal growth factor receptor 2). About 20% of breast cancers are HER2-positive. Trastuzumab (Herceptin) and other monoclonal antibodies that block HER2 receptors to keep cancer cells from growing are used to treat HER2-positive breast cancer. So, those medicines are given to women with HER2-positive breast cancer and not to other women with breast cancer. However, not all women with HER2-positive breast cancer benefit from the medicine and some will experience serious harmful effects, such as congestive heart failure (CHF). For example, for women with breast cancer detected at an early stage who have a moderate risk of cancer recurrence or death in the next three years (30%) and a moderate risk of CHF (2%), only about 10% more women who take the medicine will benefit (experience disease-free survival) and about 8% more will be harmed (experience CHF) *[Moja 2012 (SR)]*. It is not possible to predict which among those women will benefit and which will be harmed.

Basis for this concept

“Personalised medicine” and “precision medicine” are sometimes used interchangeably. “Personalised medicine” is an older term. However, “personalised” may be misinterpreted to imply that treatments are developed uniquely for each individual *[National Research Council 2011]*. Although definitions of personalised and precision medicine vary, the aim is to improve decisions about treatments by using biological information and biomarkers to identify more precisely which (subgroups of) patients will benefit or which will be harmed by a treatment *[National Research Council 2011 , Schleidgen 2013 (SR)]*. This has the potential to increase the proportion of patients who benefit from a treatment or reduce the proportion who are harmed. Treatments very rarely have the same effect on everyone *[Glasziou 2007 , Leucht 2015 (SR),* *Nagendran 2016 (SR), Pereira 2012 (SR)]*. It will not be possible in the foreseeable future to know which individuals will benefit or which will be harmed within subgroups.

Observational, non-randomized studies, such as genetic association studies, can be used to develop hypotheses about new and clinically useful ways to group patients who may respond differently to a treatment. However, there are many ways to classify patients, and only some are useful. Moreover, most reported genetic associations, which could potentially be used to group people, are not reliable *[Dolan 2010 (SR), Ioannidis 2009 (SR), Köhler 2018 (SR)), Nair 2012 (SR), Richards 2009 (SR), Serghiou 2016 (SR), Staines-Urias 2012 (SR), Trifiletti 2017 (SR)]*. Therefore, the usefulness of grouping people based on genetic associations, or other factors, needs to be evaluated using randomized trials *[National Research Council 2011]*.

Personalised medicine has been portrayed as a revolution in health care *[Marcon 2018 (SR)]*. However, there is much uncertainty about the usefulness of most personalised medicine technologies *[Holmes 2009 (SR), Kasztura 2019 (SR), Plöthner 2016 (SR)]*.

Implications

Fair treatment comparisons provide the best basis for making well-informed decisions about treatments, but there is almost always some uncertainty about who will benefit, who will not, and who will be harmed.

References

Systematic reviews

Dolan SM, Hollegaard MV, Merialdi M, Betran AP, Allen T, Abelow C, et al. Synopsis of preterm birth genetic association studies: the preterm birth genetics knowledge base (PTBGene). Public Health Genomics. 2010;13(7-8):514-23. https://doi.org/10.1159/000294202

Guyatt GH, Keller JL, Jaeschke R, Rosenbloom D, Adachi JD, Newhouse MT. The n-of-1 randomized controlled trial: clinical usefulness. Our three-year experience. Ann Intern Med. 1990;112(4):293-9. https://doi.org/10.7326/0003-4819-112-4-293

Holmes MV, Shah T, Vickery C, Smeeth L, Hingorani AD, Casas JP. Fulfilling the promise of personalized medicine? Systematic review and field synopsis of pharmacogenetic studies. PLoS One. 2009;4(12):e7960. https://doi.org/10.1371/journal.pone.0007960

Ioannidis JP. Prediction of cardiovascular disease outcomes and established cardiovascular risk factors by genome-wide association markers. Circ Cardiovasc Genet. 2009;2(1):7-15. https://doi.org/10.1161/circgenetics.108.833392

Kasztura M, Richard A, Bempong NE, Loncar D, Flahault A. Cost-effectiveness of precision medicine: a scoping review. Int J Public Health. 2019;64(9):1261-71. https://doi.org/10.1007/s00038-019-01298-x

Köhler CA, Evangelou E, Stubbs B, Solmi M, Veronese N, Belbasis L, et al. Mapping risk factors for depression across the lifespan: An umbrella review of evidence from meta-analyses and Mendelian randomization studies. J Psychiatr Res. 2018;103:189-207. https://doi.org/10.1016/j.jpsychires.2018.05.020

Leucht S, Helfer B, Gartlehner G, Davis JM. How effective are common medications: a perspective based on meta-analyses of major drugs. BMC Med. 2015;13:253. https://doi.org/10.1186/s12916-015-0494-1

Marcon AR, Bieber M, Caulfield T. Representing a "revolution": how the popular press has portrayed personalized medicine. Genet Med. 2018;20(9):950-6. https://doi.org/10.1038/gim.2017.217

Moja L, Tagliabue L, Balduzzi S, Parmelli E, Pistotti V, Guarneri V, et al. Trastuzumab containing regimens for early breast cancer. Cochrane Database Syst Rev. 2012;2012(4):Cd006243. https://doi.org/10.1002/14651858.cd006243.pub2

Nagendran M, Pereira TV, Kiew G, Altman DG, Maruthappu M, Ioannidis JP, et al. Very large treatment effects in randomised trials as an empirical marker to indicate whether subsequent trials are necessary: meta-epidemiological assessment. BMJ. 2016;355:i5432. https://doi.org/10.1136/bmj.i5432

Nair VS, Maeda LS, Ioannidis JP. Clinical outcome prediction by microRNAs in human cancer: a systematic review. J Natl Cancer Inst. 2012;104(7):528-40. https://doi.org/10.1093/jnci/djs027

Pereira TV, Horwitz RI, Ioannidis JP. Empirical evaluation of very large treatment effects of medical interventions. JAMA. 2012;308(16):1676-84. https://doi.org/10.1001/jama.2012.13444

Plöthner M, Ribbentrop D, Hartman JP, Frank M. Cost-Effectiveness of Pharmacogenomic and Pharmacogenetic Test-Guided Personalized Therapies: A Systematic Review of the Approved Active Substances for Personalized Medicine in Germany. Adv Ther. 2016;33(9):1461-80. https://doi.org/10.1007/s12325-016-0376-8

Richards JB, Kavvoura FK, Rivadeneira F, Styrkársdóttir U, Estrada K, Halldórsson BV, et al. Collaborative meta-analysis: associations of 150 candidate genes with osteoporosis and osteoporotic fracture. Ann Intern Med. 2009;151(8):528-37. https://doi.org/10.7326/0003-4819-151-8-200910200-00006

Schleidgen S, Klingler C, Bertram T, Rogowski WH, Marckmann G. What is personalized medicine: sharpening a vague term based on a systematic literature review. BMC Med Ethics. 2013;14:55. https://doi.org/10.1186/1472-6939-14-55

Serghiou S, Kyriakopoulou A, Ioannidis JP. Long noncoding RNAs as novel predictors of survival in human cancer: a systematic review and meta-analysis. Mol Cancer. 2016;15(1):50. https://doi.org/10.1186/s12943-016-0535-1

Staines-Urias E, Paez MC, Doyle P, Dudbridge F, Serrano NC, Ioannidis JP, et al. Genetic association studies in pre-eclampsia: systematic meta-analyses and field synopsis. Int J Epidemiol. 2012;41(6):1764-75. https://doi.org/10.1093/ije/dys162

Trifiletti DM, Sturz VN, Showalter TN, Lobo JM. Towards decision-making using individualized risk estimates for personalized medicine: A systematic review of genomic classifiers of solid tumors. PLoS One. 2017;12(5):e0176388. https://doi.org/10.1371/journal.pone.0176388

Research studies

Wegman AC, van der Windt DA, de Haan M, Devillé WL, Fo CT, de Vries TP. Switching from NSAIDs to paracetamol: a series of n of 1 trials for individual patients with osteoarthritis. Ann Rheum Dis. 2003;62(12):1156-61. https://doi.org/10.1136/ard.2002.002865

Other references

Glasziou P, Chalmers I, Rawlins M, McCulloch P. When are randomised trials unnecessary? Picking signal from noise. BMJ. 2007;334(7589):349-51. https://doi.org/10.1136/bmj.39070.527986.68

National Research Council. Toward Precision Medicine: Building a Knowledge Network for Biomedical Research and a New Taxonomy of Disease. Washington, DC: National Academies Press; 2011.

1.1e Do not assume that comparisons are not needed.

Explanation

Unless a treatment is compared to something else, it is not possible to know what would happen without the treatment. This makes it difficult to attribute outcomes to the treatment. Whenever comparative terms are used to describe a treatment – for example, “faster relief” or “better” – ask “compared to what?”. Sometimes people argue that a fair treatment comparison is impossible because the treatment is ‘holistic’, ‘individualised’, or ‘complex’. However, as with any other treatment, claims about the effects of such treatments depend on the results of comparing them with one or more alternatives. How trustworthy those claims are depends on how fair the comparisons are.

For example, a television commentator in the U.S. reported that “Between late-December of 2020 and last month [April 2021], a total of 3,362 people apparently died after getting the Covid vaccine in the United States.” He exclaimed: “That is an average of roughly 30 people every day,” and he went on to suggest that the vaccine was killing people *[Qiu 2021]*. There are many problems with that claim, including the lack of a comparison – how many similar people who had not been vaccinated died or would have been expected to die? Given that over 250 million doses of Covid-19 vaccines had been administered at that time *[CDC 2021]*, and that old people and others with a high chance of dying were prioritised for vaccination, it would be surprising if some of those people did not die after receiving the vaccine. That does not mean the vaccine caused them to die. The U.S. Centers for Disease Control and Prevention (CDC) reported that there were 17 reported deaths per million vaccinated people (up to May 17, 2021) *[CDC 2021]*. The proportion of Americans who died from any cause in 2019 was 8,697 per million *[CDC 2020]*. That corresponds to an average of 7,821 people dying every day. Most of them probably drank some water before dying. So, you could say that 1000s of Americans apparently died every day after drinking water. That does not mean that drinking water caused them to die.

Basis for this concept

Descriptive studies, such as case reports and case series, do not include a comparison group. They can provide clues about causation that warrant further investigation, but they rarely provide a reliable basis for drawing conclusions about treatment effects *[Dalziel 2005 (SR), Grimes 2002]*.

Even when people make a claim about the effects of a treatment without saying what it has been compared with, there is nevertheless an implied comparison; there is an assumption about what would have happened without the treatment. Often, the implied comparison is how things were before the treatment. For example, people were alive before being vaccinated and dead after being vaccinated. The problem with such before-after comparisons is that we can only rarely be certain about what would have happened without the treatment *[Glasziou 2007]*. Before-after studies are simple, easy to conduct, and common, but there is a high risk that they will suggest treatment effects that differ from actual effects *[Ho 2018 (SR)]*. One type of before-after study uses “historical controls”. These studies compare people who received a new treatment with people treated in the past. In comparing the results of studies using historical controls to the results of studies using random controls (randomized trials) of the same treatments, 44 of 56 historical control studies (79%) found the treatment of interest better than the comparison treatment, but only 10 of 50 randomized trials (20%) yielded similar findings *[Sacks 1982 (SR)]*.

Implications

Always ask which comparisons provide the basis for claims about the effects of treatments. Claims that are not based on fair comparisons are not reliable.

References

Systematic reviews

Dalziel K, Round A, Stein K, Garside R, Castelnuovo E, Payne L. Do the findings of case series studies vary significantly according to methodological characteristics? Health Technol Assess. 2005;9(2):iii-iv, 1-146. https://doi.org/10.3310/hta9020

Ho AMH, Phelan R, Mizubuti GB, Murdoch JAC, Wickett S, Ho AK, et al. Bias in before–after studies: narrative overview for anesthesiologists. Anesth Analg. 2018;126(5):1755-62. https://doi.org/10.1213/ANE.0000000000002705

Sacks H, Chalmers TC, Smith H. Randomized versus historical controls for clinical trials. The American Journal of Medicine. 1982;72(2):233-40. https://doi.org/10.1016/0002-9343(82)90815-4

Other reviews

CDC. Selected adverse events reported after Covid-19 vaccination (updated May 18, 2021). Atlanta: Centers for Disease Control and Prevention; 2021. https://www.cdc.gov/coronavirus/2019-ncov/vaccines/safety/adverse-events.html

CDC, National Center for Health Statistics. Underlying Cause of Death 1999-2019 on Centers for Disease Control and Prevention WONDER Online Database, released in 2020. Data are from the Multiple Cause of Death Files, 1999-2019, as compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program. Atlanta: Centers for Disease Control and Prevention; 2020. http://wonder.cdc.gov/ucd-icd10.html

Other references

Glasziou P, Chalmers I, Rawlins M, McCulloch P. When are randomised trials unnecessary? Picking signal from noise. BMJ. 2007;334(7589):349-51. https://doi.org/10.1136/bmj.39070.527986.68

Grimes DA, Schulz KF. Descriptive studies: what they can and cannot do. Lancet. 2002;359(9301):145-9. https://doi.org/10.1016/S0140-6736(02)07373-7

Qiu L. No, Covid-19 vaccines are not killing more people than the virus itself. New York Times. May 7, 2021. https://www.nytimes.com/live/2020/2020-election-misinformation-distortions

1.2 Seemingly logical assumptions about *research* can be misleading.

1.2a Do not assume that a plausible explanation of how or why a treatment might work is a sufficient basis for a claim about treatment effects.

Explanation

Treatments that should work in theory often do not work in practice, or may turn out to be harmful. A plausible explanation of how or why a treatment might work does not prove that it actually does work, or that it is safe. For example, cutting someone to make them bleed (bloodletting) used to be a common treatment for lots of problems. People believed it would rid the body of "bad humours", which is what they thought made people sick. But bloodletting did not help. It even killed people, including George Washington, the first president of the United States *[Morens 1999]*. His doctors drained 40% of his blood to treat a sore throat!

A more recent theory was that operating on blocked tubes (arteries) that carry blood to the brain would stop damage to the brain (strokes). That makes sense, but when that theory was tested in a fair comparison, researchers found not only that it did not help, but that some people died from the surgery *[Powers 2011 (RS)]*.

Even if there is plausible evidence that a treatment works in ways likely to be beneficial, the size of any such treatment effect, and its safety, cannot be predicted. For example, most drugs in a class of heart medicines called beta-blockers have beneficial effects in reducing recurrence of heart attacks; but two drugs in the class – pronethalol and practolol – were taken off the market because of unanticipated side effects *[Furberg 1999]*. Similarly, it cannot be assumed that a treatment works or does not work based on the type of treatment. For example, it cannot be assumed that all complementary medicines or that all modern medicines do or do not work, or that all vaccines do or do not work. On the other hand, not understanding how a treatment works does not mean that it does not work.

Basis for this concept

Protocols for randomized trials of new treatments almost always have a rationale that includes an explanation of how or why the treatment might work. A systematic review of four cohorts of randomized trials including 743 trials involving almost 300,000 patients found that only slightly more than half of the new treatments were better than established treatments and few were substantially better, despite plausible explanations why they might be better *[Djulbegovic 2012 (SR)]*.

New medicines are developed based on an understanding of how and why they are expected to work, and many medicines do, in fact work. However, among 222 novel medicines that were found to be effective and were approved by the U.S. Federal Drug Administration (FDA) from 2001 to 2010, about one-fifth were found to have unanticipated serious adverse events after they had been approved *[Downing 2017 (RS)]*.

Homeopathy has been used for over 200 years, based on the theory that patients with signs and symptoms can be helped by a homeopathic remedy that produces those signs and symptoms in healthy individuals, and that homeopathic remedies retain biological activity after repeated dilution. But systematic reviews of the effects of homeopathy have found no condition that responds convincingly better to homeopathic treatment than placebo *[Ernst 2002 (SR), Jorgensen 2013 (SR)]*.

It is argued that the use of theory will lead to more effective behaviour change interventions. However, there are dozens of different theories to choose from *[Davis 2015 (SR)]*. Interventions to change health-related behaviours typically have modest effects, and systematic reviews of randomized trials of health behaviour change interventions have not found theory-based interventions to be more effective than non-theory-based interventions *[Dalgetty 2019 (SR)]*.

Implications

Do not assume that claims about the effects of treatments based on an explanation of how they might work are correct if the treatments have not been assessed in systematic reviews of fair comparisons of treatments.

References

Systematic reviews

Dalgetty R, Miller CB, Dombrowski SU. Examining the theory-effectiveness hypothesis: A systematic review of systematic reviews. Br J Health Psychol. 2019;24(2):334-56. https://doi.org/10.1111/bjhp.12356

Davis R, Campbell R, Hildon Z, Hobbs L, Michie S. Theories of behaviour and behaviour change across the social and behavioural sciences: a scoping review. Health Psychol Rev. 2015;9(3):323-44. https://doi.org/10.1080/17437199.2014.941722

Djulbegovic B, Kumar A, Glasziou PP, Perera R, Reljic T, Dent L, et al. New treatments compared to established treatments in randomized trials. Cochrane Database Syst Rev. 2012;10(10):Mr000024. https://doi.org/10.1002/14651858.mr000024.pub3

Ernst E. A systematic review of systematic reviews of homeopathy. Br J Clin Pharmacol. 2002;54(6):577-82. https://doi.org/10.1046/j.1365-2125.2002.01699.x

Jorgensen M, Applegarth K, Wong J, Campbell S. Effectiveness of homeopathy for clinical conditions: evaluation of the evidence: overview of systematic reviews prepared for the National Health and Medical Research Council. Canberra; 2013. https://www.researchgate.net/publication/281642270_Effectiveness_of_homeopathy_for_clinical_conditions_evaluation_of_the_evidence

Research studies

Downing NS, Shah ND, Aminawung JA, Pease AM, Zeitoun JD, Krumholz HM, et al. Postmarket safety events among novel therapeutics approved by the US Food and Drug Administration Between 2001 and 2010. JAMA. 2017;317(18):1854-63. https://doi.org/10.1001/jama.2017.5150

Powers WJ, Clarke WR, Grubb RL, Jr., Videen TO, Adams HP, Jr., Derdeyn CP. Extracranial-intracranial bypass surgery for stroke prevention in hemodynamic cerebral ischemia: the Carotid Occlusion Surgery Study randomized trial. JAMA. 2011;306(18):1983-92. https://doi.org/10.1001/jama.2011.1610

Other references

Furberg CD, Herrington DM, Psaty BM. Are drugs within a class interchangeable? Lancet. 1999;354(9185):1202-4. https://doi.org/10.1016/s0140-6736(99)03190-6

Morens DM. Death of a president. N Engl J Med. 1999;341(24):1845-9. https://doi.org/10.1056/nejm199912093412413

1.2b Do not assume that association is the same as causation.

Explanation

The fact that a possible treatment outcome (i.e. a potential benefit or harm) is associated with a treatment does not mean that the treatment caused the outcome. The association or correlation could instead be due to chance or some other underlying factor. For example, people who seek and receive a treatment may be healthier and have better living conditions than those who do not seek and receive the treatment. Therefore, people receiving the treatment might appear to benefit from the treatment, but the difference in outcomes could be because they are healthier and have better living conditions, rather than because of the treatment.

An obvious example of confusing an association with causation would be to assume that going to the doctor causes people to be sick because going to the doctor is associated with being sick. It is more likely that people went to the doctor because they were sick than that going to the doctor caused them to be sick. Another obvious example would be to assume that eating ice cream causes people to drown because ice cream sales are associated with drowning. A more likely explanation for that association is that when it is hot people eat more ice cream and they also swim more. In this example, hot weather is a confounder – it is associated with the "treatment" (eating ice cream) and it affects the "outcome" (the number of people who drown).

A less obvious example of confusing an association with causation was the assumption that hormone replacement therapy (HRT) prevented cardiovascular disease (CVD). For many years, experts and doctors believed that HRT reduced the risk of CVD, based on an association found in studies that compared women who chose to take HRT and women who did not. However, large, randomized trials showed no benefit or an increased risk of CVD in women assigned to HRT. An explanation for this is that socio-economic status was a confounder in the non-randomized studies. Women with lower socio-economic status are more likely to have CVD and they are less likely to take HRT. So, a reason for the association found in the non-randomized studies was the difference in socio-economic status between the comparison groups, not the difference in whether they took HRT or not *[Humphrey 2002 (SR)]*.

Basis for this concept

Researchers, press releases from universities and journal publishers, and news reports frequently use causal language when reporting associations found in non-randomized studies of treatments *[Lazarus 2015 (RS), Oxman 2022 (SR), Yu 2020 (RS)]*. This is likely to be misleading.

As illustrated by the examples above, there is a compelling logical basis for not assuming that an association between a treatment and an outcome means that the treatment caused the outcome. However, it is less clear how often assumptions about causation based on an association are wrong or when it is correct to assume that an association ***does*** mean that a treatment caused an outcome.

When there are very strong associations, it is very unlikely that they result from confounding *[Glasziou 2007]*. However, very strong associations are uncommon *[Nagendran 2016 (SR), Oxman 2012a , Pereira 2012 (SR)]*.

When there are not very strong associations, one way of assessing the likelihood of being misled by assumptions about causation based on an association is to compare associations found in non-randomized studies to the findings of randomized trials. Non-randomized studies can only adjust for potential confounders if these are known and have been measured. On the other hand, randomly assigning people to comparison groups in large, randomized trials tends to balance the distribution of both measured and unmeasured risk factors (potential confounders) across treatment comparison groups. How much is known about potential confounders and to what extent they have been measured varies. Often, what is known and measured is limited. For example, a systematic review of non-randomized studies published in major psychiatry journals found that confounding was widely ignored in interpreting the results *[Munkholm 2020 (SR)]*.

Some reviews of comparisons between the results of non-randomized studies and randomized trials have found important differences in results, while others have found little or no difference *[Anglemyer 2014 (SR), Rush 2018 (SR)]*. There are several possible reasons for these variable findings and why both randomized trials and non-randomized studies can either overestimate or underestimate the effects of treatments *[Kleijnen 1997 , Sterne 2016]*. This includes confounding that can occur after randomisation, particularly in trials that measure long-term effects of treatments *[Hernán 2013 , Manson 2016]*. So, it is difficult to draw firm conclusions about how often assumptions about causation based on an association observed outside the context of randomized trials are misleading.

Before assuming that an outcome associated with a treatment has been caused by the treatment, other reasons for an association should be considered in a systematic review of fair comparisons as the basis for judging the extent to which other reasons for an association have been ruled out *[Sterne 2016]*.

Implications

Do not assume that an outcome associated with a treatment was caused by the treatment unless other reasons for the association have been ruled out in a systematic review of fair comparisons.

References

Systematic reviews

Anglemyer A, Horvath HT, Bero L. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. Cochrane Database Syst Rev. 2014(4). https://doi.org//10.1002/14651858.MR000034.pub2

Humphrey LL, Chan BK, Sox HC. Postmenopausal hormone replacement therapy and the primary prevention of cardiovascular disease. Ann Intern Med. 2002;137(4):273-84. https://doi.org/10.7326/0003-4819-137-4-200208200-00012

Munkholm K, Faurholt-Jepsen M, Ioannidis JPA, Hemkens LG. Consideration of confounding was suboptimal in the reporting of observational studies in psychiatry: a meta-epidemiological study. J Clin Epidemiol. 2020;119:75-84. https://doi.org/10.1016/j.jclinepi.2019.12.002

Nagendran M, Pereira TV, Kiew G, Altman DG, Maruthappu M, Ioannidis JP, et al. Very large treatment effects in randomised trials as an empirical marker to indicate whether subsequent trials are necessary: meta-epidemiological assessment. BMJ. 2016;355:i5432. https://doi.org/10.1136/bmj.i5432

Oxman M, Larun L, Gaxiola GP, Alsaid D, Qasim A, Rose CJ, et al. Quality of information in news media reports about the effects of health interventions: systematic review and meta-analyses. F1000Res. 2022;10:433. https://doi.org/10.12688/f1000research.52894.2

Pereira TV, Horwitz RI, Ioannidis JP. Empirical evaluation of very large treatment effects of medical interventions. JAMA. 2012;308(16):1676-84. https://doi.org/10.1001/jama.2012.13444

Rush CJ, Campbell RT, Jhund PS, Petrie MC, McMurray JJV. Association is not causation: treatment effects cannot be estimated from observational data in heart failure. Eur Heart J. 2018;39(37):3417-38. https://doi.org/10.1093/eurheartj/ehy407

Research studies

Lazarus C, Haneef R, Ravaud P, Boutron I. Classification and prevalence of spin in abstracts of non-randomized studies evaluating an intervention. BMC Med Res Methodol. 2015;15:85. https://doi.org/10.1186/s12874-015-0079-x

Yu B, Wang J, Guo L, Li Y. Measuring correlation-to-causation exaggeration in press releases. Proceedings of the 28th International Conference on Computational Linguistics2020. p. 4860-72. http://dx.doi.org/10.18653/v1/2020.coling-main.427

Other references

Glasziou P, Chalmers I, Rawlins M, McCulloch P. When are randomised trials unnecessary? Picking signal from noise. BMJ. 2007;334(7589):349-51. https://doi.org/10.1136/bmj.39070.527986.68

Hernán MA, Hernández-Díaz S, Robins JM. Randomized trials analyzed as observational studies. Ann Intern Med. 2013;159(8):560-2. https://doi.org/10.7326/0003-4819-159-8-201310150-00709

Kleijnen J, Gøtzsche P, Kunz RA, Oxman AD, Chalmers I. So what's so special about randomisation? In: Maynard A, Chalmers I, editors. Non-Random Reflections on Health Care Research: On the 25th Anniversary Of Archie Cochrane's Effectiveness and Efficiency. London: BMJ Publishing Group; 1997. p. 93-106.

Manson JE, Shufelt CL, Robins JM. The potential for postrandomization confounding in randomized clinical trials. JAMA. 2016;315(21):2273-4. https://doi.org/10.1001/jama.2016.3676

Oxman AD. Improving the health of patients and populations requires humility, uncertainty, and collaboration. JAMA. 2012a;308(16):1691-2. https://doi.org/10.1001/jama.2012.14477

Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 2016;355:i4919. https://doi.org/10.1136/bmj.i4919

1.2c Do not assume that more data is better data.

Explanation

Claims that are based on "big data" (data from large databases) or "real world data" (routinely collected data) can be misleading. More data simply gives a more statistically precise estimate of whatever biases there might be in a treatment comparison that uses routinely collected data. When using routinely collected data, it is only possible to control for confounders that are already known and have been measured. Unfortunately, routinely collected data often do not include sufficient detail to confidently conclude that any association found between a treatment and an outcome means that the treatment caused the outcome.

For example, routinely collected (real world) data have been used in non-randomized comparisons of different types of coronary artery bypass surgery. Twelve studies including 34,019 patients used a non-randomized study design that is believed to reduce the risk of bias due to confounders (propensity-score matching) *[Gaudino 2018 (SR)]*. They found that using two internal thoracic arteries compared to using one artery was associated with a lower risk of dying within one year. A more likely explanation is that the association was because of confounders that had not been measured. Using two arteries instead of one increases the complexity and invasiveness of the surgery. It is likely that surgeons tend to reserve this type of surgery for patients perceived as healthier and expected to live longer. This type of bias in allocating patients to different treatments (e.g., based on the individual surgeon's judgement) is very difficult to quantify. The statistics can only be adjusted for the measured confounders *[Agoritsas 2017]*. As a further illustration of this problem, a large randomized trial found little or no difference in survival after 10 years. This contrasts with 14 non-randomized studies using propensity-score matching with 24,123 patients, which found that using two arteries improved survival compared to one artery *[Gaudino 2019 (SR)]*. This was due to both lower survival in patients in randomized trials, who were allocated to the two-artery group, and higher survival in the group allocated to the one-artery group compared to the studies using "real world data".

Describing routinely collected data as "real world data" implies that data collected in carefully designed fair comparisons of treatments do not come from the real world. Databases of routinely collected data may indeed include a broader spectrum of people than data collected in fair comparisons of treatments that have narrow eligibility criteria. However, routine collection of data is rarely planned to include the information that is needed to ensure fair comparisons, and randomized trials can be designed to have wide eligibility criteria.

Basis for this concept

A systematic review of studies that evaluated the effectiveness of treatments on mortality using propensity scores found that most of the studies explored effects of treatments that had already been compared in randomized trials *[Hemkens 2016b (SR)]*. The so-called "real world" studies seemed to have little impact.

Another systematic review compared treatment effects found in non-randomized studies using routinely collected ("real world") data and propensity-score matching with those found in randomized trials *[Hemkens 2016a (SR), Sterne 2018]*. The review found that the non-randomized studies using routinely collected data systematically and substantially overestimated mortality benefits of treatments compared with subsequent trials investigating the same question. This is consistent with the findings of another systematic review comparing studies using propensity score methods with randomized trials *[Dahabreh 2012 (SR)]*. A third systematic review compared treatment effects found in non-randomized studies using "real world data" with those found in randomized trials for mortality and other outcomes *[Ewald 2020 (SR)]*. That review did not find a systematic difference in treatment effects, but it found important differences, including effects going in the opposite direction for eight of the 19 included comparisons. A fourth systematic review of comparisons between non-randomized studies using real world data and randomized trials found only two substantial differences in treatment effects out of 15 comparisons *[Mathes 2021 (SR)]*.

As with comparisons of other types of non-randomized studies with randomized trials, there are many reasons why both randomized trials and non-randomized studies can either overestimate or underestimate the effects of treatments *[Anglemyer 2014 (SR), Goodman 2017 , Kleijnen 1997 , Mathes 2021 (SR), Sterne 2016]*. So, it is difficult to draw firm conclusions about how often the results of non-randomized studies using real world data will differ substantially from the results of randomized trials. However, evaluations of treatment effects using "real world data" are unlikely to be reliable if there are not high-quality data *[Bian 2020 (SR)]*, and important confounders have not been measured *[Franklin 2019]*.

The main argument for studies using "real world data" is that the findings of randomized trials are not applicable to the real world. However, the fourth systematic review of comparisons between non-randomized studies using real world data and randomized trials found little impact of factors related to the applicability of the findings on the estimated treatment effects *[Mathes 2021 (SR)]*. Limited applicability of the randomized trials was mostly due to the trials being designed to assess the effects of treatments under ideal circumstances ("explanatory studies") (e.g., having narrow eligibility criteria) rather than under normal, everyday circumstances ("pragmatic studies") *[Thorpe 2009]*.

The finding that the results of the randomized trials appeared to be largely applicable to the "real world" is consistent with the findings of a systematic review of studies comparing outcomes in patients receiving a treatment in a randomized trial to similar patients receiving the same treatment outside of a randomized trial *[Vist 2008 (SR)]*. On average, participants in randomized trials were found to have similar outcomes compared to similar people who received a similar treatment outside of a randomized trial.

Implications

Do not assume that an association between a treatment and an outcome found using "big data" or "real world data" means that the treatment caused the outcome unless other possible reasons for the association have been ruled out.

References

Systematic reviews

Anglemyer A, Horvath HT, Bero L. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. Cochrane Database Syst Rev. 2014(4). https://doi.org//10.1002/14651858.MR000034.pub2

Bian J, Lyu T, Loiacono A, Viramontes TM, Lipori G, Guo Y, et al. Assessing the practice of data quality evaluation in a national clinical data research network through a systematic scoping review in the era of real-world data. J Am Med Inform Assoc. 2020;27(12):1999-2010. https://doi.org/10.1093/jamia/ocaa245

Dahabreh IJ, Sheldrick RC, Paulus JK, Chung M, Varvarigou V, Jafri H, et al. Do observational studies using propensity score methods agree with randomized trials? A systematic comparison of studies on acute coronary syndromes. Eur Heart J. 2012;33(15):1893-901. https://doi.org/10.1093/eurheartj/ehs114

Ewald H, Ioannidis JPA, Ladanie A, Mc Cord K, Bucher HC, Hemkens LG. Nonrandomized studies using causal-modeling may give different answers than RCTs: a meta-epidemiological study. J Clin Epidemiol. 2020;118:29-41. https://doi.org/10.1016/j.jclinepi.2019.10.012

Gaudino M, Di Franco A, Rahouma M, Tam DY, Iannaccone M, Deb S, et al. Unmeasured confounders in observational studies comparing bilateral versus single internal thoracic artery for coronary artery bypass grafting: a meta-analysis. J Am Heart Assoc. 2018;7(1). https://doi.org/10.1161/jaha.117.008010

Gaudino M, Rahouma M, Hameed I, Khan FM, Taggart DP, Flather M, et al. Disagreement between randomized and observational evidence on the use of bilateral internal thoracic artery grafting: a meta-analytic approach. J Am Heart Assoc. 2019;8(23):e014638. https://doi.org/10.1161/jaha.119.014638

Hemkens LG, Contopoulos-Ioannidis DG, Ioannidis JP. Agreement of treatment effects for mortality from routinely collected data and subsequent randomized trials: meta-epidemiological survey. BMJ. 2016a;352:i493. https://doi.org/10.1136/bmj.i493

Hemkens LG, Contopoulos-Ioannidis DG, Ioannidis JP. Current use of routinely collected health data to complement randomized controlled trials: a meta-epidemiological survey. CMAJ Open. 2016b;4(2):E132-40. https://doi.org/10.9778/cmajo.20150036

Mathes T, Rombey T, Kuss O, Pieper D. No inexplicable disagreements between real-world data-based nonrandomized controlled studies and randomized controlled trials were found. J Clin Epidemiol. 2021;133:1-13. https://doi.org/10.1016/j.jclinepi.2020.12.019

Vist GE, Bryant D, Somerville L, Birminghem T, Oxman AD. Outcomes of patients who participate in randomized controlled trials compared to similar patients receiving similar interventions who do not participate. Cochrane Database Syst Rev. 2008(3):MR000009. https://doi.org/10.1002/14651858.mr000009.pub4

Other references

Agoritsas T, Merglen A, Shah ND, O'Donnell M, Guyatt GH. Adjusted analyses in studies addressing therapy and harm: Users' guides to the medical literature. JAMA. 2017;317(7):748-59. https://doi.org/10.1001/jama.2016.20029

Franklin JM, Glynn RJ, Martin D, Schneeweiss S. Evaluating the use of nonrandomized real-world data analyses for regulatory decision making. Clin Pharmacol Ther. 2019;105(4):867-77. https://doi.org/10.1002/cpt.1351

Goodman SN, Schneeweiss S, Baiocchi M. Using design thinking to differentiate useful from misleading evidence in observational research. JAMA. 2017;317(7):705-7. https://doi.org/10.1001/jama.2016.19970

Kleijnen J, Gøtzsche P, Kunz RA, Oxman AD, Chalmers I. So what's so special about randomisation? In: Maynard A, Chalmers I, editors. Non-Random Reflections on Health Care Research: On the 25th Anniversary Of Archie Cochrane's Effectiveness and Efficiency. London: BMJ Publishing Group; 1997. p. 93-106.

Sterne J. Commentary: does the selective inversion approach demonstrate bias in the results of studies using routinely collected data? BMJ. 2018;362. https://doi.org/10.1136/bmj.k3259

Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 2016;355:i4919. https://doi.org/10.1136/bmj.i4919

Thorpe KE, Zwarenstein M, Oxman AD, Treweek S, Furberg CD, Altman DG, et al. A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers. CMAJ. 2009;180(10):E47-57. https://doi.org/10.1503/cmaj.090523

1.2d Do not assume that a single study is a sufficient basis for a claim about treatment effects.

Explanation

The results of one study considered in isolation can be misleading. A single comparison of treatments rarely provides conclusive evidence; and results are often available from other comparisons of the same treatments. Systematic reviews of all the similar comparisons ("replications") may yield different results from those based on the initial studies, and these should help to provide more reliable and statistically precise estimates of treatment differences. Even so, obtaining reliable estimates from treatment comparisons must always consider that important studies may remain unpublished, incompletely published, or inaccessible for other reasons.

Randomized trials of oral rehydration solutions (ORS) for children with diarrhoea are an example of single comparisons of treatments that did not provide conclusive evidence *[Hahn 2002 (SR)]*. Children with diarrhoea can become dehydrated. If they become seriously dehydrated, they can die. For more than 20 years, the World Health Organization (WHO) recommended a standard ORS with a large amount of sugar and salt mixed in water. However, some researchers believed that it might be better to use a smaller amount of sugar and salt (reduced osmolarity). Eleven randomized trials published between 1982 and 2001 compared ORS with reduced osmolarity to the standard solution. A key outcome was the number of children who needed an unscheduled fluid infusion, which indicates they were becoming seriously dehydrated. The results varied. It was not until the results of all the studies were carefully summarised in a systematic review that it was shown convincingly that a reduced osmolarity solution was substantially more effective than the standard solution. Based on combined results of all 11 studies, the WHO changed its recommendation.

Replication or reproducibility is sometimes used to describe the extent to which similar studies, such as the trials of reduced osmolarity ORS, have similar results. However, these terms are not well defined and can sometimes cause confusion *[Goodman 2016]*.

Basis for this concept

There are several reasons why single studies can be misleading. First, studies are often too small to provide reliable results *[Button 2013 , Dechartres 2013 (SR), IntHout 2015 (SR)]*. Small studies provide statistically less precise estimates of treatment effects than large studies; the results of small studies are more inconsistent than the results of large ones *[IntHout 2015 (SR)]*; and small studies tend to overestimate treatment effects *[Dechartres 2013 (SR)]*. Second, studies that evaluate the effects of treatments often have a high risk of bias *[Wood 2008 (SR)]*. Third, the results of studies that address the same question often have inconsistent results *[Guyatt 2011c , IntHout 2015 (SR)]*. Fourth, randomized trials with statistically significant results are published more often, and more quickly, than trials with statistically "non-significant" results *[Hopewell 2009 (SR)]*. Studies that show benefits, especially large benefits, are also more likely to be noticed than studies that do not *[Duyx 2017 (SR), Ioannidis 2005 (SR)]*.

Systematic reviews use a structured approach to identify studies (including unpublished studies), to select and critically appraise the risk of bias in relevant studies, and to collect and analyse data from the studies that are included in the review. Compared to single studies, systematic reviews can:

- Increase the statistical precision of estimates and so reduce the chances of being misled by the play of chance (random errors),
- Reduce the chances of being misled by systematic errors (biases)
- Assess the consistency of estimates of treatment effects across studies and reduce the chances of being misled by inconsistency, and
- Assess the risk of reporting biases and reduce the risk of being misled by publication bias.

Implications

The results of single comparisons of treatments can be misleading. Consider all the relevant fair comparisons when making judgements about treatment effects.

References

Systematic reviews

Dechartres A, Trinquart L, Boutron I, Ravaud P. Influence of trial sample size on treatment effect estimates: meta-epidemiological study. BMJ. 2013;346:f2304. https://www.bmj.com/content/bmj/346/bmj.f2304.full.pdf

Duyx B, Urlings MJE, Swaen GMH, Bouter LM, Zeegers MP. Scientific citations favor positive results: a systematic review and meta-analysis. J Clin Epidemiol. 2017;88:92-101. https://doi.org/10.1016/j.jclinepi.2017.06.002

Hahn S, Kim S, Garner P. Reduced osmolarity oral rehydration solution for treating dehydration caused by acute diarrhoea in children. Cochrane Database Syst Rev. 2002(1):Cd002847. https://doi.org/10.1002/14651858.cd002847

Hopewell S, Loudon K, Clarke MJ, Oxman AD, Dickersin K. Publication bias in clinical trials due to statistical significance or direction of trial results. Cochrane Database Syst Rev. 2009(1):MR000006. https://doi.org/10.1002/14651858.mr000006.pub3

IntHout J, Ioannidis JP, Borm GF, Goeman JJ. Small studies are more heterogeneous than large ones: a meta-meta-analysis. J Clin Epidemiol. 2015;68(8):860-9. https://doi.org/10.1136/bmj.f2304

Ioannidis JPA. Contradicted and initially stronger effects in highly cited clinical research. JAMA. 2005;294(2):218-28. https://doi.org/10.1001/jama.294.2.218

Wood L, Egger M, Gluud LL, Schulz KF, Jüni P, Altman DG, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. BMJ. 2008;336(7644):601-5. https://doi.org/10.1136/bmj.39465.451748.AD

Other references

Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, et al. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci. 2013;14(5):365-76. https://doi.org/10.1038/nrn3475

Goodman SN, Fanelli D, Ioannidis JP. What does research reproducibility mean? Sci Transl Med. 2016;8(341):341ps12. https://doi.org/10.1126/scitranslmed.aaf5027

Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 7. Rating the quality of evidence--inconsistency. J Clin Epidemiol. 2011c;64(12):1294-302. https://doi.org/10.1016/j.jclinepi.2011.03.017

1.2e Do not assume that fair comparisons are not applicable in practice.

Explanation

Assumptions that fair comparisons of treatments in research are not applicable in practice can be misleading. People may claim that evidence from fair comparisons of treatments cannot be applied to everyday practice. This is likely to be true if there are important differences between the fair comparisons and everyday practice. The effects of treatments are unlikely to differ substantially unless there are compelling reasons why everyday practice is so different from the fair comparisons that the treatments are unlikely to work in the same way *[Dans 1998]*.

Deciding whether there are compelling reasons depends on evidence outside of fair comparisons of treatments (for example, basic science research that demonstrates how a treatment causes an outcome) and judgement. Reasons for uncertainty about the applicability of research only become compelling when there is compelling evidence or compelling logical reasons for expecting the effects of a treatment to be substantially different in practice.

For example, human biology tends to be more similar than different across people from different countries, races, and ethnicities. So, you would expect medicines to have similar effects most of the time. Thus, it is not necessary to conduct randomized trials of medicines in every country with large samples of people from every race and ethnicity. But there are sometimes important differences. For example, the benefits of lowering elevated blood pressure in reducing strokes and other cardiovascular morbidity and mortality are well established. However, several different types of medicine are used to lower blood pressure and there has been uncertainty about which of these should be used. There has also been uncertainty about whether these medicines worked the same in Black people and in non-Black people, particularly for angiotensin-converting enzyme (ACE) inhibitors. This is because ACE inhibitors were found to be less effective for lowering blood pressure in Black people than in non-Black people. For this reason, a randomized trial designed to compare different medicines for lowering blood pressure planned to do a subgroup analysis for Black participants in the trial, which included 33,357 participants (35% Black) in the U.S. and Canada *[Wright 2005 (RS)]*. The results of this study were largely similar for Blacks and non-Blacks, except for the effect of the ACE inhibitor on strokes. Black participants assigned to the ACE inhibitor were more likely to have a stroke than Black participants assigned to the thiazide diuretic, but not non-Black participants.

Various terms are used to describe the “applicability” of research, including transferability, generalisability, external validity, and relevance. Although these terms have been defined differently, checklists designed to assess these concepts include broadly similar criteria *[Munthe-Kaas 2019 (SR)]*. These include differences between fair comparisons and everyday practice in the characteristics of the people, characteristics of the treatments, and characteristics of the context. It is possible to generate long lists of things that could potentially be different. For example, differences in patient characteristics could include differences in age, sex, education, income, race, ethnicity, weight, comorbidity, genetic markers, astrological sign, baseline risk, etc. To avoid being misled by spurious assumptions about fair comparisons not being relevant, only those factors for which there are compelling reasons why a treatment is unlikely to work the same way in practice as it did in fair comparisons should be considered when assessing the applicability of the results.

It should be noted that most often the relative effect will be similar for people with different baseline risks. Differences in baseline risk will, however, often lead to differences in the absolute effect.

Basis for this concept

There have been at least 136 comparisons of outcomes of patients who participated in randomized trials and outcomes of patients who were eligible for the trial and received a similar treatment but did not participate *[Vist 2008 (SR)]*. The comparisons include both comparisons of the ‘experimental’ or new treatment inside and outside of the trial, and 'control' treatment comparisons. On average, the outcomes of patients participating and not participating in trials were similar. Among the 136 comparisons, 21 comparisons found statistically significant differences in outcomes. Eleven of those reported better outcomes for patients within trials and ten reported worse outcomes for patients treated within trials. These results challenge the assertion that the results of randomized trials are not applicable in practice.

However, the results of some randomized trials may be less likely to be applicable than others. Some trials are largely explanatory. That is, they are designed to assess the effects of a treatment given in ideal circumstances *[Thorpe 2009]*. Those trials may be less likely to be applicable in practice than trials that are largely pragmatic, i.e., designed to assess the effects of a treatment given in the circumstances of everyday practice.

A systematic review that compared treatment effects on mortality from randomized trials conducted in more developed versus less developed countries found similar effects in 128 out of 139 cases (92%). A few cases (8%) showed, on average, more favourable treatment effects in less developed countries. The extent to which those discrepancies reflect biases in reporting or study design versus genuine differences in treatment effects (for example, due to differences in treatment implementation or baseline risk) is uncertain.

There are statistical and logical reasons for thinking that relative measures of effect are more likely to be consistent across different baseline risks. For example, a large risk difference, say 50%, is not possible in a group of people with a baseline risk that is less than 50%. There is also empirical evidence that relative measures of effect tend to be more consistent across people with different baseline risks than absolute risks *[Guyatt 2013a]*, although that evidence has been brought into question *[Poole 2015 (OR)]*. It cannot be assumed that relative risks are consistent and can be applied in practice to people with different baseline risks. However, most of the time it is more likely that relative effects are applicable in practice across groups with different baseline risks than it is that absolute risks are applicable. Moreover, this is only a concern when it is possible to identify groups of people with important differences in baseline risk.

Implications

Do not assume fair comparisons are not applicable because of differences between fair comparisons and everyday practice, unless there are compelling reasons why treatments would work differently.

References

Systematic reviews

Munthe-Kaas H, Nøkleby H, Nguyen L. Systematic mapping of checklists for assessing transferability. Syst Rev. 2019;8(1):22. https://doi.org/10.1186/s13643-018-0893-4

Vist GE, Bryant D, Somerville L, Birminghem T, Oxman AD. Outcomes of patients who participate in randomized controlled trials compared to similar patients receiving similar interventions who do not participate. Cochrane Database Syst Rev. 2008(3):MR000009. https://doi.org/10.1002/14651858.mr000009.pub4

Other reviews

Poole C, Shrier I, VanderWeele TJ. Is the risk difference really a more heterogeneous measure? Epidemiology. 2015;26(5):714-8. https://doi.org/10.1097/ede.0000000000000354

Research studies

Wright JT, Jr., Dunn JK, Cutler JA, Davis BR, Cushman WC, Ford CE, et al. Outcomes in hypertensive black and nonblack patients treated with chlorthalidone, amlodipine, and lisinopril. JAMA. 2005;293(13):1595-608. https://doi.org/10.1001/jama.293.13.1595

Other references

Dans AL, Dans LF, Guyatt GH, Richardson S, Group ftE-BMW. Users' guides to the medical literature XIV. How to decide on the applicability of clinical trial results to your patient. JAMA. 1998;279(7):545-9. https://doi.org/10.1001/jama.279.7.545

Guyatt GH, Oxman AD, Santesso N, Helfand M, Vist G, Kunz R, et al. GRADE guidelines: 12. Preparing summary of findings tables-binary outcomes. J Clin Epidemiol. 2013a;66(2):158-72. https://doi.org/10.1016/j.jclinepi.2012.01.012

Thorpe KE, Zwarenstein M, Oxman AD, Treweek S, Furberg CD, Altman DG, et al. A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers. CMAJ. 2009;180(10):E47-57. https://doi.org/10.1503/cmaj.090523

1.3 Seemingly logical assumptions about *treatments* can be misleading.

1.3a Do not assume that treatment is needed.

Explanation

Effective treatments can prevent health problems and premature death and improve the quality of life. However, nature is a great healer and people often recover from illness without treatment. Likewise, some health problems may get worse despite treatment, or treatment may even make things worse. Not using a treatment is not the same as "no treatment". Waiting to see what happens ("letting nature take its course"), with or without treating symptoms such as pain, is a treatment option.

Sore throats are an example of an illness that gets better without treatment. Sore throats caused by bacteria (strep throat) have been treated with antibiotics primarily to prevent rheumatic fever. Rheumatic fever still occurs in some parts of the world, but it is very rare in many parts of the world. In those parts of the world antibiotics are used primarily to promote faster recovery. Antibiotics have a modest effect on recovery in the first few days, but after seven days, 90% of patients are symptom-free with or without antibiotics *[Spinks 2013 (SR)]*. Moreover, antibiotics have adverse effects, including diarrhoea and rash, and widespread use of antibiotics contributes to antibiotic resistance.

Basis for this concept

Inappropriate healthcare is not easily defined or measured, but it is widespread *[Brownlee 2017 (OR)]*. This includes inappropriate use of diagnostic tests and use of treatments that are not effective or do more harm than good, as well as use of treatments that are not needed.

Antibiotics, for example, are not needed for common respiratory tract infections *[Hirschmann 2002 (OR), Tan 2008 (SR)]*. These illnesses are rarely serious and get better without treatment. Antibiotics do not help and can cause harm. Treating symptoms by other means frequently helps.

Aggressive care for older people and dying patients is another example of treatments that are not needed. Older people near the end of their life often receive non-essential medicines that can cause discomfort and serious side effects or are no longer beneficial given limited life expectancy. Nearly 50% of older adults take one or more medications that are not necessary *[Maher 2014 (OR)]*. A "good death" is free from avoidable distress and suffering for the individual, the family, and caregivers, and in accord with the individual's and family's wishes. Unwarranted medicalisation of the last phase life with surgery, intensive testing, medical procedures, polypharmacy, hospitalisation, and intensive care increases distress and suffering rather than reducing it. Most people would prefer to die at home, yet about half die in hospital worldwide *[Brownlee 2017 (OR)]*. At the same time, appropriate palliative care is underused.

Implications

Always consider the usual course of a health problem when considering treatments other than waiting to see what happens. Sometimes treatment is not needed and may even make things worse.

References

Systematic reviews

Spinks A, Glasziou PP, Del Mar CB. Antibiotics for sore throat. Cochrane Database Syst Rev. 2013;2013(11):Cd000023. https://doi.org/10.1002/14651858.cd000023.pub4

Tan T, Little P, Stokes T. Antibiotic prescribing for self limiting respiratory tract infections in primary care: summary of NICE guidance. BMJ. 2008;337:a437. https://doi.org/10.1136/bmj.a437

Other reviews

Brownlee S, Chalkidou K, Doust J, Elshaug AG, Glasziou P, Heath I, et al. Evidence for overuse of medical services around the world. Lancet. 2017;390(10090):156-68. https://doi.org/10.1016/S0140-6736(16)32585-5

Hirschmann JV. Antibiotics for common respiratory tract infections in adults. Arch Intern Med. 2002;162(3):256-64. https://doi.org/10.1001/archinte.162.3.256

Maher RL, Hanlon J, Hajjar ER. Clinical consequences of polypharmacy in elderly. Expert Opin Drug Saf. 2014;13(1):57-65. https://doi.org/10.1517/14740338.2013.827660

1.3b Do not assume that more treatment is better.

Explanation

Increasing the dose or amount of a treatment (e.g., how many vitamin pills you take) can increase harms without increasing beneficial effects.

For example, iron deficiency is an important cause of anaemia and a major contributor to the global burden of disease *[Pasricha 2021 (OR)]*. Iron supplements are effective for preventing and treating iron deficiency anaemia. However, iron supplements can injure the upper gastrointestinal tract and cause nausea, vomiting, discomfort, diarrhoea, and constipation – and higher doses of iron increase the number and severity of adverse effects *[Cancelo-Hidalgo 2013 (SR)]*.

More aggressive treatment can also increase harms without increasing the benefits. For example, radical mastectomy entails removing the breast tissue along with the nipple, lymph nodes in the armpit, and chest wall muscles underneath the breast. This was the standard of care for breast cancer surgery for almost a century. But in the 1980s, fair comparisons found that a lumpectomy was an equally viable option that was far less extensive and easier on the patient, since it removed the tumour, not the breast itself *[Cotlar 2003]*.

Basis for this concept

Adverse effects of medicines are common, and they can be serious. For example, studies have found that about 7% of hospitalised patients in the U.S. have had serious adverse drug reactions and about 0.3% have had fatal adverse drug reactions *[Lazarou 1998 (SR)]*. About three-quarters of the reported side effects are dose related. Dose-related adverse drug reactions are also common in patients who are not hospitalised, and they can occur at dosages recommended by pharmaceutical manufacturers in package inserts *[Cohen 2001 (OR)]*. This occurs because those recommendations are based on incomplete information and sometimes do not reflect research showing that a lower dose would be better *[McCormack 2011]*.

Millions of people consume dietary supplements hoping to maintain or improve their health. These include vitamins, minerals, amino acids, herbs or other botanicals, and other substances used to increase total dietary intake. Sales of dietary supplements exceeded $30 billion in the U.S. alone in 2011. However, extensive research and systematic reviews have not detected beneficial effects *[Batsis 2021 (SR), Di Lorenzo 2015 (SR), McCormick 2010 (OR), Starr 2015 (OR)]*. Moreover, routine and high-dose supplementation may not be safe.

Implications

If a treatment is believed to be beneficial, do not assume that more of it is better.

References

Systematic reviews

Batsis JA, Apolzan JW, Bagley PJ, Blunt HB, Divan V, Gill S, et al. A systematic review of dietary supplements and alternative therapies for weight loss. Obesity. 2021;29(7):1102-13. https://doi.org/10.1002/oby.23110

Cancelo-Hidalgo MJ, Castelo-Branco C, Palacios S, Haya-Palazuelos J, Ciria-Recasens M, Manasanch J, et al. Tolerability of different oral iron supplements: a systematic review. Curr Med Res Opin. 2013;29(4):291-303. https://doi.org/10.1185/03007995.2012.761599

Di Lorenzo C, Ceschi A, Kupferschmidt H, Lüde S, De Souza Nascimento E, Dos Santos A, et al. Adverse effects of plant food supplements and botanical preparations: a systematic review with critical evaluation of causality. Br J Clin Pharmacol. 2015;79(4):578-92. https://doi.org/10.1111/bcp.12519

Lazarou J, Pomeranz BH, Corey PN. Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. JAMA. 1998;279(15):1200-5. https://doi.org/10.1001/jama.279.15.1200

Other reviews

Cohen JS. Dose discrepancies between the Physicians' Desk Reference and the medical literature, and their possible role in the high incidence of dose-related adverse drug events. Archives of Internal Medicine. 2001;161(7):957-64. https://doi.org/10.1001/archinte.161.7.957

McCormick DB. Vitamin/mineral supplements: of questionable benefit for the general population. Nutr Rev. 2010;68(4):207-13. https://doi.org/10.1111/j.1753-4887.2010.00279.x

Pasricha S-R, Tye-Din J, Muckenthaler MU, Swinkels DW. Iron deficiency. Lancet. 2021;397(10270):233-48. https://doi.org/10.1016/S0140-6736(20)32594-0

Starr RR. Too little, too late: ineffective regulation of dietary supplements in the United States. Am J Public Health. 2015;105(3):478-85. https://doi.org/10.2105/ajph.2014.302348

Other references

Cotlar AM, Dubose JJ, Rose DM. History of surgery for breast cancer: radical to the sublime. Curr Surg. 2003;60(3):329-37. https://doi.org/10.1016/s0149-7944(02)00777-8

McCormack JP, Allan GM, Virani AS. Is bigger better? An argument for very low starting doses. CMAJ. 2011;183(1):65-9. https://doi.org/10.1503/cmaj.091481

1.3c Do not assume that a treatment is helpful or safe based on how widely used it is or has been.

Explanation

Treatments that have not been properly evaluated but are widely used or have been used for a long time are often assumed to work. Sometimes, however, they may be unsafe or of doubtful benefit.

Bloodletting, taking blood from a patient to prevent or cure illness, was the most common procedure performed by physicians and surgeons for almost two thousand years *[Science Museum 2009]*. As late as 1923, it was recommended in *Principles and the Practice of Medicine [Stewart 2019]*. In addition to not being helpful, bloodletting was not safe. People were killed from blood loss, including George Washington, the first president of the U.S. *[Morens 1999]*. It could also lead to severe or even fatal infections.

Medicine to reduce heart rhythm abnormalities is a more recent example of a widely-used treatment that was deadly. Because heart rhythm abnormalities are associated with an increased risk of early death after a heart attack, it was believed that medicines that reduced these abnormalities would also reduce early deaths. These medicines were used for many years before it was discovered that they increase the risk of sudden death. It has been estimated that, at the peak of their use in the late 1980s, they may have been killing as many as 70,000 people every year in the U.S. alone *[Moore 1995]*.

Basis for this concept

The reversal of established medical practice is common and occurs across all classes of medical practice. Reviews of fair comparisons published in leading medical journals (JAMA, Lancet, and the New England Journal of Medicine) between 2003 and 2017 identified 542 "medical reversals" – evidence that established practices were ineffective or harmful *[Herrera-Perez 2019 (SR), Prasad 2013 (SR)]*. Similar studies have found that common practice is commonly shown to be ineffective or harmful in specific areas of practice *[Haslam 2021 (SR), Herrera-Perez 2020 (SR)]*.

It is difficult to estimate how many contemporary medical practices are not any better than or are worse than doing nothing or doing something else that is simpler or less expensive *[Ioannidis 2013]*. About a quarter of original articles published in the New England Journal of Medicine evaluated established practices. Of those, about 40% found that established practices were not helpful or not safe and about 40% reaffirmed established practices. It is uncertain how representative these estimates are of evaluations published in other journals or, more importantly, of common practices generally. Nonetheless, many widely-used treatments are not helpful or are not safe *[Luo 2013 (RS)]*. Estimates of waste due to overtreatment or low-value care in the U.S. healthcare system range from $75.7 billion to $101.2 billion per year *[Shrank 2019 (SR)]*.

Complementary and alternative medicine is widely used and has been for a long time *[Ernst 2000 (SR), Frass 2012 (SR)]*. Because many complementary and alternative treatments are poorly evaluated, it is uncertain whether they are effective or safe. For example, homeopathy has been used for over 200 years, but systematic reviews of the effects of homeopathy have found no condition that responds convincingly better to homeopathic treatment than to placebo [Ernst 2002 (SR), Jorgensen 2013 (SR)]. Similarly, herbal remedies are widely used and have been for a long time, but the effectiveness of many herbal remedies is uncertain *[Hu 2011 (SR)]*. Moreover, some herbal remedies have been found to have adverse effects [Lee 2016 (SR)]. Acupuncture, which has been used for about 3,000 years, has been shown to be effective for some conditions, but not for others *[Tait 2002 (SR)]*. And although acupuncture is relatively safe, it can have both minor and serious adverse effects. Chiropractic treatments, which have been used for over 100 years, have been shown to be effective for some upper extremity conditions, but not for other conditions *[Salehi 2015 (SR)]*. Chiropractic treatments can also have both minor and serious adverse effects *[Gouveia 2009 (SR)]*.

Implications

Do not assume that treatments are beneficial or safe simply because they are widely used or have been used for a long time, unless this has been shown in systematic reviews of fair comparisons of treatments.

References

Systematic reviews

Ernst E. Prevalence of use of complementary/alternative medicine: a systematic review. Bull World Health Organ. 2000;78(2):252-7. http://www.ncbi.nlm.nih.gov/pmc/articles/pmc2560678/

Frass M, Strassl RP, Friehs H, Müllner M, Kundi M, Kaye AD. Use and acceptance of complementary and alternative medicine among the general population and medical personnel: a systematic review. Ochsner J. 2012;12(1):45-56. http://www.ncbi.nlm.nih.gov/pmc/articles/pmc3307506/

Gouveia LO, Castanho P, Ferreira JJ. Safety of chiropractic interventions: a systematic review. Spine (Phila Pa 1976). 2009;34(11):E405-13. https://doi.org/10.1097/brs.0b013e3181a16d63

Haslam A, Gill J, Crain T, Herrera-Perez D, Chen EY, Hilal T, et al. The frequency of medical reversals in a cross-sectional analysis of high-impact oncology journals, 2009-2018. BMC Cancer. 2021;21(1):889. https://doi.org/10.1186/s12885-021-08632-8

Herrera-Perez D, Fox-Lee R, Bien J, Prasad V. Frequency of medical reversal among published randomized controlled trials assessing cardiopulmonary resuscitation (CPR). Mayo Clin Proc. 2020;95(5):889-910. https://doi.org/10.1016/j.mayocp.2020.01.036

Herrera-Perez D, Haslam A, Crain T, Gill J, Livingston C, Kaestner V, et al. A comprehensive review of randomized clinical trials in three medical journals reveals 396 medical reversals. eLife. 2019;8:e45183. https://doi.org/10.7554/eLife.45183

Hu J, Zhang J, Zhao W, Zhang Y, Zhang L, Shang H. Cochrane systematic reviews of Chinese herbal medicines: an overview. PLoS One. 2011;6(12):e28696. https://doi.org/10.1371/journal.pone.0028696

Prasad V, Vandross A, Toomey C, Cheung M, Rho J, Quinn S, et al. A decade of reversal: an analysis of 146 contradicted medical practices. Mayo Clin Proc. 2013;88(8):790-8. https://doi.org/10.1016/j.mayocp.2013.05.012

Salehi A, Hashemi N, Imanieh MH, Saber M. Chiropractic:is it efficient in treatment of diseases? Review of systematic reviews. Int J Community Based Nurs Midwifery. 2015;3(4):244-54. http://www.ncbi.nlm.nih.gov/pmc/articles/pmc4591574/

Shrank WH, Rogstad TL, Parekh N. Waste in the US health care system: estimated costs and potential for savings. JAMA. 2019;322(15):1501-9. https://doi.org/10.1001/jama.2019.13978

Tait PL, Brooks LJ, Harstall C. Acupuncture: evidence from systematic reviews and meta-analyses: Alberta Heritage Foundation for Medical Research Edmonton, Alberta, Canada; 2002.

Research studies

Luo XM, Tang JL, Hu YH, Li LM, Wang YL, Wang WZ, et al. How often are ineffective interventions still used in clinical practice? A cross-sectional survey of 6,272 clinicians in China. PLoS One. 2013;8(3):e52159. https://doi.org/10.1371/journal.pone.0052159

Other references

Ioannidis JP. How many contemporary medical practices are worse than doing nothing or doing less? Mayo Clin Proc. 2013;88(8):779-81. https://doi.org/10.1016/j.mayocp.2013.05.010

Moore TJ. Deadly Medicine: Why Tens of Thousands of Heart Patients Died in America's Worst Drug Disaster. New York: Simon & Schuster; 1995.

Morens DM. Death of a president. N Engl J Med. 1999;341(24):1845-9. https://doi.org/10.1056/nejm199912093412413

Science Museum. Bloodletting, 2009. http://www.sciencemuseum.org.uk/broughttolife/techniques/bloodletting.aspx]

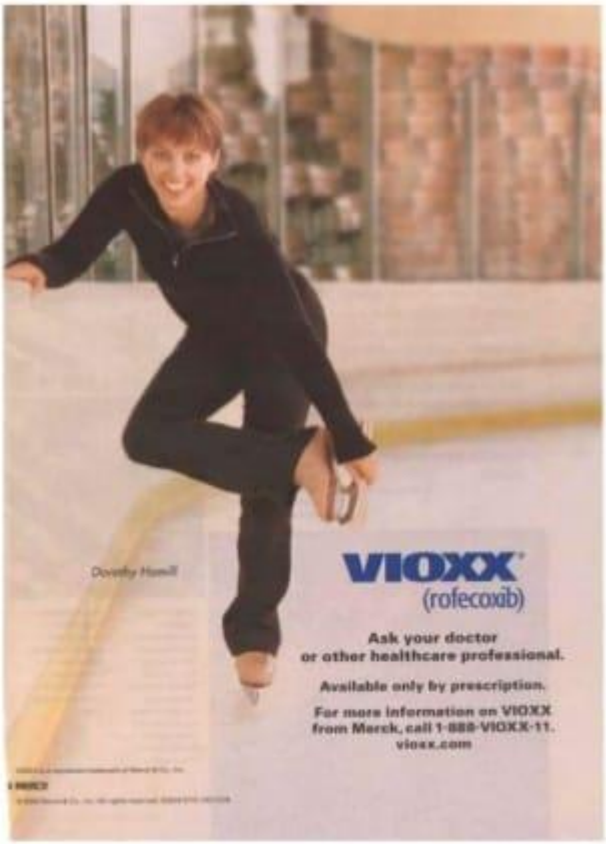
Stewart O. Bloodletting: a brief historical perspective and modern medical applications. Clinical Correlations, October. 2019;31. https://www.clinicalcorrelations.org/2019/10/31/bloodletting-a-brief-historical-perspective-and-modern-medical-applications/

1.3d Do not assume that a treatment is better based on how new or technologically impressive it is.

Explanation

New treatments can be assumed to be better simply because they are new, more expensive, or technologically impressive. However, on average, they are only very slightly likely to be better than other available treatments. Some side effects of treatments, for example, take time to appear, and without long term follow-up it may not be possible to know whether they will appear.

Vioxx (rofecoxib) was a new non-steroidal anti-inflammatory drug (NSAID) prescribed to decrease pain and inflammation in arthritis and acute pain in adults. Fair comparisons showed that more people who took Vioxx for eight weeks had relief from arthritis symptoms than people who took a 'sugar pill' or placebo, and that it worked just as well as Naprosyn *[Garner 2005 (SR)]*. Vioxx was approved by the U.S. Federal Drug Administration (FDA) in 1999. The producer of Vioxx spent $161 million advertising Vioxx with advertisements like this:


Dorothy Hamill
VIOXX
(rofecoxib)
Ask your doctor
or other healthcare professional.
Available only by prescription.
For more information on VIOXX
from Merck, call 1-888-VIOXX-11.
vioxx.com

``Just one small pill let me resume my life. I could get up in the morning without pain. I could take my daughter to the park, lace up my skates and perform again. It was a miracle."

DOROTHY HAMILL, OLYMPIC ICESKATING CHAMPION, IN A TV AD FOR VIOXX

However, Vioxx was withdrawn from the market in 2004 after it was shown that long-term use increased the risk of heart attack and stroke.

Basis for this concept

About 4% of new medicines approved in Canada between 1990 and 2009 were withdrawn because of adverse effects after two to eight years *[Lexchin 2014 (RS)]*. Worldwide, the average time between introduction of a medicine and its withdrawal due to safety is about 20 years (SD±14 years) *[Craveiro 2020 (SR)]*. Worldwide, among medicines launched between 1951 and 2007, 83 were withdrawn because of drug-attributed deaths between 1957 and 2001 *[Onakpoya 2017 (SR)]*. Among 353 medicines withdrawn from any country between 1950 and 2015 because of an adverse effect, only 40 were withdrawn worldwide *[Onakpoya 2016b (SR)]*. The median time between the first launch and worldwide withdrawal of a medicine was four years (interquartile range four to 24 years). The interval between launch date and reports of adverse drug reactions has shortened over the past few decades *[Onakpoya 2016a (SR)]*. This may be in part because of more people being exposed more quickly, leading to quicker detection of adverse reactions. However, withdrawal of medicines following reports of suspected serious adverse reactions has not improved consistently, and harmful medicines are less likely to be withdrawn in African countries.

It is more difficult to document the proportion of new non-pharmacological treatments that are found to be harmful. Only slightly more than half of new treatments that are evaluated in randomized trials have been found to be better than established treatments, and few were substantially better *[Djulbegovic 2012 (SR)]*. This suggests that a large proportion of new treatments are unlikely to be substantially better than other available treatments. Large effects of medical treatments on outcomes that are important to patients are uncommon *[Pereira 2012 (SR)]*. Many new non-pharmaceutical treatments are not evaluated in randomized trials, so it is uncertain how effective or safe they are. New treatments with limited or no evidence of benefit are frequently introduced into practice. For example, about half of the recommendations in major cardiology guidelines are based on low-certainty evidence or expert opinion *[Tricoci 2009 (SR)]*. Similarly, about half of the recommendations in *UpToDate,* a widely used medical textbook, are based on low-certainty evidence *[Agoritsas 2017 (RS)]*.

Implications

Do not assume that a treatment is better or safer simply because it is new, brand-named, expensive, or technologically impressive.

References

Systematic reviews

Craveiro NS, Lopes BS, Tomás L, Almeida SF. Drug withdrawal due to safety: a review of the data supporting withdrawal decision. Curr Drug Saf. 2020;15(1):4-12. https://doi.org/10.2174/1574886314666191004092520

Djulbegovic B, Kumar A, Glasziou PP, Perera R, Reljic T, Dent L, et al. New treatments compared to established treatments in randomized trials. Cochrane Database Syst Rev. 2012;10(10):Mr000024. https://doi.org/10.1002/14651858.mr000024.pub3

Garner SE, Fidan D, Frankish RR, Judd M, Towheed T, Tugwell P, et al. Rofecoxib for rheumatoid arthritis. Cochrane Database Syst Rev. 2005(1):CD003685. https://doi.org//10.1002/14651858.CD003685.pub2

Onakpoya IJ, Heneghan CJ, Aronson JK. Post-marketing withdrawal of 462 medicinal products because of adverse drug reactions: a systematic review of the world literature. BMC Med. 2016a;14:10. https://doi.org/10.1186/s12916-016-0553-2

Onakpoya IJ, Heneghan CJ, Aronson JK. Worldwide withdrawal of medicinal products because of adverse drug reactions: a systematic review and analysis. Crit Rev Toxicol. 2016b;46(6):477-89. https://doi.org/10.3109/10408444.2016.1149452

Onakpoya IJ, Heneghan CJ, Aronson JK. Post-marketing regulation of medicines withdrawn from the market because of drug-attributed deaths: an analysis of justification. Drug Saf. 2017;40(5):431-41.

Pereira TV, Horwitz RI, Ioannidis JP. Empirical evaluation of very large treatment effects of medical interventions. JAMA. 2012;308(16):1676-84. https://doi.org/10.1001/jama.2012.13444

Tricoci P, Allen JM, Kramer JM, Califf RM, Smith SC, Jr. Scientific evidence underlying the ACC/AHA clinical practice guidelines. JAMA. 2009;301(8):831-41. https://doi.org/10.1001/jama.2009.205

Research studies

Agoritsas T, Merglen A, Heen AF, Kristiansen A, Neumann I, Brito JP, et al. UpToDate adherence to GRADE criteria for strong recommendations: an analytical survey. BMJ Open. 2017;7(11):e018593. https://doi.org/10.1136/bmjopen-2017-018593

Lexchin J. How safe are new drugs? Market withdrawal of drugs approved in Canada between 1990 and 2009. Open Med. 2014;8(1):e14-9. http://www.ncbi.nlm.nih.gov/pmc/articles/pmc4085091/

1.3e Do not assume that earlier detection of 'disease' is better.

Explanation

People often assume that early detection of disease and 'treating' people who are at statistical risk of disease lead to better outcomes. However, screening people to detect disease or treating people at statistical risk of disease is only helpful if two conditions are met. First, there must be an effective treatment. Second, people who are treated before the disease becomes apparent must do better than people who are treated after the disease becomes apparent. Screening and treating people at statistical risk of a disease can lead to overdiagnosis and overtreatment. Screening tests can be inaccurate (e.g., misclassifying people who do not have a disease as if they do have the disease). Screening or treating a statistical risk factor as if it is a 'disease' can also cause harm by labelling people as being sick when they are not, and because of side effects of the tests and treatments.

Screening for phenylketonuria (PKU) is an example of early detection of disease that is better than late detection. PKU is a rare inherited disorder. People with PKU cannot metabolise phenylalanine. Untreated, PKU results in severe intellectual disability, epilepsy, and behavioural problems. PKU can be detected in newborn babies with a drop of blood. Treatment includes a special diet and regular blood tests. With early diagnosis and treatment, most children with PKU can live healthy lives *[van Wegberg 2017 (SR)]*.

Screening women without symptoms for ovarian cancer is an example of early detection that does more harm than good. In randomized trials with nearly 300,000 women, there was not an important difference in similar numbers of women who died from ovarian cancer as women who were screened and those who were not *[Henderson 2018 (SR)]*. Harms of screening included surgery (with major surgical complications) in women found to not have cancer.

Basis for this concept

Screening to detect disease earlier can harm people in several different ways, including:

- *Undesirable effects of the screening tests*
  Screening tests can be bothersome and may occasionally cause harm. For example, screening mammography for breast cancer can cause discomfort or pain, and some women decide not to have mammograms because they can be painful *[Nelson 2016 (SR)]*. Screening for breast cancer also can cause anxiety, distress, and other psychological responses. More invasive tests can sometimes cause more serious harm. For example, screening colonoscopy for colon cancer is estimated to cause three perforations and 15 major bleeds for every 10,000 patients screened *[Lin 2021 (SR)]*.

- *False-positive and false-negative test results*
  Tests that are positive, indicating disease, when in fact the individual does not have disease can result in adverse psychological effects. For example, 42% of women screened for breast cancer every other year for 10 years have at least one false-positive mammography, and 6% receive a biopsy because of a false-positive test *[Nelson 2016 (SR)]*. Compared to women with normal results, women with false-positive results are more likely to experience breast cancer specific worry, worries that affected their mood or daily activities, and lower mental functioning and vitality. False-negative test results can result in a delay in recognising and treating breast cancer.

- *Undesirable effects of the treatment*
  Surgery, radiation, and chemotherapy, which are used to treat cancer, all have serious adverse effects. For example, men with prostate cancer detected by screening can be treated with surgery and radiotherapy, which can cause impotence and urinary incontinence *[Michaelson 2008 (OR)]*.

- *Overdiagnosis, overtreatment, and overmedicalisation*
  Overdiagnosis means the detection of a condition or problem that would never cause a person harm during their lifetime. Overtreatment means that people receive more extensive or invasive treatment than is required to improve health outcomes. For example, men with prostate cancer detected by screening can be treated with surgery and radiotherapy. These can cause impotence and urinary incontinence when the cancer may not have caused them any harm in their lifetime.

  Overmedicalisation is wrongly defining and treating human conditions and problems as medical conditions. This can result in both overdiagnosis and overtreatment. For example, lowering the threshold for a risk factor such as high blood pressure or gestational diabetes can result in many people who may never experience any harm caused by the condition being diagnosed as being "at risk" and treated unnecessarily. The lower the baseline risk is, say for having a stroke or heart attack, the lower the absolute effect will be, assuming the relative effect is the same. So, with lower thresholds, the likelihood of desirable effects decreases, while the likelihood of undesirable effects stays the same *[Doust 2020]*. For example, there is evidence that attention-deficit/hyperactivity disorder (ADHD) is being diagnosed more frequently and that the increase is due to milder cases being diagnosed and treated with drugs *[Kazda 2021 (SR)]*. The benefits of pharmacological treatment for youth with milder symptoms are uncertain and may be outweighed by the harms.

- *Labelling*
  People who are labelled as having a condition or a disease may experience adverse effects simply from being labelled. For example, a study with 33,000 adults found that individuals who were aware that they had hypertension had more psychological distress than individuals who were unaware that they had hypertension *[Hamer 2010 (RS)]*. Other studies have found increased absenteeism and decreased psychological well-being associated with labelling, although the results are not consistent *[Guirguis-Blake 2021 (SR), Macdonald 1984 (OR)]*.

The evidence for and against screening varies. Out of 87 recommendations (for 64 topics) made by the U.S. Preventive Services Taskforce last updated between 2011 and 2021, there was high certainty that the benefits substantially outweigh the harms for eight *[U.S. Preventive Services Task Force 2021 (SR)]*. There was high-certainty evidence of a moderate net benefit or moderate certainty of a moderate to substantial net benefit for another 23 recommendations. There was at least moderate certainty that the net benefit is small for four recommendations. For 14 recommendations, there was moderate or high certainty that there is no net benefit or that the harms outweigh the benefits, and for 38 statements there was insufficient evidence to assess the balance of the benefits and harms of screening.

Implications

Do not assume that early detection of disease is worthwhile if it has not been assessed in systematic reviews of fair comparisons between people who were screened and people who were not screened.

References

Systematic reviews

Guirguis-Blake JM, Evans CV, Webber EM, Coppola EL, Perdue LA, Weyrich MS. Screening for hypertension in adults: updated evidence report and systematic review for the US Preventive Services Task Force. JAMA. 2021;325(16):1657-69. https://doi.org/10.1001/jama.2020.21669

Henderson JT, Webber EM, Sawaya GF. Screening for ovarian cancer: an updated evidence reviewfor the U.S. Preventive Services Task Force. Rockville (MD): Agency for Healthcare Research and Quality (US); 2018. http://www.ncbi.nlm.nih.gov/books/nbk493399/

Kazda L, Bell K, Thomas R, McGeechan K, Sims R, Barratt A. Overdiagnosis of attention-deficit/hyperactivity disorder in children and adolescents: a systematic scoping review. JAMA Netw Open. 2021;4(4):e215335. https://doi.org/10.1001/jamanetworkopen.2021.5335

Lin JS, Perdue LA, Henrikson NB, Bean SI, Blasi PR. Screening for colorectal cancer: an evidence update for the U.S. Preventive Services Task Force. Rockville (MD): Agency for Healthcare Research and Quality (US); 2021. http://www.ncbi.nlm.nih.gov/books/nbk570913/

Nelson HD, Pappas M, Cantor A, Griffin J, Daeges M, Humphrey L. Harms of breast cancer screening: systematic review to update the2009 U.S. Preventive Services Task Force recommendation. Ann Intern Med. 2016;164(4):256-67. https://doi.org/10.7326/m15-0970

U.S. Preventive Services Task Force. Recommendations. 2021. https://www.uspreventiveservicestaskforce.org/uspstf/topic_search_results?topic_status=P&grades%5B%5D=A&grades%5B%5D=B&grades%5B%5D=C&grades%5B%5D=D&grades%5B%5D=I&type%5B%5D=5&searchterm=

van Wegberg AMJ, MacDonald A, Ahring K, Bélanger-Quintana A, Blau N, Bosch AM, et al. The complete European guidelines on phenylketonuria: diagnosis and treatment. Orphanet J Rare Dis. 2017;12(1):162. https://doi.org/10.1186/s13023-017-0685-2

Other reviews

Macdonald LA, Sackett DL, Haynes RB, Taylor DW. Labelling in hypertension: a review of the behavioural and psychological consequences. J Chronic Dis. 1984;37(12):933-42. https://doi.org/10.1016/0021-9681(84)90070-5

Michaelson MD, Cotter SE, Gargollo PC, Zietman AL, Dahl DM, Smith MR. Management of complications of prostate cancer treatment. CA Cancer J Clin. 2008;58(4):196-213. https://doi.org/10.3322/ca.2008.0002

Research studies

Hamer M, Batty GD, Stamatakis E, Kivimaki M. Hypertension awareness and psychological distress. Hypertension. 2010;56(3):547-50. https://doi.org/10.1161/hypertensionaha.110.153775

Other references

Doust JA, Bell KJL, Glasziou PP. Potential consequences of changing disease classifications. JAMA. 2020;323(10):921-2. https://doi.org/10.1001/jama.2019.22373

1.4 Trust in a source alone can be misleading.

1.4a Do not assume that personal experiences alone are sufficient.

Explanation

People can be led to believe that improvements in a health problem (for example, recovery from a disease) resulted from having received a treatment. Similarly, they might believe that an undesirable health outcome was due to having received a treatment. However, the fact that an individual recovered after receiving a treatment does not mean that the treatment caused the improvement, or that other people receiving the same treatment will also improve. The improvement (or the undesirable health outcome) might have occurred even without treatment.

One reason that personal experiences – including a series of personal experiences – are sometimes misleading is that experiences, such as pain, fluctuate and tend to return to a more normal or average level. This is sometimes referred to as "regression to the mean". For example, people often treat symptoms such as pain when they are very bad and would improve anyway without treatment. The same applies to a series of experiences. For example, if there is a spike in the number of traffic crashes someplace, traffic lights may be installed to reduce these. A subsequent reduction may leave the impression that the traffic lights caused this change. However, it is possible that the number of crashes would have returned to a more normal level without the traffic lights.

If you have a splinter that is causing pain and the pain goes away right away after you pull out the splinter, you can be confident that pulling out the splinter (the treatment) caused the outcome (no more pain). This is because the outcome happened right after the treatment and without the treatment the pain was constant and would very likely continue *[Glasziou 2007]*. However, few conditions are constant (unchanging without treatment) and respond quickly to treatment. So, for example, it is impossible to know based on your personal experience whether you did or did not have a stroke or cancer when you are 70 because of your diet when you were younger.

Unless an outcome rarely, if ever, occurs without treatment, it is not possible to know based on personal experience whether the treatment caused the outcome, even if the outcome occurs shortly after the treatment. For example, tension-type headaches are very common. In adults who have frequent headaches, about 5%, 20%, and 44% are likely to be pain-free within one, two, and four hours respectively without taking paracetamol (acetaminophen) *[Stephens 2016 (SR)]*. So, if an individual with frequent tension-type headaches took paracetamol and the headache went away, it would not be possible for that individual to know whether it was because of the medicine or if it would have gone away just as quickly without the medicine.

Basis for this concept

Most treatments have at best modest effects *[Nagendran 2016 (SR), Pereira 2012 (SR)]*, which cannot be reliably detected by personal experience. For example, randomized trials have compared a placebo to "no treatment" for a variety of conditions, treatments, and outcomes. On average 41% of participants allocated to "no treatment" in these trials had a good outcome independent of a possible placebo effect *[Hróbjartsson 2010 (SR)]*. Based on personal experience, it would not have been possible for individuals in those trials to know whether a good outcome was because of the treatment or would have occurred without the treatment.

Personal experience can provide compelling evidence of beneficial effects for conditions that are constant (unchanging or consistent without treatment) and respond quickly to treatment [*Glasziou 2007*]. However, there are not many examples of conditions like that. Consequently, case reports (the experience of an individual) and case series (reports of several individuals) rarely provide a reliable basis for concluding that a treatment has a beneficial effect or is safe, and they can be misleading. For example, hundreds of case reports and series reported "successful" extracranial-intracranial arterial bypass surgery for stroke prevention, but a large randomized trial found that the surgery was ineffective and harmful [*Haynes 1990*]. Moreover, it appears likely that cases are reported selectively, they may represent outliers, and they are often poorly reported [*Agha 2016 (SR)*, *Albrecht 2005 (SR)*, *Albrecht 2009 (SR)*, *Oliveira 2006 (SR)*, *Richason 2009 (SR)*]. Case reports can generate hypotheses about the effects of treatments and lead to research to test those hypotheses, but it is uncertain how often those hypotheses turn out to be correct [*Albrecht 2005 (SR)*, *Dalziel 2005 (SR)*, *Olaku 2011 (SR)*].

Case reports and case series cannot establish causation for common outcomes, but they can provide compelling evidence for rare adverse events [*Hauben 2007*]. For example, sudden, unexplained perforation of the colon is extremely rare [*Namikawa 2011 (RS)*]. So, it is safe to assume that perforations following colonoscopy are caused by the colonoscopy and case series without a comparison group can provide reliable estimates of the risk of perforation [*Lin 2021 (SR)*]. There is very limited empirical evidence of how reliable case reports and case series are with respect to adverse effects [*Vandenbroucke 2001*]. A review of anecdotal reports of suspected adverse drug reactions published in four high-profile journals in 1963 found that 35 of 47 reports were clearly correct [*Venning 1982 (SR)*]. Modelling studies suggest that for rare events, coincidental associations between taking a medicine and the event are so unlikely that more than three reports constitute a strong warning requiring further investigation [*Begaud 1994* , *Tubert 1992*].

Implications

If an individual improved after receiving a treatment it does not necessarily mean that the treatment caused the improvement, or that other people receiving the same treatment will also improve.

References

Systematic reviews

Agha RA, Fowler AJ, Lee SY, Gundogan B, Whitehurst K, Sagoo HK, et al. Systematic review of the methodological and reporting quality of case series in surgery. Br J Surg. 2016;103(10):1253-8. https://doi.org/10.1002/bjs.10235

Albrecht J, Meves A, Bigby M. Case reports and case series from Lancet had significant impact on medical literature. J Clin Epidemiol. 2005;58(12):1227-32. https://doi.org/10.1016/j.jclinepi.2005.04.003

Albrecht J, Meves A, Bigby M. A survey of case reports and case series of therapeutic interventions in the Archives of Dermatology. Int J Dermatol. 2009;48(6):592-7. https://doi.org/10.1111/j.1365-4632.2009.04031.x

Dalziel K, Round A, Stein K, Garside R, Castelnuovo E, Payne L. Do the findings of case series studies vary significantly according to methodological characteristics? Health Technol Assess. 2005;9(2):iii-iv, 1-146. https://doi.org/10.3310/hta9020

Hróbjartsson A, Gøtzsche PC. Placebo interventions for all clinical conditions. Cochrane Database Syst Rev. 2010;2010(1):Cd003974. https://doi.org/10.1002/14651858.cd003974.pub3

Lin JS, Perdue LA, Henrikson NB, Bean SI, Blasi PR. Screening for colorectal cancer: an evidence update for the U.S. Preventive Services Task Force. Rockville (MD): Agency for Healthcare Research and Quality (US); 2021. http://www.ncbi.nlm.nih.gov/books/nbk570913/

Nagendran M, Pereira TV, Kiew G, Altman DG, Maruthappu M, Ioannidis JP, et al. Very large treatment effects in randomised trials as an empirical marker to indicate whether subsequent trials are necessary: meta-epidemiological assessment. BMJ. 2016;355:i5432. https://doi.org/10.1136/bmj.i5432

Olaku O, White JD. Herbal therapy use by cancer patients: a literature review on case reports. Eur J Cancer. 2011;47(4):508-14. https://doi.org/10.1016/j.ejca.2010.11.018

Oliveira GJ, Leles CR. Critical appraisal and positive outcome bias in case reports published in Brazilian dental journals. J Dent Educ. 2006;70(8):869-74. https://doi.org/10.1002/j.0022-0337.2006.70.8.tb04153.x

Pereira TV, Horwitz RI, Ioannidis JP. Empirical evaluation of very large treatment effects of medical interventions. JAMA. 2012;308(16):1676-84. https://doi.org/10.1001/jama.2012.13444

Richason TP, Paulson SM, Lowenstein SR, Heard KJ. Case reports describing treatments in the emergency medicine literature: missing and misleading information. BMC Emerg Med. 2009;9:10. https://doi.org/10.1186/1471-227x-9-10

Stephens G, Derry S, Moore RA. Paracetamol (acetaminophen) for acute treatment of episodic tension-type headache in adults. Cochrane Database Syst Rev. 2016;2016(6):Cd011889. https://doi.org/10.1002/14651858.cd011889.pub2

Venning GR. Validity of anecdotal reports of suspected adverse drug reactions: the problem of false alarms. BMJ. 1982;284(6311):249-52. https://doi.org/10.1136/bmj.284.6311.249

Research studies

Namikawa T, Ozaki S, Okabayashi T, Dabanaka K, Okamoto K, Mimura T, et al. Clinical characteristics of the idiopathic perforation of the colon. J Clin Gastroenterol. 2011;45(9):e82-6. https://doi.org/10.1097/mcg.0b013e31820ca4c2

Other references

Begaud B, Moride Y, Tubert-Bitter P, Chaslerie A, Haramburu F. False-positives in spontaneous reporting: should we worry about them? Br J Clin Pharmacol. 1994;38(5):401-4. https://doi.org/10.1111/j.1365-2125.1994.tb04373.x

Glasziou P, Chalmers I, Rawlins M, McCulloch P. When are randomised trials unnecessary? Picking signal from noise. BMJ. 2007;334(7589):349-51. https://doi.org/10.1136/bmj.39070.527986.68

Hauben M, Aronson JK. Gold standards in pharmacovigilance: the use of definitive anecdotal reports of adverse drug reactions as pure gold and high-grade ore. Drug Saf. 2007;30(8):645-55. https://doi.org/10.2165/00002018-200730080-00001

Haynes RB. Loose connections between peer-reviewed clinical journals and clinical practice. Ann Intern Med. 1990;113(9):724-8. https://doi.org/10.7326/0003-4819-113-9-724

Tubert P, Bégaud B, Péré JC, Haramburu F, Lellouch J. Power and weakness of spontaneous reporting: a probabilistic approach. J Clin Epidemiol. 1992;45(3):283-6. https://doi.org/10.1016/0895-4356(92)90088-5

Vandenbroucke JP. In defense of case reports and case series. Ann Intern Med. 2001;134(4):330-4. https://doi.org/10.7326/0003-4819-134-4-200102200-00017

1.4b Do not assume that your beliefs are correct.

Explanation

People often look for and use information to support their own beliefs, including beliefs about the effects of treatments. This is sometimes called ‘confirmation bias’. Confirmation bias can occur when people want a claim about treatment effects to be true. By focussing on evidence or arguments that support their existing beliefs and ignoring evidence or arguments that challenge these, people believe claims that confirm what they believe or wanted to be true without thinking critically about the basis for the claims.

When looking for health information, many people search the Internet. However, the information they select, and their perception of that information may be biased based on their prior beliefs. For example, parents of young children are more likely to select information about vaccination that is consistent than information that is inconsistent with their prior beliefs, and they perceive information that is consistent with their prior beliefs as being more credible, useful, and convincing *[Meppelink 2019 (RS)]*.

Basis for this concept

A great deal of empirical evidence supports the idea that confirmation bias is extensive and strong *[Nickerson 1998 (OR)]*. The evidence also supports the view that once one has a belief, the primary motivation in seeking and evaluating information is to defend or justify that belief. People tend to seek information that they consider supportive of their existing beliefs and to interpret information in ways that endorse those beliefs. Conversely, they tend not to seek and perhaps even to avoid information that contradicts their beliefs.

Confirmation bias may explain people’s tendency to believe that a treatment was responsible for a desired result. People decide to use a treatment to bring about a health-related result. If the desired result occurs, the natural tendency is to attribute it to the treatment, which was used based on the belief it would cause the desired result. They often do not seriously consider the possibility that the result might have occurred without the treatment.

There are several explanations for confirmation bias. One is that people find it easier to believe propositions they would like to be true than propositions they would prefer to be false. Another is that people tend to avoid cognitive dissonance – anxiety that results from holding contradictory beliefs. People do not naturally adopt a falsifying strategy of hypothesis testing. Our natural tendency seems to be to look for evidence that is directly supportive of hypotheses we favour.

Confirmation bias is also found in the scientific literature. Citation bias is the selective citation of scientific articles based on their results *[Gøtzsche 2022 (OR)]*. Studies of citation bias have found that articles in which the authors explicitly concluded to have found support for their hypothesis were cited 2.7 times as often as articles that did not *[Duyx 2017 (SR)]*. This can lead to wrong conclusions and decisions.

Implications

Don’t be misled by your own beliefs or rely on them unless they are based on the results of systematic reviews of fair comparisons of treatments.

References

Systematic reviews

Duyx B, Urlings MJE, Swaen GMH, Bouter LM, Zeegers MP. Scientific citations favor positive results: a systematic review and meta-analysis. J Clin Epidemiol. 2017;88:92-101. https://doi.org/10.1016/j.jclinepi.2017.06.002

Other reviews

Gøtzsche PC. Citation bias: questionable research practice or scientific misconduct? J R Soc Med. 2022;115(1):31-5. https://doi.org/10.1177/01410768221075881

Nickerson RS. Confirmation bias: a ubiquitous phenomenon in many guises. Rev Gen Psychol. 1998;2(2):175-220. https://doi.org/10.1037%2F1089-2680.2.2.175

Research studies

Meppelink CS, Smit EG, Fransen ML, Diviani N. "I was right about vaccination": confirmation bias and health literacy in online health information seeking. J Health Commun. 2019;24(2):129-40. https://doi.org/10.1080/10810730.2019.1583701

1.4c Do not assume that opinions alone are sufficient.

Explanation

People often disagree about the effects of treatments, including doctors, researchers, and patients. This may be because their opinions are not always based on systematic reviews of fair comparisons of treatments. Who makes a treatment claim, how likable they are, or how much experience and expertise they have do not provide a reliable basis for assessing how reliable their claim is. This does not mean that conflicting opinions should be given equal weight – or that the existence of conflicting opinions means that no conclusion can be reached. How much weight to give an opinion should be based on the strength of the evidence supporting it.

Experts, just like everyone else, do not always base what they say on systematic reviews. For example, experts did not begin to recommend aspirin after a heart attack until years after there was strong evidence supporting its use *[Antman 1992 (SR)]*. Conversely, experts continued to recommend medicines to reduce heart rhythm abnormalities years after there was strong evidence that they increased the risk of early death after a heart attack.

Basis for this concept

More than two thirds of Americans often hear conflicting medical information from family and friends *[The Merck Manuals 2021 (RS)]*. Health professionals also often have conflicting opinions, and gaps between research findings and health professional practice are well documented *[Bero 1998 (SR), Bloom 2005 (SR), Boaz 2011 (SR), Grimshaw 2001 (SR)]*. Passive dissemination of research evidence does not adequately ensure that the opinions and practices of health professionals are consistent with the best available evidence.

New research evidence of the effects of treatments is published daily, making it difficult to keep up to date *[Bastian 2010]*. Studies that have examined how often clinicians have questions have found that, on average, clinicians have about one question for every two patients they see, but they only seek answers to about half of those questions *[Del Fiol 2014 (SR)]*.

There is an enormous amount of information about treatments on the Internet, much of which is unreliable *[Eysenbach 2002 (SR), Glenton 2005 (RS)]*. Frequently, the basis for claims about the effects of treatments is not provided. Very few online sources of information about treatments are explicitly based on systematic reviews of fair comparisons, making it difficult to know which opinions to trust *[Oxman 2019 (SR)]*.

Experts may disagree more than non-experts when assessing the quality of reviews of research evidence written by others; and reviews of research evidence written by experts may be, on average, of inferior scientific quality compared to reviews by non-experts *[Oxman 1993 (RS)]*. It appears that the greater the expertise of review authors, the more likely the quality is to be poor. Poor quality of reviews by experts may be related to the strength of their prior opinions and the amount of time they spend preparing a review. Like others, experts are prone to confirmation bias, and articles written by experts tend to selectively cite other articles that support their opinions *[Duyx 2017 (SR)]*. Expert opinion is nearly always based on evidence. The evidence can, for example, be a systematic review of fair comparisons, anecdotal experience, or laboratory studies. The problem is that unless experts are explicit about the basis of their opinions, it is not possible to critically appraise the claims that they make *[Schunemann 2019]*. Many clinical practice guidelines that use the term "expert opinion" when evidence is insufficient or do not provide an explanation *[Ponce 2017 (SR)]*. The expert opinions are based on various types of evidence, most often indirect evidence. Indirect evidence does not directly support the recommendation, for example because of differences between the study participants and the people for whom the recommendation is being made, the treatments compared in the studies and recommendation, or the outcome measure in the study and the outcome of interest *[Guyatt 2011b]*.

New randomized trials of treatments are typically planned and implemented by experts, and protocols for new trials are typically reviewed and approved or not approved by experts. This seems sensible, but the opinions of those experts are not always informed by systematic reviews of previous research, sometimes resulting in unnecessary harm and wasted resources *[Clarke 2014 (SR), Lund 2016]*.

Implications

Do not rely on the opinions of experts or other authorities about the effects of treatments unless they have taken account of the results of systematic reviews of fair comparisons of treatments.

References

Systematic reviews

Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. Treatments for myocardial infarction. JAMA. 1992;268(2):240-8. https://doi.org/10.1001/jama.1992.03490020088036

Bero LA, Grilli R, Grimshaw JM, Harvey E, Oxman AD, Thomson MA. Closing the gap between research and practice: an overview of systematic reviews of interventions to promote the implementation of research findings. The Cochrane Effective Practice and Organization of Care Review Group. BMJ. 1998;317(7156):465-8. https://doi.org/10.1136/bmj.317.7156.465

Bloom BS. Effects of continuing medical education on improving physician clinical care and patient health: a review of systematic reviews. Int J Technol Assess Health Care. 2005;21(3):380-5. https://doi.org/10.1017/s026646230505049x

Boaz A, Baeza J, Fraser A. Effective implementation of research into practice: an overview of systematic reviews of the health literature. BMC Res Notes. 2011;4:212. https://doi.org/10.1186/1756-0500-4-212

Clarke M, Brice A, Chalmers I. Accumulating research: a systematic account of how cumulative meta-analyses would have provided knowledge, improved health, reduced harm and saved resources. PLoS One. 2014;9(7):e102670. https://doi.org/10.1371/journal.pone.0102670

Del Fiol G, Workman TE, Gorman PN. Clinical questions raised by clinicians at the point of care: a systematic review. JAMA Intern Med. 2014;174(5):710-8. https://doi.org/10.1001/jamainternmed.2014.368

Duyx B, Urlings MJE, Swaen GMH, Bouter LM, Zeegers MP. Scientific citations favor positive results: a systematic review and meta-analysis. J Clin Epidemiol. 2017;88:92-101. https://doi.org/10.1016/j.jclinepi.2017.06.002

Eysenbach G, Powell J, Kuss O, Sa ER. Empirical studies assessing the quality of health information for consumers on the world wide web: a systematic review. JAMA. 2002;287(20):2691-700. https://doi.org/10.1001/jama.287.20.2691

Grimshaw JM, Shirran L, Thomas R, Mowatt G, Fraser C, Bero L, et al. Changing provider behavior: an overview of systematic reviews of interventions. Med Care. 2001;39(8 Suppl 2):II2-45. https://journals.lww.com/lww-medicalcare/Fulltext/2001/08002/Changing_Provider_Behavior__An_Overview_of.2.aspx

Oxman AD, Paulsen EJ. Who can you trust? A review of free online sources of "trustworthy" information about treatment effects for patients and the public. BMC Med Inform Decis Mak. 2019;19(1):35. https://doi.org/10.1186/s12911-019-0772-5

Ponce OJ, Alvarez-Villalobos N, Shah R, Mohammed K, Morgan RL, Sultan S, et al. What does expert opinion in guidelines mean? a meta-epidemiological study. Evid Based Med. 2017;22(5):164-9. https://doi.org/10.1136/ebmed-2017-110798

Research studies

Glenton C, Paulsen EJ, Oxman AD. Portals to Wonderland: health portals lead to confusing information about the effects of health care. BMC Med Inform Decis Mak. 2005;5:7. https://doi.org/10.1186/1472-6947-5-7

Oxman AD, Guyatt GH. The science of reviewing research. Ann N Y Acad Sci. 1993;703:125-33; discussion 33-4. https://doi.org/10.1111/j.1749-6632.1993.tb26342.x

The Merck Manuals. Merck Manuals survey: more than two thirds of Americans often hear conflicting medical information from family and friends. CISION PR Newswire. https://www.prnewswire.com/news-releases/merck-manuals-survey-more-than-two-thirds-of-americans-often-hear-conflicting-medical-information-from-family-and-friends-301395024.html

Other references

Bastian H, Glasziou P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? PLoS Med. 2010;7(9):e1000326. https://doi.org/10.1371/journal.pmed.1000326

Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 8. Rating the quality of evidence--indirectness. J Clin Epidemiol. 2011b;64(12):1303-10. https://doi.org/10.1016/j.jclinepi.2011.04.014

Lund H, Brunnhuber K, Juhl C, Robinson K, Leenaars M, Dorch BF, et al. Towards evidence based research. BMJ. 2016;355:i5440. https://doi.org/10.1136/bmj.i5440

Schunemann HJ, Zhang Y, Oxman AD. Distinguishing opinion from evidence in guidelines. BMJ. 2019;366:l4606. https://doi.org/10.1136/bmj.l4606

1.4d Do not assume that peer review and publication is sufficient.

Explanation

Even though a comparison of treatments – whether in a single study or in a review of similar studies – has been published in a prestigious journal, it may not be a fair comparison and the results may not be reliable. Peer review (assessment of a study by others working in the same field) does not guarantee that published studies are reliable. Assessments vary and may not be systematic. Similarly, just because a study is widely publicised does not mean that it is trustworthy.

Sometimes, research that has been peer reviewed and published is so untrustworthy that it is retracted. About half of all retractions involve misconduct, including fabrication or falsification *[Brainard 2018 , Budd 2011]*. Perhaps the most widely-known example of a widely-publicised paper that was subsequently retracted was a small study published in The Lancet which suggested that measles, mumps and rubella vaccination might cause autism *[Flaherty 2011]*. Publication of that paper contributed to vaccine scepticism and led to a decrease in vaccinated children, outbreaks of measles, serious illness, and at least four deaths that could have been prevented.

Although a small proportion of published papers are retracted, many more are corrected or refuted by more reliable research *[Oransky 2021]*. Journals rely on peer review to ensure the quality of the research they publish. However, peer review is highly variable, inconsistent, and flawed *[Smith 2006 , Smith 2010]*. For the most part it is done by volunteers. Few peer reviewers have formal training and they commonly do not detect major errors. For example, the British Medical Journal (BMJ) sent three papers, each of which had nine major methodological errors inserted, to about 600 peer reviewers *[Schroter 2008 (RS)]*. On average, the peer reviewers detected about one-third of the errors in each paper. Half of the peer reviewers were given brief training, which had only a slight impact on improving error detection.

Basis for this concept

Published information written for busy decision makers sometimes contains misleading information on the effects of treatments *[Antman 1992 (SR)]*. Published, peer-reviewed comparisons of treatments often have a high risk of bias, which can result in overestimating or underestimating the effects and cost-effectiveness of treatments *[Bell 2006 (SR), Page 2016a (SR), Savović 2012a (SR), Savović 2012b (SR)]*. Before accepting the results of published randomized trials or systematic reviews, decision makers should critically appraise their methods to identify sources of bias *[Guyatt 1993 , Oxman 1994]*.

Published reports of randomized trials frequently fail to consider the results in the context of prior trials *[Robinson 2011 (SR)]*, and sometimes selectively cite other research *[Duyx 2017 (SR)]*. In addition, published reports of randomized trials are often inconsistent with their protocols, and "statistically significant" results are more likely to be reported than results that are not statistically significant *[Dwan 2013 (SR)]*.

Reports of randomized trials are often inadequate for assessing the validity of study results *[Haidich 2011 (SR), Hopewell 2010 (SR), Mills 2005 (SR)]*. Although reporting of randomized trials has improved, there is still room for further improvement *[To 2013 (SR)]*.

Editorial peer review is used as a tool to assess and improve the quality of submissions to journals. However, there is very little evidence of the effects of peer review on the quality of published research evidence *[Jefferson 2007]*. Judgements about the quality of information are often based on the reputation of the journal. However, this does not guarantee high quality information. Journal impact factor, a measure that reflects the prestige of a journal, may have little or no association with the quality of published research *[Masic 2020 (RS), Pölkki 2014 (SR), Saginur 2020 (SR)]*.

Published studies that show benefits, especially large benefits, are more likely to be noticed than studies that do not *[Duyx 2017 (SR)*, *Ioannidis 2005 (SR)]*, but they are not necessarily trustworthy. Many published studies are too small to have reliable results, and small studies are more likely to report extreme results than large studies *[Schwab 2021 (SR)]*. Subsequent studies, which often contradict those studies or show smaller benefits, *[Ioannidis 2005 (SR)*, *Serra-Garcia 2021 (SR)]*, are accorded less attention *[Serra-Garcia 2021 (SR)]*. Research reports commonly emphasise findings that suggest benefits, while ignoring other findings *[Chiu 2017 (SR)]*. Press releases are often designed to attract favourable media attention and news reports of those studies do the same *[Yavchitz 2012 (RS)]*. News reports about published comparisons of treatments often do not consider the reliability of the results *[Oxman 2022 (SR)]*.

Implications

Always consider whether a published comparison of the effects of treatments is fair and whether the results are reliable. Peer review is a poor indicator of reliability.

References

Systematic reviews

Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. Treatments for myocardial infarction. JAMA. 1992;268(2):240-8. https://doi.org/10.1001/jama.1992.03490020088036

Bell CM, Urbach DR, Ray JG, Bayoumi A, Rosen AB, Greenberg D, et al. Bias in published cost effectiveness studies: systematic review. BMJ. 2006;332(7543):699-703. https://doi.org/10.1136/bmj.38737.607558.80

Chiu K, Grundy Q, Bero L. 'Spin' in published biomedical literature: a methodological systematic review. PLoS Biol. 2017;15(9):e2002173. https://doi.org/10.1371/journal.pbio.2002173

Duyx B, Urlings MJE, Swaen GMH, Bouter LM, Zeegers MP. Scientific citations favor positive results: a systematic review and meta-analysis. J Clin Epidemiol. 2017;88:92-101. https://doi.org/10.1016/j.jclinepi.2017.06.002

Dwan K, Gamble C, Williamson PR, Kirkham JJ. Systematic review of the empirical evidence of study publication bias and outcome reporting bias - an updated review. PLoS One. 2013;8(7):e66844. https://doi.org/10.1371/journal.pone.0066844

Haidich AB, Birtsou C, Dardavessis T, Tirodimos I, Arvanitidou M. The quality of safety reporting in trials is still suboptimal: survey of major general medical journals. J Clin Epidemiol. 2011;64(2):124-35. https://doi.org/10.1016/j.jclinepi.2010.03.005

Hopewell S, Dutton S, Yu LM, Chan AW, Altman DG. The quality of reports of randomised trials in 2000 and 2006: comparative study of articles indexed in PubMed. BMJ. 2010;340:c723. https://doi.org/10.1136/bmj.c723

Ioannidis JPA. Contradicted and initially stronger effects in highly cited clinical research. JAMA. 2005;294(2):218-28. https://doi.org/10.1001/jama.294.2.218

Mills EJ, Wu P, Gagnier J, Devereaux PJ. The quality of randomized trial reporting in leading medical journals since the revised CONSORT statement. Contemp Clin Trials. 2005;26(4):480-7. https://doi.org/10.1016/j.cct.2005.02.008

Oxman M, Larun L, Gaxiola GP, Alsaid D, Qasim A, Rose CJ, et al. Quality of information in news media reports about the effects of health interventions: systematic review and meta-analyses. F1000Res. 2022;10:433. https://doi.org/10.12688/f1000research.52894.2

Page MJ, Higgins JP, Clayton G, Sterne JA, Hróbjartsson A, Savović J. Empirical evidence of study design biases in randomized trials: systematic review of meta-epidemiological studies. PLoS One. 2016a;11(7):e0159267. https://doi.org/10.1371/journal.pone.0159267

Pölkki T, Kanste O, Kääriäinen M, Elo S, Kyngäs H. The methodological quality of systematic reviews published in high-impact nursing journals: a review of the literature. J Clin Nurs. 2014;23(3-4):315-32. https://doi.org/10.1111/jocn.12132

Robinson KA, Goodman SN. A systematic examination of the citation of prior research in reports of randomized, controlled trials. Ann Intern Med. 2011;154(1):50-5. https://doi.org/10.7326/0003-4819-154-1-201101040-00007

Saginur M, Fergusson D, Zhang T, Yeates K, Ramsay T, Wells G, et al. Journal impact factor, trial effect size, and methodological quality appear scantly related: a systematic review and meta-analysis. Syst Rev. 2020;9(1):53. https://doi.org/10.1186/s13643-020-01305-w

Savović J, Jones HE, Altman DG, Harris RJ, Jüni P, Pildal J, et al. Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. Ann Intern Med. 2012b;157(6):429-38. https://doi.org/10.7326/0003-4819-157-6-201209180-00537

Schwab S, Kreiliger G, Held L. Assessing treatment effects and publication bias across different specialties in medicine: a meta-epidemiological study. BMJ Open. 2021;11(9):e045942. https://doi.org/10.1136/bmjopen-2020-045942

Serra-Garcia M, Gneezy U. Nonreplicable publications are cited more than replicable ones. Sci Adv. 2021;7(21):eabd1705. https://advances.sciencemag.org/content/advances/7/21/eabd1705.full.pdf

To MJ, Jones J, Emara M, Jadad AR. Are reports of randomized controlled trials improving over time? A systematic review of 284 articles published in high-impact general and specialized medical journals. PLoS One. 2013;8(12):e84779. https://doi.org/10.1371/journal.pone.0084779

Research studies

Masic I, Jankovic SM. Meta-analysing methodological quality of published research: importance and effectiveness. Stud Health Technol Inform. 2020;272:229-32. https://doi.org/10.3233/shti200536

Savović J, Jones H, Altman D, Harris R, Jűni P, Pildal J, et al. Influence of reported study design characteristics on intervention effect estimates from randomised controlled trials: combined analysis of meta-epidemiological studies. Health Technol Assess. 2012a;16(35):1-82. https://doi.org/10.3310/hta16350

Schroter S, Black N, Evans S, Godlee F, Osorio L, Smith R. What errors do peer reviewers detect, and does training improve their ability to detect them? J R Soc Med. 2008;101(10):507-14. https://doi.org/10.1258/jrsm.2008.080062

Yavchitz A, Boutron I, Bafeta A, Marroun I, Charles P, Mantz J, et al. Misrepresentation of randomized controlled trials in press releases and news coverage: a cohort study. PLoS Med. 2012;9(9):e1001308. https://doi.org/10.1371/journal.pmed.1001308

Other references

Brainard J, You J. What a massive database of retracted papers reveals about science publishing's 'death penalty'. Science. 2018;25(1):1-5. https://www.science.org/content/article/what-massive-database-retracted-papers-reveals-about-science-publishing-s-death-penalty

Budd JM, Coble ZC, Anderson KM. Retracted publications in biomedicine: cause for concern. Association of College & Research Libraries Conference Program; Philadelphia,2011. https://www.ala.org/acrl/files/conferences/confsandpreconfs/national/2011/papers/retracted_publicatio.pdf

Flaherty DK. The vaccine-autism connection: a public health crisis caused by unethical medical practices and fraudulent science. Ann Pharmacother. 2011;45(10):1302-4. https://doi.org/10.1345/aph.1q318

Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature. II. How to use an article about therapy or prevention. A. Are the results of the study valid? Evidence-Based Medicine Working Group. JAMA. 1993;270(21):2598-601. https://doi.org/10.1001/jama.1993.03510210084032

Jefferson T, Rudin M, Brodney Folse S, Davidoff F. Editorial peer review for improving the quality of reports of biomedical studies. Cochrane Database Syst Rev. 2007(2). https://doi.org//10.1002/14651858.MR000016.pub3

Oransky I, Fremes SE, Kurlansky P, Gaudino M. Retractions in medicine: the tip of the iceberg. Eur Heart J. 2021. https://doi.org/10.1093/eurheartj/ehab398

Oxman AD, Cook DJ, Guyatt GH. Users' guides to the medical literature. VI. How to use an overview. Evidence-Based Medicine Working Group. JAMA. 1994;272(17):1367-71. https://doi.org/10.1001/jama.272.17.1367

Smith R. Peer review: a flawed process at the heart of science and journals. J R Soc Med. 2006;99(4):178-82. https://doi.org/10.1258/jrsm.99.4.178

Smith R. Classical peer review: an empty gun. Breast Cancer Res. 2010;12 Suppl 4(Suppl 4):S13. https://doi.org/10.1186/bcr2742

1.4e Do not assume that there are no competing interests.

Explanation

People with an interest in promoting a treatment (in addition to wanting to help people) – for example, to make money – may promote treatments by exaggerating benefits, ignoring potential harmful effects, cherry picking which information is used, or making false claims. Conversely, people may be opposed to a treatment for a range of reasons, such as cultural practices.

Tamiflu (oseltamivir) is an example of how financial conflicts of interest can result in misleading claims about the effects of a treatment *[Doshi 2012 , Loder 2014]*. Tamiflu was approved for seasonal influenza by the U.S. Food and Drug Administration in 1999. Several randomized trials and systematic reviews emphasised the benefits and safety of Tamiflu. Most of them were funded by Roche, which also marketed and promoted Tamiflu. In 2005 and 2009, the fear of pandemic flu led to recommendations to stockpile Tamiflu and billions of dollars were spent on this. After battling with the company for over four years, a team of review authors finally accessed the complete data held by the company. After carefully reviewing all the documents, they found no compelling evidence to support claims that oseltamivir reduces the risk of complications of influenza, such as pneumonia and hospital admission, claims that had been used to justify international stockpiling of the drug *[Jefferson 2014 (SR)]*. Tamiflu was found to slightly reduce the time to alleviation of flu symptoms in adults and to slightly reduce the risk of flu symptoms in people exposed to the flu. It was also found to have adverse effects that potentially outweighed the benefits. As a result of biased reporting of the research and misinformed recommendations and decisions, billions of dollars were wasted.

Basis for this concept

Financial conflicts of interests can lead to bias in several ways *[Dunn 2016 (OR)]*. Researchers with conflicts of interest are more likely to choose less effective control comparison treatments, leading to more favourable results for a new drug *[Dunn 2013 (SR), Hugenholtz 2006 (SR), Lathyris 2010 (SR)]*. They may be more likely to selectively report outcomes that favour the treatment and not to publish the results of a trial if it does not favour the treatment *[Dwan 2013 (SR)]*. They also may be more likely to draw conclusions and recommend the treatment *[Als-Nielsen 2003 (SR), Chartres 2016 (SR), Yank 2007 (SR)]*.

Studies of pharmaceuticals, devices, and dental implants that have been sponsored by the manufacturing company have more favourable results and conclusions than studies sponsored by other sources of support *[Lundh 2017 (SR), Popelut 2010 (SR), Saltaji 2021 (SR)]*.

Review authors may also be more likely to interpret results favourably when they have financial conflicts of interest *[Barnes 1998 (SR), Bes-Rastrollo 2013 (SR), Dunn 2014 (SR), Jørgensen 2006 (SR), Mandrioli 2016 (SR)]*. Cost-effectiveness studies funded by industry are more likely to present favourable results than other studies *[Bell 2006 (SR)]*, and authors of clinical practice guidelines may be more likely to recommend a treatment when they have a financial conflict of interest *[Nejstgaard 2020 (SR), Norris 2011 (SR), Tabatabavakili 2021 (SR)]*.

Promotion of treatments is regulated in many countries. Nonetheless, advertisements are frequently misleading *[Every-Palmer 2014 , Faerber 2012 (RS), Folsom 2010 (RS), Huang 2007 (OR), Klara 2018 (RS), Morganroth 2009 (RS), Othman 2009 (SR), Salas 2008 (RS), Sansgiry 1999 (RS), Spielmans 2008 (RS), Vendra 2019 (RS), Wayant 2020 (RS)]*. Because vitamin and mineral supplements are regulated as foods rather than treatments in the U.S., they are not regulated in the same way as treatments. Thus, supplement manufacturers can market, sell, and obtain substantial profit from a supplement despite uncertain benefits and potential harms *[McCormick 2010 (OR)]*. Expenditures on supplements in the U.S. were estimated to be $21–25 billion a year in 2010, and increasing.

A majority of health news reports do not consider conflicts of interest *[Oxman 2022 (SR)]*.

The assumption that non-financial conflicts of interest can influence the outcomes of treatment comparisons, reviews, and recommendations is logical, but in contrast to financial conflicts of interest, there is little evidence of biased effects of non-financial conflicts of interest *[Akl 2014 (RS), Bero 2014]*.

Implications

Ask if people making claims that a treatment is effective have conflicting interests. If they do, be careful not to be misled by their claims about the effects of treatments.

References

Systematic reviews

Als-Nielsen B, Chen W, Gluud C, Kjaergard LL. Association of funding and conclusions in randomized drug trials: a reflection of treatment effect or adverse events? JAMA. 2003;290(7):921-8. https://doi.org/10.1001/jama.290.7.921

Barnes DE, Bero LA. Why review articles on the health effects of passive smoking reach different conclusions. JAMA. 1998;279(19):1566-70. https://doi.org/10.1001/jama.279.19.1566

Bell CM, Urbach DR, Ray JG, Bayoumi A, Rosen AB, Greenberg D, et al. Bias in published cost effectiveness studies: systematic review. BMJ. 2006;332(7543):699-703. https://doi.org/10.1136/bmj.38737.607558.80

Bes-Rastrollo M, Schulze MB, Ruiz-Canela M, Martinez-Gonzalez MA. Financial conflicts of interest and reporting bias regarding the association between sugar-sweetened beverages and weight gain: a systematic review of systematic reviews. PLoS Med. 2013;10(12):e1001578; dicsussion e. https://doi.org/10.1371/journal.pmed.1001578

Chartres N, Fabbri A, Bero LA. Association of Industry Sponsorship With Outcomes of Nutrition Studies: A Systematic Review and Meta-analysis. JAMA Intern Med. 2016;176(12):1769-77. https://doi.org/10.1001/jamainternmed.2016.6721

Dunn AG, Arachi D, Hudgins J, Tsafnat G, Coiera E, Bourgeois FT. Financial conflicts of interest and conclusions about neuraminidase inhibitors for influenza: an analysis of systematic reviews. Ann Intern Med. 2014;161(7):513-8. https://doi.org/10.7326/m14-0933

Dunn AG, Mandl KD, Coiera E, Bourgeois FT. The effects of industry sponsorship on comparator selection in trial registrations for neuropsychiatric conditions in children. PLoS One. 2013;8(12):e84951. https://doi.org/10.1371/journal.pone.0084951

Dwan K, Gamble C, Williamson PR, Kirkham JJ. Systematic review of the empirical evidence of study publication bias and outcome reporting bias - an updated review. PLoS One. 2013;8(7):e66844. https://doi.org/10.1371/journal.pone.0066844

Hugenholtz GW, Heerdink ER, Stolker JJ, Meijer WE, Egberts AC, Nolen WA. Haloperidol dose when used as active comparator in randomized controlled trials with atypical antipsychotics in schizophrenia: comparison with officially recommended doses. J Clin Psychiatry. 2006;67(6):897-903. https://doi.org/10.4088/jcp.v67n0606

Jefferson T, Jones MA, Doshi P, Del Mar CB, Hama R, Thompson MJ, et al. Neuraminidase inhibitors for preventing and treating influenza in adults and children. Cochrane Database Syst Rev. 2014;2014(4):Cd008965. https://doi.org/10.1002/14651858.cd008965.pub4

Jørgensen AW, Hilden J, Gøtzsche PC. Cochrane reviews compared with industry supported meta-analyses and other meta-analyses of the same drugs: systematic review. BMJ. 2006;333(7572):782. https://doi.org/10.1136/bmj.38973.444699.0b

Lathyris DN, Patsopoulos NA, Salanti G, Ioannidis JP. Industry sponsorship and selection of comparators in randomized clinical trials. Eur J Clin Invest. 2010;40(2):172-82. https://doi.org/10.1111/j.1365-2362.2009.02240.x

Lundh A, Lexchin J, Mintzes B, Schroll JB, Bero L. Industry sponsorship and research outcome. Cochrane Database Syst Rev. 2017(2). https://doi.org//10.1002/14651858.MR000033.pub3

Mandrioli D, Kearns CE, Bero LA. Relationship between Research Outcomes and Risk of Bias, Study Sponsorship, and Author Financial Conflicts of Interest in Reviews of the Effects of Artificially Sweetened Beverages on Weight Outcomes: A Systematic Review of Reviews. PLoS One. 2016;11(9):e0162198. https://doi.org/10.1371/journal.pone.0162198

Nejstgaard CH, Bero L, Hróbjartsson A, Jørgensen AW, Jørgensen KJ, Le M, et al. Association between conflicts of interest and favourable recommendations in clinical guidelines, advisory committee reports, opinion pieces, and narrative reviews: systematic review. BMJ. 2020;371:m4234. https://doi.org/10.1136/bmj.m4234

Norris SL, Holmer HK, Ogden LA, Burda BU. Conflict of interest in clinical practice guideline development: a systematic review. PLoS One. 2011;6(10):e25153. https://doi.org/10.1371/journal.pone.0025153

Othman N, Vitry A, Roughead EE. Quality of pharmaceutical advertisements in medical journals: a systematic review. PLoS One. 2009;4(7):e6350. https://doi.org/10.1371/journal.pone.0006350

Oxman M, Larun L, Gaxiola GP, Alsaid D, Qasim A, Rose CJ, et al. Quality of information in news media reports about the effects of health interventions: systematic review and meta-analyses. F1000Res. 2022;10:433. https://doi.org/10.12688/f1000research.52894.2

Popelut A, Valet F, Fromentin O, Thomas A, Bouchard P. Relationship between sponsorship and failure rate of dental implants: a systematic approach. PLoS One. 2010;5(4):e10274. https://doi.org/10.1371/journal.pone.0010274

Saltaji H, Armijo-Olivo S, Cummings GG, Amin M, Major PW, da Costa BR, et al. Influence of sponsorship bias on treatment effect size estimates in randomized trials of oral health interventions: a meta-epidemiological study. J Evid Based Dent Pract. 2021;21(2):101544. https://doi.org/10.1016/j.jebdp.2021.101544

Tabatabavakili S, Khan R, Scaffidi MA, Gimpaya N, Lightfoot D, Grover SC. Financial Conflicts of Interest in Clinical Practice Guidelines: A Systematic Review. Mayo Clin Proc Innov Qual Outcomes. 2021;5(2):466-75. https://doi.org/10.1016/j.mayocpiqo.2020.09.016

Yank V, Rennie D, Bero LA. Financial ties and concordance between results and conclusions in meta-analyses: retrospective cohort study. BMJ. 2007;335(7631):1202-5. https://doi.org/10.1136/bmj.39376.447211.be

Other reviews

Dunn AG, Coiera E, Mandl KD, Bourgeois FT. Conflict of interest disclosure in biomedical research: A review of current practices, biases, and the role of public registries in improving transparency. Res Integr Peer Rev. 2016;1. https://doi.org/10.1186/s41073-016-0006-7

Huang CK, Miller TA. The truth about over-the-counter topical anti-aging products: a comprehensive review. Aesthet Surg J. 2007;27(4):402-12; quiz 13-5. https://doi.org/10.1016/j.asj.2007.05.005

McCormick DB. Vitamin/mineral supplements: of questionable benefit for the general population. Nutr Rev. 2010;68(4):207-13. https://doi.org/10.1111/j.1753-4887.2010.00279.x

Research studies

Akl EA, El-Hachem P, Abou-Haidar H, Neumann I, Schünemann HJ, Guyatt GH. Considering intellectual, in addition to financial, conflicts of interest proved important in a clinical practice guideline: a descriptive study. J Clin Epidemiol. 2014;67(11):1222-8. https://doi.org/10.1016/j.jclinepi.2014.05.006

Faerber AE, Kreling DH. Now you see it. Now you don't: fair balance and adequate provision in advertisements for drugs before and after the switch from prescription to over-the-counter. Health Commun. 2012;27(1):66-74. https://doi.org/10.1080/10410236.2011.569001

Folsom C, Fesperman SF, Tojuola B, Sultan S, Dahm P. Direct-to-consumer advertising for urological pharmaceuticals: a cross-sectional analysis of print media. Urology. 2010;75(5):1029-33. https://doi.org/10.1016/j.urology.2009.10.053

Klara K, Kim J, Ross JS. Direct-to-consumer broadcast advertisements for pharmaceuticals: off-label promotion and adherence to FDA guidelines. J Gen Intern Med. 2018;33(5):651-8. https://doi.org/10.1007/s11606-017-4274-9

Morganroth P, Wilmot AC, Miller C. JAAD online. Over-the-counter scar products for postsurgical patients: disparities between online advertised benefits and evidence regarding efficacy. J Am Acad Dermatol. 2009;61(6):e31-47. https://doi.org/10.1016/j.jaad.2009.02.046

Salas M, Martin M, Pisu M, McCall E, Zuluaga A, Glasser SP. Analysis of US Food and Drug Administration warning letters: false promotional claims relating to prescription and over-the-counter medications. Pharmaceut Med. 2008;22(2). https://doi.org/10.1007/bf03256691

Sansgiry S, Sharp WT, Sansgiry SS. Accuracy of information on printed over-the-counter drug advertisements. Health Mark Q. 1999;17(2):7-18. https://doi.org/10.1300/j026v17n02_02

Spielmans GI, Thielges SA, Dent AL, Greenberg RP. The accuracy of psychiatric medication advertisements in medical journals. J Nerv Ment Dis. 2008;196(4):267-73. https://doi.org/10.1097/nmd.0b013e31816a436b

Vendra V, Vaisbuch Y, Mudry AC, Jackler RK. Over-the-counter tinnitus "cures": marketers' promises do not ring true. Laryngoscope. 2019;129(8):1898-906. https://doi.org/10.1002/lary.27677

Wayant C, Aran G, Johnson BS, Vassar M. Evaluation of selective outcome reporting bias in efficacy endpoints in print and television advertisements for oncology drugs. J Gen Intern Med. 2020;35(10):2853-7. https://doi.org/10.1007/s11606-020-06028-1

Other references

Bero L. What is in a name? Nonfinancial influences on the outcomes of systematic reviews and guidelines. J Clin Epidemiol. 2014;67(11):1239-41. https://doi.org/10.1016/j.jclinepi.2014.06.015

Doshi P, Jefferson T, Del Mar C. The imperative to share clinical study reports: recommendations from the Tamiflu experience. PLoS Med. 2012;9(4):e1001201. https://doi.org/10.1371/journal.pmed.1001201

Every-Palmer S, Duggal R, Menkes DB. Direct-to-consumer advertising of prescription medication in New Zealand. NZ Med J. 2014;127(1401):102-10. https://www.proquest.com/openview/652f0e94979ec70b273d1c064c4e5f37/1?pq-origsite=gscholar&cbl=1056335

Loder E, Tovey D, Godlee F. The Tamiflu trials. BMJ. 2014;348:g2630. https://doi.org/10.1136/bmj.g2630

2. Comparisons

Studies should make fair comparisons, designed to minimize the risk of systematic errors (biases) and random errors (the play of chance).

2.1 Comparisons of treatments should be fair.

2.1a Consider whether the people being compared were similar.

Explanation

If people in treatment comparison groups differ in ways other than the treatments being compared, the apparent effects of the treatments might reflect those differences rather than actual treatment effects. Differences in the characteristics of the people in the comparison groups at the beginning of the comparison might result in estimates of treatment effects that appear either larger or smaller than they actually are. A method such as allocating people to different treatments by assigning them random numbers (the equivalent of flipping a coin) is the best way to ensure that the groups being compared are similar in terms of both measured and unmeasured characteristics.

If people are not randomly allocated to treatment comparison groups, differences between the groups other than the treatments may result in estimates of treatment effects appearing larger or smaller than they actually are because of confounders or other differences. For example, patients who are most ill (e.g., have severe pain) may be more likely to be given a new treatment than patients who are less ill. There may appear to be a sharp response to treatment in the most ill patients because of regression to the mean. If they are compared to patients who are less ill and receive an older treatment, the new treatment may appear to be more effective than it actually is compared to the older treatment. Differences in recall ("recall bias") can also lead to over- or under-estimates of effects in case-control and retrospective cohort studies that are based on recollection of exposure to a treatment.

As described in relation to Concept 1.2b, the effect of hormone replacement therapy (HRT) on cardiovascular disease (CVD) is an example of overestimation of a treatment effect in non-randomized studies. For many years experts and doctors believed that HRT reduced the risk of CVD, based on non-randomized studies. But the results of large, randomized trials provided no support for this belief and sometimes suggested an increased risk of CVD in women assigned to HRT. This may be because women of lower socio-economic status are more likely to have CVD and less likely to take HRT. So, a reason for the apparent beneficial effect of HRT on CVD in non-randomized studies is the difference in socioeconomic status between the comparison groups, rather than the difference in whether they took HRT or not *[Humphrey 2002 (SR)]*.

Quinidine is an example of a treatment for which a beneficial effect appeared smaller in non-randomized studies when compared to those in randomized studies. Quinidine was frequently used to treat heart rhythm abnormalities (atrial fibrillation).[1] A systematic review of randomized and non-randomized studies found that the beneficial effect of maintaining a normal heart rhythm was 54% less after three months and 76%

[1] Although quinidine was effective for maintaining a normal heart rhythm, it has been replaced by safer and more effective medicines.

less after 12 months in non-randomized studies when compared with randomized studies [Reimold 1992 (SR)]. One possible explanation for the apparently smaller effects in the non-randomized studies is that patients with the most symptoms and the highest risk may have been more likely to receive quinidine in the non-randomized studies.

Aspirin is an example of a treatment where a harmful effect appeared larger in non-randomized studies when compared to randomized studies. Randomized studies have shown that low-dose aspirin reduces the risk of stroke in people at high risk (with symptoms and signs of vascular disease) but not in people at low risk. A systematic review of randomized and non-randomized studies found an increased risk of stroke in people at low risk who took aspirin, whereas randomized studies did not find an increased risk [Hart 2000 (SR)]. Aspirin use in the non-randomized studies was largely self-selected and it is possible that people who chose to take aspirin had a higher risk of stroke than those who did not, even after statistical adjustment for risk factors that were known and had been measured.

Basis for this concept

Random allocation of people to comparison groups is unbiased with respect to prognosis (characteristics of participants that can predict the course and outcome of a condition) and responsiveness to the treatment. No other way of creating comparison groups has these properties because it cannot be assumed that all factors relevant to prognosis and responsiveness to treatment have been distributed in an unbiased way between comparison groups [Kleijnen 1997]. However, when a small number of people are randomly allocated, important differences between comparison groups can occur by chance. Moreover, both randomized studies and non-randomized studies can be misleading for other reasons [Sterne 2016], including those addressed by Key Concepts 2.1b to 2.1g.

Comparisons of the results of randomized and non-randomized studies have found that carefully designed and implemented non-randomized studies and randomized studies sometimes give similar estimates of the effects of treatments [Anglemyer 2014 (SR), Bun 2020 (SR), Concato 2000 (SR), Golder 2011 (SR), Schwingshackl 2021 (SR)]. However, non-randomized comparisons of treatments can overestimate effects, underestimate effects, mask effects, or reverse the direction of effects [Deeks 2003 (SR), Ewald 2020 (SR), Hemkens 2016a (SR), Ioannidis 2001 (SR), Kunz 1998 (SR), Odgaard-Jensen 2011 (SR)]. It is a paradox that the unpredictability of randomization is the best protection against the unpredictability of the extent and direction of bias in treatment comparisons that are not properly randomized.

To ensure that people in treatment comparison groups are similar, in addition to randomly allocating enough people, it is important to ensure that random allocation is properly implemented. Researchers have investigated the impact of two key elements of random allocation: adequate generation of a random sequence (to ensure that the allocation sequence is unpredictable, and that people are allocated by chance), and concealed allocation (to ensure that the random sequence is properly implemented, and that participation is not influenced by knowing the treatment assignment prior to enrolment in the study. A systematic review combined the data from seven studies that investigated the influence of these and other characteristics of randomized trials on effect estimates [Savović 2012b (SR)]. It included 234 meta-analyses containing 1,973 randomized trials. It found that, on average, effects were overestimated in trials with inadequate or unclear (compared with adequate) random-sequence generation and with inadequate or unclear (compared with adequate) allocation concealment. A systematic review of 24 studies found similar results [Page 2016a (SR)]. A review of 56 studies that examined associations between 58 different trial characteristics and effect estimates found that allocation concealment, sequence generation, and small sample size were the characteristics most consistently associated with treatment effect estimates [Dechartres 2016 (SR)]. However, it is not generally possible to predict the magnitude, or even the direction, of bias in studies with inadequate or unclear random-sequence generation or allocation concealment [Armijo-Olivo 2015 (SR), Bialy 2014 (SR), Bolvig 2018 (SR), Ginnerup-Nielsen 2016 (SR), Hartling 2014 (SR), Koletsi 2016 (SR), Odgaard-Jensen 2011 (SR), Saltaji 2018 (SR), Wang 2021 (SR)].

Implications

Be cautious about relying on the results of non-randomized treatment comparisons (for example, if the people being compared chose which treatment they received). Be particularly cautious when you cannot be confident that the characteristics of the comparison groups are similar. If people were *not* randomly allocated to treatment comparison groups, ask if there were important differences between the groups that might have resulted in the estimates of treatment effects appearing either larger or smaller than they actually are.

References

Systematic reviews

Anglemyer A, Horvath HT, Bero L. Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. Cochrane Database Syst Rev. 2014(4). https://doi.org//10.1002/14651858.MR000034.pub2

Armijo-Olivo S, Saltaji H, da Costa BR, Fuentes J, Ha C, Cummings GG. What is the influence of randomisation sequence generation and allocation concealment on treatment effects of physical therapy trials? A meta-epidemiological study. BMJ Open. 2015;5(9):e008562. https://doi.org/10.1136/bmjopen-2015-008562

Bialy L, Vandermeer B, Lacaze-Masmonteil T, Dryden DM, Hartling L. A meta-epidemiological study to examine the association between bias and treatment effects in neonatal trials. Evid Based Child Health. 2014;9(4):1052-9. https://doi.org/10.1002/ebch.1985

Bolvig J, Juhl CB, Boutron I, Tugwell P, Ghogomu EAT, Pardo JP, et al. Some Cochrane risk-of-bias items are not important in osteoarthritis trials: a meta-epidemiological study based on Cochrane reviews. J Clin Epidemiol. 2018;95:128-36. https://doi.org/10.1016/j.jclinepi.2017.11.026

Bun RS, Scheer J, Guillo S, Tubach F, Dechartres A. Meta-analyses frequently pooled different study types together: a meta-epidemiological study. J Clin Epidemiol. 2020;118:18-28. https://doi.org/10.1016/j.jclinepi.2019.10.013

Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. N Engl J Med. 2000;342(25):1887-92. https://doi.org/10.1056/nejm200006223422507

Dechartres A, Trinquart L, Faber T, Ravaud P. Empirical evaluation of which trial characteristics are associated with treatment effect estimates. J Clin Epidemiol. 2016;77:24-37. https://doi.org/10.1016/j.jclinepi.2016.04.005

Deeks JJ, Dinnes J, D'Amico R, Sowden AJ, Sakarovitch C, Song F, et al. Evaluating non-randomised intervention studies. Health Technol Assess. 2003;7(27):iii-x, 1-173. https://doi.org/10.3310/hta7270

Ewald H, Ioannidis JPA, Ladanie A, Mc Cord K, Bucher HC, Hemkens LG. Nonrandomized studies using causal-modeling may give different answers than RCTs: a meta-epidemiological study. J Clin Epidemiol. 2020;118:29-41. https://doi.org/10.1016/j.jclinepi.2019.10.012

Ginnerup-Nielsen E, Christensen R, Thorborg K, Tarp S, Henriksen M. Physiotherapy for pain: a meta-epidemiological study of randomised trials. Br J Sports Med. 2016;50(16):965-71. https://doi.org/10.1136/bjsports-2015-095741

Golder S, Loke YK, Bland M. Meta-analyses of adverse effects data derived from randomised controlled trials as compared to observational studies: methodological overview. PLoS Med. 2011;8(5):e1001026. https://doi.org/10.1371/journal.pmed.1001026

Hart RG, Halperin JL, McBride R, Benavente O, Man-Son-Hing M, Kronmal RA. Aspirin for the primary prevention of stroke and other major vascular events: meta-analysis and hypotheses. Arch Neurol. 2000;57(3):326-32. https://doi.org/10.1001/archneur.57.3.326

Hartling L, Hamm MP, Fernandes RM, Dryden DM, Vandermeer B. Quantifying bias in randomized controlled trials in child health: a meta-epidemiological study. PLoS One. 2014;9(2):e88008. https://doi.org/10.1371/journal.pone.0088008

Hemkens LG, Contopoulos-Ioannidis DG, Ioannidis JP. Agreement of treatment effects for mortality from routinely collected data and subsequent randomized trials: meta-epidemiological survey. BMJ. 2016a;352:i493. https://doi.org/10.1136/bmj.i493

Humphrey LL, Chan BK, Sox HC. Postmenopausal hormone replacement therapy and the primary prevention of cardiovascular disease. Ann Intern Med. 2002;137(4):273-84. https://doi.org/10.7326/0003-4819-137-4-200208200-00012

Ioannidis JP, Haidich AB, Pappa M, Pantazis N, Kokori SI, Tektonidou MG, et al. Comparison of evidence of treatment effects in randomized and nonrandomized studies. JAMA. 2001;286(7):821-30. https://doi.org/10.1001/jama.286.7.821

Koletsi D, Spineli LM, Lempesi E, Pandis N. Risk of bias and magnitude of effect in orthodontic randomized controlled trials: a meta-epidemiological review. Eur J Orthod. 2016;38(3):308-12. https://doi.org/10.1093/ejo/cjv049

Kunz R, Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. BMJ. 1998;317(7167):1185-90. https://doi.org/10.1136/bmj.317.7167.1185

Odgaard-Jensen J, Vist GE, Timmer A, Kunz R, Akl EA, Schunemann H, et al. Randomisation to protect against selection bias in healthcare trials. Cochrane Database Syst Rev. 2011(4):MR000012. https://doi.org/10.1002/14651858.mr000012.pub3

Page MJ, Higgins JP, Clayton G, Sterne JA, Hróbjartsson A, Savović J. Empirical evidence of study design biases in randomized trials: systematic review of meta-epidemiological studies. PLoS One. 2016a;11(7):e0159267. https://doi.org/10.1371/journal.pone.0159267

Reimold SC, Chalmers TC, Berlin JA, Antman EM. Assessment of the efficacy and safety of antiarrhythmic therapy for chronic atrial fibrillation: observations on the role of trial design and implications of drug-related mortality. Am Heart J. 1992;124(4):924-32. https://doi.org/10.1016/0002-8703(92)90974-z

Saltaji H, Armijo-Olivo S, Cummings GG, Amin M, da Costa BR, Flores-Mir C. Impact of Selection Bias on Treatment Effect Size Estimates in Randomized Trials of Oral Health Interventions: A Meta-epidemiological Study. J Dent Res. 2018;97(1):5-13. https://doi.org/10.1177/0022034517725049

Savović J, Jones HE, Altman DG, Harris RJ, Jüni P, Pildal J, et al. Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. Ann Intern Med. 2012b;157(6):429-38. https://doi.org/10.7326/0003-4819-157-6-201209180-00537

Wang Z, Alahdab F, Farah M, Seisa M, Firwana M, Rajjoub R, et al. Association of study design features and treatment effects in trials of chronic medical conditions: a meta-epidemiological study. BMJ Evid Based Med. 2021. https://doi.org/10.1136/bmjebm-2021-111667

Research studies

Schwingshackl L, Balduzzi S, Beyerbach J, Bröckelmann N, Werner SS, Zähringer J, et al. Evaluating agreement between bodies of evidence from randomised controlled trials and cohort studies in nutrition research: meta-epidemiological study. BMJ. 2021;374:n1864. https://doi.org/10.1136/bmj.n1864

Other references

Kleijnen J, Gøtzsche P, Kunz RA, Oxman AD, Chalmers I. So what’s so special about randomisation? In: Maynard A, Chalmers I, editors. Non-Random Reflections on Health Care Research: On the 25th Anniversary Of Archie Cochrane's Effectiveness and Efficiency. London: BMJ Publishing Group; 1997. p. 93-106.

Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 2016;355:i4919. https://doi.org/10.1136/bmj.i4919

2.1b Consider whether the people being compared were cared for similarly.

Explanation

If people in one treatment comparison group receive additional treatments or more care and attention ("co-intervention") than people in the other comparison group, differences in outcomes may reflect those differences rather than the effects of the treatments being compared. For example, in a randomized trial of cognitive behavioural therapy (CBT) for hypochondriasis (persistent fear or belief that one has a serious, undiagnosed illness) compared with no cognitive therapy, a detailed letter of advice was sent to the primary care physicians whose patients were allocated to receive CBT *[Barsky 2004 (RS), Thomson 2007 (SR)]*. Thus, it was not possible to attribute any differences in outcomes to CBT alone since the letter could have altered how the primary care physicians managed patients allocated to CBT. In addition, patients in the CBT group received more attention than those who did not receive CBT. So, it is uncertain how much of the observed difference in outcomes was due to non-specific attention, support, concern, and positive expectation and not specifically to CBT.

Treatment providers who are aware of the treatment to which people are allocated may treat people differently based on their beliefs about the effectiveness of the treatments that are being compared. Their inclinations for or against the treatment can be transferred to the people receiving care and this could have an impact the outcome of interest. One way of preventing co-intervention is to keep treatment providers and patients unaware of ("blind" to) which people have been allocated to which treatment. However, this is not always possible. For example, a randomized comparison of acupuncture to relieve symptoms of irritable bowel syndrome compared three groups prior to administering genuine acupuncture to two of the groups *[Kaptchuk 2008 (RS)]*. Two groups received sham acupuncture. This blinded the recipients of care, but not the providers. To assess the impact of the providers' attitudes about the treatment, in one group, the providers were instructed to interact minimally with the patients, explaining that it was "a scientific study" for which they had been "instructed not to converse with patients". In the other group, they communicated with the patients in a warm, friendly manner, actively listened, showed empathy, and communicated confidence and positive expectation. The third group was put on a waiting list. The proportion of patients reporting adequate relief was 28% in the waiting list group, 44% in the sham acupuncture + minimal interaction group, and 62% in the sham acupuncture + positive communication group.

Basis for this concept

People who can potentially be "blinded" include the people receiving the treatments being compared, the people delivering the treatments, data collectors, people who assess the outcomes, data analysts, the data safety and monitoring committee, and manuscript writers. Because "double blinding" has multiple definitions and is interpreted in different ways *[Devereaux 2001 (RS), Schulz 2002]*, it is best to consider specifically who was blinded and how that could lead to overestimation or underestimation of treatment effects. A systematic review that compared effects in blinded and non-blinded studies in 142 meta-analyses *[Moustgaard 2020 (SR)]* categorised the comparisons and the potential for "performance bias" (the risk of co-intervention and placebo effects) and measurement bias, based on who was not blinded and the type of outcome. The five categories and corresponding potential biases were:

	Who was not blinded	Types of outcomes	Potential biases
1	Recipients of care	Patient reported	Measurement and performance
2	Recipients of care	Assessed by blinded observers	Performance
3	Providers of care	Assessed by providers	Measurement and performance
4	Providers	Assessed by blinded patients or observers	Performance
5	Outcome assessors	Outcomes requiring judgement (“subjective”)	Measurement

The review found similar average effects for the two comparisons where there was a risk of performance bias and not measurement bias (comparison 2 and 4). This could reflect limitations of these comparisons, including possible confounding by other characteristics of the trials included in each meta-analysis, a small number of meta-analyses included in each comparison (14 and 13 respectively) and wide confidence intervals. It also is likely that a lack of blinding is sometimes associated with similar estimates, sometimes with overestimates of effects, and sometimes with underestimates of effects.

Another systematic review combined the data from seven studies that investigated the influence of blinding and other characteristics of randomized trials on treatment effect estimates *[Savović 2012b (SR)]*. It included 234 meta-analyses containing 1973 randomized trials. It found that, on average, lack of or unclear “double-blinding“ (compared to double-blinding) was associated with average treatment effects that were 13% larger – despite differences in definitions of double-blinding. Exaggerated estimates of treatment effects were found primarily for subjective outcomes and not for objective outcomes. The extent to which that was due to measurement bias rather than performance bias is uncertain. Two other reviews have also found that, on average, treatment effects appeared to be exaggerated in randomized studies with lack of unclear implementation of double-blinding *[Martin 2021 (SR), Page 2016a (SR)]*, while other reviews have had inconclusive results *[Dechartres 2016 (SR), Wang 2021 (SR)]*. One other systematic review found that for subjective outcomes, effect estimates appeared to be exaggerated in trials with lack of or unclear blinding of participants (versus blinding of participants), but not for mortality *[Page 2016a (SR)]*. In contrast to that review, a systematic review of the association between lack of blinding and mortality results in critical care found slightly larger effect estimates in nonblinded trials *[Martin 2021 (SR)]*. A possible explanation for this finding is that physicians’ beliefs in a favourable effect of new treatments might influence the timing of their decisions about end-of-life versus life-support practices. All these reviews included comparisons between studies and have a high risk of confounding by other characteristics of the trials included in each meta-analysis.

Within-trial comparisons are at low risk of confounding, when participants are randomized to be blinded or not to be blinded. A systematic review of randomized trials that included sub-studies that randomly allocated patients to be blinded or not blinded included 12 trials in its main analysis *[Hróbjartsson 2014a (SR)]*. It found that, on average, not blinding patients led to moderately exaggerated effect estimates in randomized trials of complementary and alternative treatments with patient-reported outcomes. It is uncertain to what extent this was due to measurement bias rather than performance bias. There are, however, other studies, like the acupuncture example in the explanation above, that indicate that attention from care providers and their attitudes can sometimes influence outcomes (e.g., *[Guyatt 1984 (RS), Kaptchuk 2008 (RS), Thomas 1987 (RS)]*). So, if care providers are not blinded, their attitudes for or against a treatment can impact the outcome of interest.

It is not always possible to blind providers and recipients of care in randomized trials, and it is rarely possible in non-randomized studies such as cohort studies or case-control studies. However, it is possible to blind participants in, for example, comparisons of surgical and technical treatments, treatments that involve attention, devices, and physical therapy *[Armijo-Olivo 2017 (SR), Monaghan 2021 , Wartolowska 2014 (SR)]*, as well as in drug trials. When blinding is not possible, it is important to consider the possibility that there were differences in the treatments received in the treatment comparison groups besides the treatments being compared.

Implications

Be cautious about relying on the results of treatment comparisons if people in the groups that are being compared were not cared for similarly (apart from the treatments being compared). The results of such comparisons can be misleading.

References

Systematic reviews

Armijo-Olivo S, Fuentes J, da Costa BR, Saltaji H, Ha C, Cummings GG. Blinding in Physical Therapy Trials and Its Association with Treatment Effects: A Meta-epidemiological Study. Am J Phys Med Rehabil. 2017;96(1):34-44. https://doi.org/10.1097/phm.0000000000000521

Dechartres A, Trinquart L, Faber T, Ravaud P. Empirical evaluation of which trial characteristics are associated with treatment effect estimates. J Clin Epidemiol. 2016;77:24-37. https://doi.org/10.1016/j.jclinepi.2016.04.005

Hróbjartsson A, Emanuelsson F, Skou Thomsen AS, Hilden J, Brorson S. Bias due to lack of patient blinding in clinical trials. A systematic review of trials randomizing patients to blind and nonblind sub-studies. Int J Epidemiol. 2014a;43(4):1272-83. https://doi.org/10.1093/ije/dyu115

Martin GL, Trioux T, Gaudry S, Tubach F, Hajage D, Dechartres A. Association Between Lack of Blinding and Mortality Results in Critical Care Randomized Controlled Trials: A Meta-Epidemiological Study. Crit Care Med. 2021;49(10):1800-11. https://doi.org/10.1097/ccm.0000000000005065

Moustgaard H, Clayton GL, Jones HE, Boutron I, Jørgensen L, Laursen DRT, et al. Impact of blinding on estimated treatment effects in randomised clinical trials: meta-epidemiological study. BMJ. 2020;368:l6802. https://doi.org/10.1136/bmj.l6802

Page MJ, Higgins JP, Clayton G, Sterne JA, Hróbjartsson A, Savović J. Empirical evidence of study design biases in randomized trials: systematic review of meta-epidemiological studies. PLoS One. 2016a;11(7):e0159267. https://doi.org/10.1371/journal.pone.0159267

Savović J, Jones HE, Altman DG, Harris RJ, Jüni P, Pildal J, et al. Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. Ann Intern Med. 2012b;157(6):429-38. https://doi.org/10.7326/0003-4819-157-6-201209180-00537

Thomson AB, Page LA. Psychotherapies for hypochondriasis. Cochrane Database Syst Rev. 2007;2007(4):Cd006520. https://doi.org/10.1002/14651858.cd006520.pub2

Wang Z, Alahdab F, Farah M, Seisa M, Firwana M, Rajjoub R, et al. Association of study design features and treatment effects in trials of chronic medical conditions: a meta-epidemiological study. BMJ Evid Based Med. 2021. https://doi.org/10.1136/bmjebm-2021-111667

Wartolowska K, Judge A, Hopewell S, Collins GS, Dean BJ, Rombach I, et al. Use of placebo controls in the evaluation of surgery: systematic review. BMJ. 2014;348:g3253. https://doi.org/10.1136/bmj.g3253

Research studies

Barsky AJ, Ahern DK. Cognitive behavior therapy for hypochondriasis: a randomized controlled trial. JAMA. 2004;291(12):1464-70. https://doi.org/10.1001/jama.291.12.1464

Devereaux PJ, Manns BJ, Ghali WA, Quan H, Lacchetti C, Montori VM, et al. Physician interpretations and textbook definitions of blinding terminology in randomized controlled trials. JAMA. 2001;285(15):2000-3. https://doi.org/10.1001/jama.285.15.2000

Guyatt GH, Pugsley SO, Sullivan MJ, Thompson PJ, Berman L, Jones NL, et al. Effect of encouragement on walking test performance. Thorax. 1984;39(11):818-22. https://doi.org/10.1136/thx.39.11.818

Kaptchuk TJ, Kelley JM, Conboy LA, Davis RB, Kerr CE, Jacobson EE, et al. Components of placebo effect: randomised controlled trial in patients with irritable bowel syndrome. BMJ. 2008;336(7651):999-1003. https://doi.org/10.1136/bmj.39524.439618.25

Thomas KB. General practice consultations: is there any point in being positive? BMJ. 1987;294(6581):1200-2. https://doi.org/10.1136/bmj.294.6581.1200

Other references

Monaghan TF, Agudelo CW, Rahman SN, Wein AJ, Lazar JM, Everaert K, et al. Blinding in Clinical Trials: Seeing the Big Picture. Medicina (Kaunas). 2021;57(7). https://doi.org/10.3390/medicina57070647

Schulz KF, Chalmers I, Altman DG. The landscape and lexicon of blinding in randomized trials. Ann Intern Med. 2002;136(3):254-9. https://doi.org/10.7326/0003-4819-136-3-200202050-00022

2.1c Consider whether the people being compared knew which treatments they received.

Explanation

People in a treatment group may behave differently or experience improvements or deterioration because they know the treatment to which they have been assigned. If this phenomenon is associated with an improvement in their symptoms it is known as a placebo effect; if it is associated with a harmful effect it is known as a nocebo effect. If individuals know that they are receiving a treatment that they believe is either better or worse than an alternative (that is, they are not "blinded"), some or all the apparent effects of treatments may be due either to placebo or nocebo effects. For example, a systematic review found 10 randomized trials of acupuncture that included both a "no acupuncture" group and a "sham acupuncture" (placebo) group *[Hróbjartsson 2014a (SR)]*. The non-blinded comparison (of acupuncture compared to no acupuncture) resulted in an overestimate of the effect of acupuncture compared to the blinded comparison (of acupuncture compared to sham acupuncture).

Patients who are aware of the treatment to which they are allocated may also seek additional care or behave differently based on which treatment they receive and their prior beliefs about the effectiveness of the treatment. If they believe a treatment is effective and they are allocated to "no treatment", they may decide to use the treatment anyway (resulting in "contamination"), to use some other treatment, or to withdraw from the study (resulting in "attrition bias"). For example, in a randomized trial, a new type of counselling to help people lose weight was compared to "usual care". People allocated to the counselling were satisfied with their allocation, whereas those allocated to usual care were disappointed *[McCambridge 2014 (RS)]*. Their disappointment may have led some participants to "take control" and change their diet or to seek support elsewhere. This could have resulted in underestimating the effect of the counselling compared to usual care.

Basis for this concept

A systematic review of placebo treatments found 202 studies that randomized participants to a placebo or "no treatment" *[Hróbjartsson 2010 (SR)]*. On average, across 44 studies that reported yes/no (dichotomous) outcomes there was a small effect, but the effect varied. There was also a small effect on average across studies with patient-reported outcomes, with variation in the effect. In trials that reported pain as an outcome, the effect was very variable. Larger effects were associated with physical placebos, such as sham acupuncture. This is consistent with a systematic review of physiotherapy for pain, which found that trials that compared physiotherapy to a sham treatment had smaller effects than trials that compared physiotherapy to "no treatment" *[Ginnerup-Nielsen 2016 (SR)]*. Trials with sham surgery have demonstrated that the act of performing surgery can have a large placebo effect *[Sihvonen 2013 (RS)]*.

In the Hróbjartsson review *[Hróbjartsson 2010 (SR)]*, trials with the explicit purpose of studying placebo effects and trials that did not inform patients about the possible placebo treatment also had larger effects. Randomized trials that have evaluated the effects of a placebo with and without the care provider being positive support the finding that what is communicated about a placebo (or a treatment) has an impact on the placebo effect. One trial found a placebo effect with sham acupuncture with minimal communication (compared to being on a waiting list) and an even larger effect with sham acupuncture with positive communication *[Kaptchuk 2008 (RS)]*. The other trial did not find an effect of a placebo tablet (compared to no placebo) in patients with a variety of different symptoms without a diagnosis, but did find an effect of the general practitioner being positive (with or without the placebo) compared to not being positive *[Thomas 1987 (RS)]*. These studies suggest that the placebo effect depends on patient expectations and that those expectations are influenced by what is communicated to the patient.

Several systematic reviews have investigated the influence of blinding and other characteristics of randomized trials on effect estimates, as described in the basis for Concept 2.1b. Some have found that, on average, studies with inadequate blinding have larger effect estimates than studies with adequate blinding, primarily for subjective outcomes *[Page 2016a (SR), Savović 2012b (SR)]*, whereas others have had inconclusive results *[Dechartres 2016 (SR), Moustgaard 2020 (SR), Wang 2021 (SR)]*. The extent to which overestimation of effect sizes in randomized trials with inadequate blinding is due to measurement bias or co-intervention rather than a placebo effect is uncertain. Moreover, these reviews are based on comparisons between studies and have a high risk of confounding by other characteristics of the trials included in each meta-analysis. Consequently, the extent to which comparisons of treatments with inadequate or no blinding of patients are misleading because of placebo effects is uncertain. It likely varies and is difficult to predict. The risk of being misled is probably greater for patient-reported outcomes, such as pain, and likely depends on the patients' beliefs about the treatments being compared and what they are told.

It is not always possible to blind the people who receive the treatments in randomized trials, and it is rarely possible in non-randomized studies such as cohort studies or case-control studies. When information about whether study participants' exposure to a treatment is collected retrospectively, apparent treatment effects may be either overestimates or underestimates of an effect or association, because of "recall bias". For example, a claim that the measles, mumps, and rubella vaccine caused autism in children received a great deal of attention following publication of a fraudulent study. After this, parents of autistic children tended to recall the start of autism as being soon after the child was vaccinated more often than parents of similar children who were diagnosed prior to publication of that study *[Andrews 2002 (RS)]*.

It is possible to blind participants in randomized trials for many different types of treatments, not just in drug trials. For example, participants can be blinded in comparisons of surgical and technical treatments, treatments that involve attention, devices, and physical therapy *[Armijo-Olivo 2017 (SR), Monaghan 2021 , Wartolowska 2014 (SR)]*. When blinding is not possible, it is important to consider the possibility of placebo and nocebo effects, especially for patient-reported outcomes.

Implications

Be cautious about relying on the results of treatment comparisons if the participants knew which treatment they had received. This may have affected their expectations or behaviour. The results of such comparisons can be misleading.

References

Systematic reviews

Armijo-Olivo S, Fuentes J, da Costa BR, Saltaji H, Ha C, Cummings GG. Blinding in Physical Therapy Trials and Its Association with Treatment Effects: A Meta-epidemiological Study. Am J Phys Med Rehabil. 2017;96(1):34-44. https://doi.org/10.1097/phm.0000000000000521

Dechartres A, Trinquart L, Faber T, Ravaud P. Empirical evaluation of which trial characteristics are associated with treatment effect estimates. J Clin Epidemiol. 2016;77:24-37. https://doi.org/10.1016/j.jclinepi.2016.04.005

Ginnerup-Nielsen E, Christensen R, Thorborg K, Tarp S, Henriksen M. Physiotherapy for pain: a meta-epidemiological study of randomised trials. Br J Sports Med. 2016;50(16):965-71. https://doi.org/10.1136/bjsports-2015-095741

Hróbjartsson A, Emanuelsson F, Skou Thomsen AS, Hilden J, Brorson S. Bias due to lack of patient blinding in clinical trials. A systematic review of trials randomizing patients to blind and nonblind sub-studies. Int J Epidemiol. 2014a;43(4):1272-83. https://doi.org/10.1093/ije/dyu115

Hróbjartsson A, Gøtzsche PC. Placebo interventions for all clinical conditions. Cochrane Database Syst Rev. 2010;2010(1):Cd003974. https://doi.org/10.1002/14651858.cd003974.pub3

Moustgaard H, Clayton GL, Jones HE, Boutron I, Jørgensen L, Laursen DRT, et al. Impact of blinding on estimated treatment effects in randomised clinical trials: meta-epidemiological study. BMJ. 2020;368:l6802. https://doi.org/10.1136/bmj.l6802

Page MJ, Higgins JP, Clayton G, Sterne JA, Hróbjartsson A, Savović J. Empirical evidence of study design biases in randomized trials: systematic review of meta-epidemiological studies. PLoS One. 2016a;11(7):e0159267. https://doi.org/10.1371/journal.pone.0159267

Savović J, Jones HE, Altman DG, Harris RJ, Jüni P, Pildal J, et al. Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. Ann Intern Med. 2012b;157(6):429-38. https://doi.org/10.7326/0003-4819-157-6-201209180-00537

Wang Z, Alahdab F, Farah M, Seisa M, Firwana M, Rajjoub R, et al. Association of study design features and treatment effects in trials of chronic medical conditions: a meta-epidemiological study. BMJ Evid Based Med. 2021. https://doi.org/10.1136/bmjebm-2021-111667

Wartolowska K, Judge A, Hopewell S, Collins GS, Dean BJ, Rombach I, et al. Use of placebo controls in the evaluation of surgery: systematic review. BMJ. 2014;348:g3253. https://doi.org/10.1136/bmj.g3253

Research studies

Andrews N, Miller E, Taylor B, Lingam R, Simmons A, Stowe J, et al. Recall bias, MMR, and autism. Arch Dis Child. 2002;87(6):493-4. https://doi.org/10.1136/adc.87.6.493

Kaptchuk TJ, Kelley JM, Conboy LA, Davis RB, Kerr CE, Jacobson EE, et al. Components of placebo effect: randomised controlled trial in patients with irritable bowel syndrome. BMJ. 2008;336(7651):999-1003. https://doi.org/10.1136/bmj.39524.439618.25

McCambridge J, Sorhaindo A, Quirk A, Nanchahal K. Patient preferences and performance bias in a weight loss trial with a usual care arm. Patient Educ Couns. 2014;95(2):243-7. https://doi.org/10.1016/j.pec.2014.01.003

Sihvonen R, Paavola M, Malmivaara A, Itälä A, Joukainen A, Nurmi H, et al. Arthroscopic partial meniscectomy versus sham surgery for a degenerative meniscal tear. N Engl J Med. 2013;369(26):2515-24. https://doi.org/10.1056/nejmoa1305189

Thomas KB. General practice consultations: is there any point in being positive? BMJ. 1987;294(6581):1200-2. https://doi.org/10.1136/bmj.294.6581.1200

Other references

Monaghan TF, Agudelo CW, Rahman SN, Wein AJ, Lazar JM, Everaert K, et al. Blinding in Clinical Trials: Seeing the Big Picture. Medicina (Kaunas). 2021;57(7). https://doi.org/10.3390/medicina57070647

2.1d Consider whether outcomes were assessed similarly in the people being compared.

Explanation

If a possible treatment outcome is assessed differently in two treatment comparison groups, differences in that outcome may be due to *how* the outcome was assessed rather than *because* of the treatments received by people in each group. For example, if outcome assessors believe that a particular treatment works and they know which patients have received that treatment, they may be more likely to record better outcomes in those who have received the treatment. One way of preventing this is to keep outcome assessors unaware of ("blind" to) which people have been allocated to which treatment.

For example, a randomized trial compared laser surgery to medical treatment for patients with angina (chest pain caused by reduced blood flow to the heart) *[Oesterle 2000 (RS)]*. The severity of angina after one year was assessed by the investigators who were aware of treatment assignment (i.e., unblinded) and by trained interviewers who were not aware (blinded). Comparison of the non-blinded investigators' assessments to the blinded interviewers' assessments showed that the investigators assessed the angina as being less severe much more often in the laser surgery group than in the medical treatment group. Twenty-eight percent of the apparent angina improvement could be attributed to bias.

Systematic differences in outcome assessment ("measurement bias") can make treatment effects appear either larger or smaller than they actually are. Blinding is less important for "objective" outcomes, like death, than for "subjective" outcomes, like pain.

Basis for this concept

Comparisons for blinded and non-blinded outcome assessment within randomized trials, like the above example, have been summarised in three systematic reviews *[Hróbjartsson 2012 (SR), Hróbjartsson 2013 (SR), Hróbjartsson 2014b (SR)]*. For yes/no (dichotomous) outcomes, treatment effects were, on average, larger when assessed non-blinded compared to blinded assessments *[Hróbjartsson 2012 (SR)]*. For outcomes that were assessed using a measurement scale, treatment effects were, on average, also larger when assessed non-blinded compared to blinded assessments *[Hróbjartsson 2013 (SR)]*. The same was found for time-to-event outcomes *[Hróbjartsson 2014b (SR)]*. But in some situations, treatment effects were smaller when assessed nonblinded compared to blinded. A likely explanation for this is that the new treatment being evaluated was more practical and less expensive than the established treatment, and the investigators' expectation was that it might not be as beneficial.

Several systematic reviews have investigated the influence of blinding and other characteristics of randomized trials on effect estimates, as described in the basis for Concept 2.1b. Some have found that, on average, studies with inadequate blinding of outcome assessors or inadequate "double blinding" have larger effect estimates than studies with adequate blinding, primarily for subjective outcomes *[Page 2016a (SR), Savović 2012b (SR)]*. Others have had inconclusive results *[Dechartres 2016 (SR), Moustgaard 2020 (SR), Wang 2021 (SR)]*. Because these reviews are based on comparisons between studies, they have a high risk of confounding by other characteristics of the trials included in each meta-analysis. So, the reviews of comparisons within randomized trials provide a more reliable basis for this concept.

Although it is not always possible to blind participants in randomized trials, it generally is possible to blind outcome assessors. However, for some outcome measures, such as patient-reported outcomes, this is not possible if the patients participating in a trial cannot be blinded. It is also sometimes possible to blind outcome assessors in non-randomized studies. When blinding is not possible, it is important to consider the possibility of measurement bias.

Implications

Be cautious about relying on the results of treatment comparisons if outcomes were not assessed in the same way in the different treatment comparison groups. The results of such comparisons can be misleading.

References

Systematic reviews

Dechartres A, Trinquart L, Faber T, Ravaud P. Empirical evaluation of which trial characteristics are associated with treatment effect estimates. J Clin Epidemiol. 2016;77:24-37. https://doi.org/10.1016/j.jclinepi.2016.04.005

Hróbjartsson A, Thomsen AS, Emanuelsson F, Tendal B, Hilden J, Boutron I, et al. Observer bias in randomised clinical trials with binary outcomes: systematic review of trials with both blinded and non-blinded outcome assessors. BMJ. 2012;344:e1119. https://doi.org/10.1136/bmj.e1119

Hróbjartsson A, Thomsen AS, Emanuelsson F, Tendal B, Hilden J, Boutron I, et al. Observer bias in randomized clinical trials with measurement scale outcomes: a systematic review of trials with both blinded and nonblinded assessors. CMAJ. 2013;185(4):E201-11. https://doi.org/10.1503/cmaj.120744

Hróbjartsson A, Thomsen AS, Emanuelsson F, Tendal B, Rasmussen JV, Hilden J, et al. Observer bias in randomized clinical trials with time-to-event outcomes: systematic review of trials with both blinded and non-blinded outcome assessors. Int J Epidemiol. 2014b;43(3):937-48. https://doi.org/10.1093/ije/dyt270

Moustgaard H, Clayton GL, Jones HE, Boutron I, Jørgensen L, Laursen DRT, et al. Impact of blinding on estimated treatment effects in randomised clinical trials: meta-epidemiological study. BMJ. 2020;368:l6802. https://doi.org/10.1136/bmj.l6802

Page MJ, Higgins JP, Clayton G, Sterne JA, Hróbjartsson A, Savović J. Empirical evidence of study design biases in randomized trials: systematic review of meta-epidemiological studies. PLoS One. 2016a;11(7):e0159267. https://doi.org/10.1371/journal.pone.0159267

Savović J, Jones HE, Altman DG, Harris RJ, Jüni P, Pildal J, et al. Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. Ann Intern Med. 2012b;157(6):429-38. https://doi.org/10.7326/0003-4819-157-6-201209180-00537

Wang Z, Alahdab F, Farah M, Seisa M, Firwana M, Rajjoub R, et al. Association of study design features and treatment effects in trials of chronic medical conditions: a meta-epidemiological study. BMJ Evid Based Med. 2021. https://doi.org/10.1136/bmjebm-2021-111667

Research studies

Oesterle SN, Sanborn TA, Ali N, Resar J, Ramee SR, Heuser R, et al. Percutaneous transmyocardial laser revascularisation for severe angina: the PACIFIC randomised trial. Potential Class Improvement From Intramyocardial Channels. Lancet. 2000;356(9243):1705-10. https://doi.org/10.1016/s0140-6736(00)03203-7

2.1e Consider whether outcomes were assessed reliably.

Explanation

Some outcomes are easy to assess, such as births and deaths. Others are more difficult, such as depression or quality of life. For treatment comparisons to be meaningful, outcomes that are meaningful to people should be assessed using methods that have been shown to be reliable.

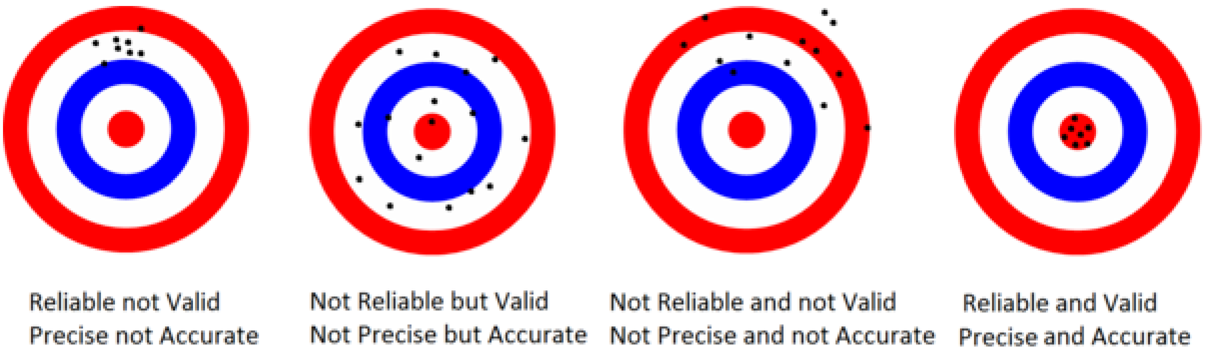
Unreliable outcome measures result in outcome misclassification or measurement error. When misclassification is similar in the groups of people being compared ("non-differential"), this tends to lead to underestimation of effects. For example, a vaccine cannot be expected to protect against infections other than those for which it was developed. So, for example, influenza vaccines are less effective for preventing 'influenza-like' illness (much of which is not caused by influenza viruses) than for preventing influenza that is confirmed by a laboratory test *[Demicheli 2018 (SR)]*. As the proportion of influenza-like illnesses that are caused by influenza viruses decreases, the difference will increase between the effects of vaccines on influenza-like illness and laboratory confirmed influenza.

Basis for this concept

Non-differential (unbiased) misclassification of outcomes has been shown to lead to underestimation of treatment effects in simulations and research studies *[Blackwelder 1991 (RS), Copeland 1977 , De Smedt 2018 , Hunnicutt 2016 (SR), Petersen 2021 (SR), Rogers 2006 (RS), Walraven 2018 (RS)]*. Underestimation of effects increases as the accuracy of the assessment decreases and as the occurrence of the outcome decreases.

Non-randomized studies of the association between treatments (and other factors) and outcomes mention the concept of measurement error in about half of the studies published in top-ranked journals *[Brakenhoff 2018 (SR)]*. An additional problem in non-randomized studies is error in measuring exposure to the treatments being studied and to confounders. Unlike treatment outcome measurement error, exposure measurement error does not always lead to underestimation of effects. Few studies investigate the impact of measurement error, so it is difficult to judge the robustness of the reported effect estimates.

A target is often used to explain the difference between precision or the extent of random errors (sometimes referred to as reliability) and "validity" or systematic errors (sometimes referred to as accuracy) in outcome measures or diagnostic tests (Figure).



For self-reported outcomes, systematic errors can be caused by social desirability bias *[Althubaiti 2016]*. Self-reporting of an outcome (or a treatment, in non-randomized studies) can be influenced by social desirability or approval, especially when anonymity and confidentiality cannot be guaranteed. For example, self-reporting of behaviours such as diet, smoking, sexual behaviours, drug use, or compliance with a prescribed treatment can be influenced by the study participants' perceptions of what the investigators or others view as good or bad behaviour. This can result in over-reporting of “good behaviours” and underreporting of “bad behaviours”.

Measuring outcomes that are important to patients often depends on patient-reported outcomes *[Calvert 2013 , Garratt 2002 (SR), Johnston 2021]*. When patient-reported quality of life is reported in randomized trials, the reported effects on quality of life sometimes are not in agreement with the primary outcome measures *[Contopoulos-Ioannidis 2009 (SR)]*.

However, for estimates of the effects of treatments on patient-reported outcomes to be reliable, the patient-reported outcome measures used must be reliable and valid *[Gagnier 2021]*. Outcomes that are measured using an outcome measure that has not been shown to be reliable and valid can result in misleading effect estimates. For example, randomized trials of treatments for schizophrenia that used unpublished outcome measures were more likely to report that a treatment was superior to the comparison (control) treatment compared to trials that had used a published (evaluated) outcome measure *[Marshall 2000 (SR)]*.

Patient-reported outcome measures do not always reflect what is meaningful and important to patients. It is important to ensure that patients understand them and that they capture what is important to them. For example, the McGill Pain Questionnaire is widely used in randomized trials, however it was developed for clinician reporting and never underwent qualitative evaluation with direct patient input. Interviews with patients likely would have revealed difficulties understanding the options for responding to the question “How strong is your pain?” than the ones that are used (1 Mild, 2 Discomforting, 3 Distressing, 4 Horrible, 5 Excruciating) *[Basch 2011]*.

It is also important that patient-reported outcome measures are comprehensive include all aspects of an outcome that are important to patients and relevant (do not include aspects that are unimportant or irrelevant). For example, a systematic review of patient-reported outcome measures for postpartum recovery included 15 outcome measures *[Sultan 2021 (SR)]*. The obstetric-specific outcome measures included between four and 12 aspects (“domains”) of outpatient postpartum recovery. They were all missing at least one domain, such as pain, psychosocial distress, sleep, motherhood experience, fatigue, or sexual function. On the other hand, some outcome measures include domains that are unlikely to be relevant in some settings, such as “satisfaction with pollution” and “satisfaction with transportation”.

The number of patient-reported outcome measures is growing rapidly, and there are now well over 1,000 systematic reviews of those measures *[COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN) 2019]*. Variation in the outcome measures that are reported in different comparisons of the same treatments makes it difficult to synthesise and interpret the results in systematic reviews. A core outcome set is a standardised set of outcomes, agreed by stakeholders, including patients that should be the minimum outcomes measured and reported in all trials in particular health areas *[Matvienko-Sikar 2021 (SR)]*. Core outcome sets are often not used or reported in randomized trials. Greater use of core outcome sets could improve evaluations of treatment effects and systematic reviews. Use of core outcome sets could also reduce the risk of selective outcome reporting (see Concept 2.2b), enhance research transparency, and help to ensure that important outcomes are assessed using reliable outcome measures.

Implications

Be cautious about relying on the results of treatment comparisons if outcomes have not been assessed using methods that have been shown to be reliable.

References

Systematic reviews

Brakenhoff TB, Mitroiu M, Keogh RH, Moons KGM, Groenwold RHH, van Smeden M. Measurement error is often neglected in medical literature: a systematic review. J Clin Epidemiol. 2018;98:89-97. https://doi.org/10.1016/j.jclinepi.2018.02.023

Contopoulos-Ioannidis DG, Karvouni A, Kouri I, Ioannidis JP. Reporting and interpretation of SF-36 outcomes in randomised trials: systematic review. BMJ. 2009;338:a3006. https://doi.org/10.1136/bmj.a3006

Demicheli V, Jefferson T, Ferroni E, Rivetti A, Di Pietrantonj C. Vaccines for preventing influenza in healthy adults. Cochrane Database Syst Rev. 2018;2(2):Cd001269. https://doi.org/10.1002/14651858.cd001269.pub6

Garratt A, Schmidt L, Mackintosh A, Fitzpatrick R. Quality of life measurement: bibliographic study of patient assessed health outcome measures. BMJ. 2002;324(7351):1417. https://doi.org/10.1136/bmj.324.7351.1417

Hunnicutt JN, Ulbricht CM, Chrysanthopoulou SA, Lapane KL. Probabilistic bias analysis in pharmacoepidemiology and comparative effectiveness research: a systematic review. Pharmacoepidemiol Drug Saf. 2016;25(12):1343-53. https://doi.org/10.1002/pds.4076

Marshall M, Lockwood A, Bradley C, Adams C, Joy C, Fenton M. Unpublished rating scales: a major source of bias in randomised controlled trials of treatments for schizophrenia. Br J Psychiatry. 2000;176:249-52. https://doi.org/10.1192/bjp.176.3.249

Matvienko-Sikar KL, Fellow HRBR, Avery KSL, Blazeby JP, Devane DP, Dodd SL, et al. Use ofcore outcome sets was low in clinical trials published in major medical journals. J Clin Epidemiol. 2021. https://doi.org/10.1016/j.jclinepi.2021.10.012

Petersen JM, Ranker LR, Barnard-Mayers R, MacLehose RF, Fox MP. A systematic review of quantitative bias analysis applied to epidemiological research. Int J Epidemiol. 2021;50(5):1708-30. https://doi.org/10.1093/ije/dyab061

Sultan P, Sharawi N, Blake L, Ando K, Sultan E, Aghaeepour N, et al. Use of Patient-Reported Outcome Measures to Assess Outpatient Postpartum Recovery: A Systematic Review. JAMA Netw Open. 2021;4(5):e2111600. https://doi.org/10.1001/jamanetworkopen.2021.11600

Research studies

Blackwelder WC, Storsaeter J, Olin P, Hallander HO. Acellular pertussis vaccines. Efficacy and evaluation of clinical case definitions. Am J Dis Child. 1991;145(11):1285-9. https://doi.org/10.1001/archpedi.1991.02160110077024

Rogers WO, Atuguba F, Oduro AR, Hodgson A, Koram KA. Clinical case definitions and malaria vaccine efficacy. J Infect Dis. 2006;193(3):467-73. Clinical Case Definitions and Malaria Vaccine Efficacy

Walraven CV. A comparison of methods to correct for misclassification bias from administrative database diagnostic codes. Int J Epidemiol. 2018;47(2):605-16. https://doi.org/10.1093/ije/dyx253

Other references

Althubaiti A. Information bias in health research: definition, pitfalls, and adjustment methods. J Multidiscip Healthc. 2016;9:211-7. https://doi.org/10.2147/jmdh.s104807

Basch E, Abernethy AP, Reeve BB. Assuring the patient centeredness of patient-reported outcomes: content validity in medical product development and comparative effectiveness research. Value Health. 2011;14(8):965-6. https://doi-org.ezproxy.uio.no/10.1016/j.jval.2011.10.002

Calvert M, Blazeby J, Altman DG, Revicki DA, Moher D, Brundage MD. Reporting of patient-reported outcomes in randomized trials: the CONSORT PRO extension. JAMA. 2013;309(8):814-22. https://doi.org/10.1001/jama.2013.879

COnsensus-based Standards for the selection of health Measurement INstruments (COSMIN). COSMIN Database of Systematic Reviews, 2019. Accessed: 22 November 2021. https://www.cosmin.nl/tools/database-systematic-reviews/

Copeland KT, Checkoway H, McMichael AJ, Holbrook RH. Bias due to misclassification in the estimation of relative risk. Am J Epidemiol. 1977;105(5):488-95. https://doi.org/10.1093/oxfordjournals.aje.a112408

De Smedt T, Merrall E, Macina D, Perez-Vilar S, Andrews N, Bollaerts K. Bias due to differential and non-differential disease- and exposure misclassification in studies of vaccine effectiveness. PLoS One. 2018;13(6):e0199180. https://doi.org/10.1371/journal.pone.0199180

Gagnier JJ, Lai J, Mokkink LB, Terwee CB. COSMIN reporting guideline for studies on measurement properties of patient-reported outcome measures. Qual Life Res. 2021;30(8):2197-218. https://doi.org/10.1007/s11136-021-02822-4

Johnston B, Patrick D, Devji T, Maxwell L, Bingham III C, Beaton D, et al. Patient-reported outcomes. Cochrane Handbook for Systematic Reviews of Interventions version 6,2. 2021. https://training.cochrane.org/handbook/current/chapter-18

2.1f Consider whether outcomes were assessed in all (or nearly all) the people being compared.

Explanation

People in treatment comparisons who are not followed up to the end of the study may have worse outcomes than those who completed follow-up. For example, they may have dropped out because the treatment was not working or because of side effects. If those people are excluded from the comparison, the findings of the study may be misleading.

For example, in a randomized trial of hip protectors for preventing hip fracture, about 20% of participants were lost to follow-up *[Dumville 2006 (RS)]*. The authors dealt with this problem for the main outcome (hip fracture) by accessing the general practice records of patients who were lost to follow-up. However, for other outcomes, such as quality of life, the necessary information had not been recorded, so this was not possible. Therefore, effect estimates for those outcomes could be misleading. Slightly more participants were lost to follow-up in the group assigned to use hip protectors than in the group assigned not to use hip protectors (28% versus 22%). This difference increased the likelihood that participants in the comparison groups were no longer similar, even though they were similar at the start of the trial, as would be expected with random allocation (see Concept 2.1a). By looking at the baseline characteristics of study participants, one can see, for example, that more volunteers, people with poor or fair health, and people with a previous fracture had been lost from the control group than had been lost from the intervention group. It is possible to adjust for those variables in statistical analyses of the results. However, because differences in attrition are difficult to predict, such analyses are rarely planned. Moreover, adjustment can only be made for variables (potential confounders) that have been measured at baseline. Thus, the apparent effect of hip protectors on quality of life is far less certain than the effect on hip fractures.

Basis for this concept

Loss to follow-up in randomized trials can make the results misleading if the unavailability of data is associated with the likelihood of outcome events. Substantial loss to follow-up can lead to overestimates or underestimates of treatment effects. A systematic review of randomized trials published in the top five medical journals found that plausible assumptions regarding outcomes of patients lost to follow-up could change the interpretation of results of as many as one-third of the included trials *[Akl 2012 (SR)]*.

Several systematic reviews have found that, on average, randomized trials reporting higher levels of attrition (loss to follow-up) were likely to overestimate treatment effects compared to trials with lower levels of attrition *[Armijo-Olivo 2021 (SR), Armijo-Olivo 2020 (SR), Nüesch 2009 (SR)]*. Other systematic reviews have reported underestimation of effects or inconclusive findings about the association between attrition and effect sizes *[Hartling 2014 (SR), Page 2016a (SR), Savović 2012b (SR), Wang 2021 (SR)]*. All these reviews included comparisons between studies and have a high risk of confounding by other characteristics of the studies that were compared. Nonetheless, they are consistent with the logical explanation of how excluding people who were lost to follow-up can be misleading, and it is likely that the direction and magnitude of bias due to attrition varies *[Nüesch 2009 (SR)]*.

Most randomized trials report the number of participants lost to follow-up, but many do not report analyses that take account of loss to follow-up or assess the robustness of analyses that exclude participants who were lost to follow-up *[Barretto Dos Santos Lopes Batista 2019 (SR), Wood 2004 (SR)]*. The best way to prevent "attrition bias" is through efforts to retain participants in studies *[Gillies 2021 (SR)]*.

Missing data from loss to follow-up can be dealt with statistically by various methods including, for example, imputing values based on assumptions about the missing data to give a conservative estimate of the treatment effect. However, the risk of bias still remains when trials do not collect adequate data to yield accurate estimates *[Hollis 1999 (SR)]*, and even small numbers of participants lost to follow-up can have an impact on the results of treatment comparisons *[Walsh 2015]*.

Implications

Be cautious about relying on the results of treatment comparisons if many people were lost to follow-up, or if there was a big difference between the comparison groups in the proportions of people lost to follow-up.

References

Systematic reviews

Akl EA, Briel M, You JJ, Sun X, Johnston BC, Busse JW, et al. Potential impact on estimated treatment effects of information lost to follow-up in randomised controlled trials (LOST-IT): systematic review. BMJ. 2012;344:e2809. https://doi.org/10.1136/bmj.e2809

Armijo-Olivo S, da Costa BR, Ha C, Saltaji H, Cummings GG, Fuentes J. Are biases related to attrition, missing data, and the use of intention to treat related to the magnitude of treatment effects in physical therapy trials? A meta-epidemiological study. Am J Phys Med Rehabil. 2021. https://doi.org/10.1097/phm.0000000000001837

Armijo-Olivo S, Machalicek W, Dennett L, Ballenberger N. Influence of attrition, missing data, compliance, and related biases and analyses strategies on treatment effects in randomized controlled trials in rehabilitation: a methodological review. Eur J Phys Rehabil Med. 2020;56(6):799-816. https://doi.org/10.23736/s1973-9087.20.06428-x

Barretto Dos Santos Lopes Batista K, Thiruvenkatachari B, O'Brien K. Intention-to-treat analysis: Are we managing dropouts and missing data properly in research on orthodontic treatment? A systematic review. Am J Orthod Dentofacial Orthop. 2019;155(1):19-27.e3. https://doi.org/10.1016/j.ajodo.2018.08.013

Gillies K, Kearney A, Keenan C, Treweek S, Hudson J, Brueton VC, et al. Strategies to improve retention in randomised trials. Cochrane Database Syst Rev. 2021;3(3):Mr000032. https://doi.org/10.1002/14651858.mr000032.pub3

Hartling L, Hamm MP, Fernandes RM, Dryden DM, Vandermeer B. Quantifying bias in randomized controlled trials in child health: a meta-epidemiological study. PLoS One. 2014;9(2):e88008. https://doi.org/10.1371/journal.pone.0088008

Hollis S, Campbell F. What is meant by intention to treat analysis? Survey of published randomised controlled trials. BMJ. 1999;319(7211):670-4. https://doi.org/10.1136/bmj.319.7211.670

Nüesch E, Trelle S, Reichenbach S, Rutjes AW, Bürgi E, Scherer M, et al. The effects of excluding patients from the analysis in randomised controlled trials: meta-epidemiological study. BMJ. 2009;339:b3244. https://doi.org/10.1136/bmj.b3244

Page MJ, Higgins JP, Clayton G, Sterne JA, Hróbjartsson A, Savović J. Empirical evidence of study design biases in randomized trials: systematic review of meta-epidemiological studies. PLoS One. 2016a;11(7):e0159267. https://doi.org/10.1371/journal.pone.0159267

Savović J, Jones HE, Altman DG, Harris RJ, Jüni P, Pildal J, et al. Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. Ann Intern Med. 2012b;157(6):429-38. https://doi.org/10.7326/0003-4819-157-6-201209180-00537

Wang Z, Alahdab F, Farah M, Seisa M, Firwana M, Rajjoub R, et al. Association of study design features and treatment effects in trials of chronic medical conditions: a meta-epidemiological study. BMJ Evid Based Med. 2021. https://doi.org/10.1136/bmjebm-2021-111667

Wood AM, White IR, Thompson SG. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. Clin Trials. 2004;1(4):368-76. https://doi.org/10.1191/1740774504cn032oa

Research studies

Dumville JC, Torgerson DJ, Hewitt CE. Reporting attrition in randomised controlled trials. BMJ. 2006;332(7547):969-71. https://doi.org/10.1136/bmj.332.7547.969

Other references

Walsh M, Devereaux PJ, Sackett DL. Clinician trialist rounds: 28. When RCT participants are lost to follow-up. Part 1: Why even a few can matter. Clin Trials. 2015;12(5):537-9. https://doi.org/10.1177/1740774515597702

2.1g Consider whether people’s outcomes were analysed in the group to which they were allocated.

Explanation

Random allocation to treatment comparison groups helps to ensure that people in the comparison groups have similar characteristics before they receive treatment (see Concept 2.1a). However, people sometimes do not receive or take the treatment allocated to them. The characteristics of such people often differ from those who do take the treatments allocated to them. Excluding from the analysis people who did not receive the treatments allocated to them may mean that like is no longer being compared with like. This may lead to an underestimate or an overestimate of treatment differences relative to what would have been the case if everyone had received treatment that had been intended for them.

For example, in a comparison of surgery and drug treatments, people who die while waiting for surgery should be counted in the surgery group, even though they did not receive surgery. This may seem counter-intuitive. But if they are excluded and people who die during the same time in the drug group are not excluded, it will not be a fair comparison.

The New York Health Insurance Plan (HIP) randomized trial of screening for breast cancer provides a striking illustration of how people who comply with a treatment (in this case, screening mammography) may be different from those who do not. The study found similar numbers of deaths after five years among women offered screening and those who were not offered screening (Table) *[Shapiro 1977 (RS)]*. Some women offered screening chose not to be screened. If those women are excluded from the comparison, it appears that there were fewer deaths in the screened group compared to the women who were not offered screening (22 versus 30 per 1,000 women). However, that comparison is misleading because there were important differences between the women offered screening who chose to be screened and those who chose not to be screened. Those differences resulted in almost twice as many deaths among women who chose not to be screened compared to women who chose to be screened (40 versus 22 per 1,000 women).

*Table. Total number of deaths after five years in the HIP randomized trial of breast cancer screening**

Comparison group	Group size	Deaths per 1,000 women
Offered screening	31,000	28
Chose to be screened	20,200	22
Chose not to be screened	10,800	40
Not offered screening	31,000	30

* Data from Table 1 in *[Freedman 2004]*.

Basis for this concept

A systematic review of randomized trials published in the top five medical journals reported the results in three ways *[Mostazir 2021 (SR)]*:

1) Including outcomes in all the study participants allocated to each of the treatment comparison groups (“intention-to-treat” analysis)

   In this analysis, study participants who dropped out of the study, did not adhere to the study treatment to which they were allocated, or even took the wrong study treatment, are included in the treatment comparison group to which they were randomly allocated.

2) Only including outcomes in participants who adhered to the trial protocol, including the treatment to which they were allocated (“per-protocol” analysis).

The aim of this analysis is to answer the question: "What is the effect if participants are fully compliant?" However, because it excludes participants who were not compliant, the treatment comparison groups will no longer be similar if the people who do not comply and are not included in the analysis, as illustrated by the breast cancer screening example above.

3) Using a statistical model to compare people who complied in the "treatment group" to people in the "control group" who would have complied to the study treatment (using the "complier average causal effect" (CACE) method).

On average, the "per-protocol" analyses generated larger estimates of treatment effects than the intention-to-treat analyses. Differences between the two analyses increased with increasing degrees of non-compliance. However, the CACE effect estimates were similar to the intention-to-treat estimates, suggesting that the "per-protocol" analyses likely overestimated the impact of non-compliance on the intention-to-treat effect estimates. In other words, they were biased.

Other systematic reviews have compared the results of randomized trials that included all participants in the analysis (intention-to-treat) to the results of trials comparing the same treatments but after excluding some participants. These reviews have found that, on average, trials that did not report intention-to-treat analyses over-estimated or underestimated treatment effects compared to those that used intention-to-treat analyses *[Abraha 2015 (SR)]* treatment effects *[Armijo-Olivo 2021 (SR)]*. Other reviews were inclusive about the average impact of excluding participants *[Balk 2002 (SR), de Almeida 2019 (SR), Siersma 2007 (RS), van Tulder 2009 (SR)]*, or found that this often resulted in biased estimates of treatment effects, but the extent and direction of bias was unpredictable *[Nüesch 2009 (SR)]*. One systematic review found that many systematic reviews of the effects of treatments include at least one randomized trial that did not report an intention-to-treat analysis, and that those trials were more likely to have "positive" ("statistically significant") findings, industry sponsorship, and authors with conflicts of interest *[Abraha 2017 (SR)]*. All the reviews that compare randomized trials, have a high risk of confounding by other characteristics of the trials. Nonetheless, they support the logical arguments for being cautious about analyses of randomized trials that exclude some participants from the treatment group to which they were allocated.

Implications

Be cautious about relying on the results of treatment comparisons if patients' outcomes have not been counted in the group to which the patients were allocated.

References

Systematic reviews

Abraha I, Cherubini A, Cozzolino F, De Florio R, Luchetta ML, Rimland JM, et al. Deviation from intention to treat analysis in randomised trials and treatment effect estimates: meta-epidemiological study. BMJ. 2015;350:h2445. https://doi.org/10.1136/bmj.h2445

Abraha I, Cozzolino F, Orso M, Marchesi M, Germani A, Lombardo G, et al. A systematic review found that deviations from intention-to-treat are common in randomized trials and systematic reviews. J Clin Epidemiol. 2017;84:37-46. https://doi.org/10.1016/j.jclinepi.2016.11.012

Armijo-Olivo S, da Costa BR, Ha C, Saltaji H, Cummings GG, Fuentes J. Are biases related to attrition, missing data, and the use of intention to treat related to the magnitude of treatment effects in physical therapy trials? A meta-epidemiological study. Am J Phys Med Rehabil. 2021. https://doi.org/10.1097/phm.0000000000001837

Balk EM, Bonis PA, Moskowitz H, Schmid CH, Ioannidis JP, Wang C, et al. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. JAMA. 2002;287(22):2973-82. https://doi.org/10.1001/jama.287.22.2973

de Almeida MO, Saragiotto BT, Maher C, Costa LOP. Allocation Concealment and Intention-To-Treat Analysis Do Not Influence the Treatment Effects of Physical Therapy Interventions in Low Back Pain Trials: a Meta-epidemiologic Study. Arch Phys Med Rehabil. 2019;100(7):1359-66. https://doi.org/10.1016/j.apmr.2018.12.036

Mostazir M, Taylor G, Henley WE, Watkins ER, Taylor RS. Per-Protocol analyses produced larger treatment effect sizes than intention to treat: a meta-epidemiological study. J Clin Epidemiol. 2021;138:12-21. https://doi.org/10.1016/j.jclinepi.2021.06.010

Nüesch E, Trelle S, Reichenbach S, Rutjes AW, Bürgi E, Scherer M, et al. The effects of excluding patients from the analysis in randomised controlled trials: meta-epidemiological study. BMJ. 2009;339:b3244. https://doi.org/10.1136/bmj.b3244

van Tulder MW, Suttorp M, Morton S, Bouter LM, Shekelle P. Empirical evidence of an association between internal validity and effect size in randomized controlled trials of low-back pain. Spine. 2009;34(16):1685-92. https://doi.org/10.1097/brs.0b013e3181ab6a78

Research studies

Shapiro S. Evidence on screening for breast cancer from a randomized trial. Cancer. 1977;39(6 Suppl):2772-82. https://doi.org/10.1002/1097-0142(197706)39:6%3C2772::aid-cncr2820390665%3E3.0.co;2-k

Siersma V, Als-Nielsen B, Chen W, Hilden J, Gluud LL, Gluud C. Multivariable modelling for meta-epidemiological assessment of the association between trial quality and treatment effects estimated in randomized clinical trials. Stat Med. 2007;26(14):2745-58. https://doi.org/10.1002/sim.2752

Other references

Freedman DA, Petitti DB, Robins JM. On the efficacy of screening for breast cancer. Int J Epidemiol. 2004;33(1):43-55. https://doi.org/10.1093/ije/dyg275

2.2 Reviews of the effects of treatments should be fair.

2.2a Consider whether systematic methods were used.

Explanation

A systematic review is a summary of research evidence (studies) which uses systematic and explicit methods to summarise the research on the effects of a treatment (or some other topic). A systematic review addresses a clearly formulated question using a structured approach to identify, select, and critically appraise relevant studies, and to collect and analyse data from the studies that are included in the review. Systematic reviews begin with protocols, which should be registered and searchable in registries such as Prospero *[Booth 2012]*.

Even reviews that purport to be systematic may not be. Reviews that do not use systematic methods may result in biased or imprecise estimates of the effects of treatments because the selection of studies for inclusion may be biased, or the methods may result in some studies not being found. In addition, the appraisal of the quality of some studies may be biased, or the synthesis of the results of the selected studies may be inadequate or inappropriate.

For example, if a systematic review of giving blood thinners to patients with an acute heart attack had been done in the late 1970s, it would have established the effectiveness of that treatment about 10 years before the results a very large randomized trial became available *[Antman 1992 (SR)]*. If those results had been acted upon, thousands of premature deaths could have been avoided. Instead, recommendations were based on unsystematic reviews of the evidence. Similarly, the harmful effects of medicines to reduce heart rhythm abnormalities in patients with an acute heart attack could have been recognised years earlier. And thousands of deaths caused by those medicines could have been prevented if those results had been acted upon.

Basis for this concept

Many reviews of the effects of treatments are unsystematic. For example, a systematic review of reviews of two treatments for arthritis found that 91% of 281 published reviews were unsystematic and did not report methods and conflicts of interest in sufficient detail *[Roundtree 2009 (SR)]*. A "cumulative meta-analysis" starts with the results of the first study evaluating a treatment, typically a randomized trial, and adds other studies one at a time. This shows how the overall effect estimate changes as each new study is added. A systematic review of more than 1,500 cumulative meta-analyses shows that, had researchers systematically assessed what was already known, some beneficial and harmful effects of treatments (such as blood thinners and medicines to reduce heart rhythm abnormalities for acute heart attacks) could have been identified earlier than they were. This could have reduced unnecessary research as well as improving health outcomes *[Clarke 2014 (SR)]*.

A review of reports of 1,523 trials published from 1963 to 2004 found that fewer than 25% of preceding trials were cited *[Robinson 2011 (SR)]*. Other research has shown that authors selectively cite studies based on their results when research is not systematically reviewed *[Duyx 2017 (SR), Leng 2018 (RS), Urlings 2019 (RS), Urlings 2021 (RS)]*. Explicit criteria for deciding which studies to include in a review, adequate searches for studies, and efforts to minimize error in selecting studies can reduce selective inclusion of studies in research reviews. In contrast to selective citation of research in unsystematic reviews, a review of a random sample of systematic reviews did not find evidence of selective inclusion of studies *[Page 2016 (RS)]*.

The starting point for a systematic review is a clearly formulated question. A widely used framework for this is PICO, which stands for Population, Intervention (treatment), Comparator (comparison treatment), and Outcomes *[Cumpston 2020 , Huang 2006 (RS)]*. In addition to being helpful formulating the question, this framework can be helpful for specifying inclusion criteria for studies, designing search strategies for finding relevant research, and extracting and analysing data from included studies.

To avoid missing relevant studies, it is important to conduct an adequate search, particularly using bibliographic databases *[Ewald 2020 (RS), Lefebvre 2021 , Marshall 2019 (RS)]*. People can easily make errors when screening the results of searches to decide which studies to include and extracting data from those studies *[E 2020 (RS), Gartlehner 2020 (RS), Robson 2019 (SR), Waffenschmidt 2019 (SR), Wang 2020 (SR)]*. Having two people screen and select studies for inclusion, extract data from included studies, and assess the risk of bias in included studies can reduce those errors. Increasingly, automation is being used to do this, with the potential to save time and increase accuracy *[Scott 2021 (RS), Tsafnat 2014 (OR)]*. Using statistical or structured methods to synthesise study results can reduce errors such as giving inappropriate weight to studies that support the authors' prior views or being misled by inappropriate analyses such as vote counting (counting the number of "positive" and "negative" studies *[Oxman 1994]*. Once the results have been reliably summarised, it is important to interpret and report the results without misrepresentation of the findings (spin) *[Page 2021 , Rucker 2021 (SR)]*.

There is an endless amount of information on the Internet about treatments. However, most of that information is not based on systematic reviews and there is a lot of misinformation. A review of English language websites intended for patients and the public, which provide information on a broad scope of treatments found two sources that provide information about treatments that is explicitly based on systematic reviews *[Oxman 2019 (SR)]*. Sources such as those are essential, to make it easy for people to find reliable information about the effects of treatments. Although an increasing number of systematic reviews are being published, many are poorly conducted and reported *[Page 2016b (SR), Rosenberger 2021 (SR)]*. The results of reliable systematic reviews can be difficult for most people (including health professionals) to find without user-friendly sources of information about the effects of treatments that is based on reliable, up-to-date systematic reviews.

Implications

Whenever possible, use up-to-date systematic reviews of fair comparisons to inform decisions rather than non-systematic reviews of fair comparisons of treatments.

References

Systematic reviews

Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. Treatments for myocardial infarction. JAMA. 1992;268(2):240-8. https://doi.org/10.1001/jama.1992.03490020088036

Clarke M, Brice A, Chalmers I. Accumulating research: a systematic account of how cumulative meta-analyses would have provided knowledge, improved health, reduced harm and saved resources. PLoS One. 2014;9(7):e102670. https://doi.org/10.1371/journal.pone.0102670

Duyx B, Urlings MJE, Swaen GMH, Bouter LM, Zeegers MP. Scientific citations favor positive results: a systematic review and meta-analysis. J Clin Epidemiol. 2017;88:92-101. https://doi.org/10.1016/j.jclinepi.2017.06.002

Oxman AD, Paulsen EJ. Who can you trust? A review of free online sources of "trustworthy" information about treatment effects for patients and the public. BMC Med Inform Decis Mak. 2019;19(1):35. https://doi.org/10.1186/s12911-019-0772-5

Page MJ, Shamseer L, Altman DG, Tetzlaff J, Sampson M, Tricco AC, et al. Epidemiology and reporting characteristics of systematic reviews of biomedical research: a cross-sectional study. PLoS Med. 2016b;13(5):e1002028. https://doi.org/10.1371/journal.pmed.1002028

Robinson KA, Goodman SN. A systematic examination of the citation of prior research in reports of randomized, controlled trials. Ann Intern Med. 2011;154(1):50-5. https://doi.org/10.7326/0003-4819-154-1-201101040-00007

Robson RC, Pham B, Hwee J, Thomas SM, Rios P, Page MJ, et al. Few studies exist examining methods for selecting studies, abstracting data, and appraising quality in a systematic review. J Clin Epidemiol. 2019;106:121-35. https://doi.org/10.1016/j.jclinepi.2018.10.003

Rosenberger KJ, Xu C, Lin L. Methodological assessment of systematic reviews and meta-analyses on COVID-19: A meta-epidemiological study. J Eval Clin Pract. 2021;27(5):1123-33. https://doi.org/10.1111/jep.13578

Roundtree AK, Kallen MA, Lopez-Olivo MA, Kimmel B, Skidmore B, Ortiz Z, et al. Poor reporting of search strategy and conflict of interest in over 250 narrative and systematic reviews of two biologic agents in arthritis: a systematic review. J Clin Epidemiol. 2009;62(2):128-37. https://doi.org/10.1016/j.jclinepi.2008.08.003

Rucker B, Umbarger E, Ottwell R, Arthur W, Brame L, Woodson E, et al. Evaluation of spin in the abstracts of systematic reviews and meta-analyses focused on tinnitus. Otol Neurotol. 2021:1237-44. https://doi.org/10.1097/mao.0000000000003178

Waffenschmidt S, Knelangen M, Sieben W, Bühn S, Pieper D. Single screening versus conventional double screening for study selection in systematic reviews: a methodological systematic review. BMC Med Res Methodol. 2019;19(1):132. https://doi.org/10.1186/s12874-019-0782-0

Wang Z, Nayfeh T, Tetzlaff J, O'Blenis P, Murad MH. Error rates of human reviewers during abstract screening in systematic reviews. PLoS One. 2020;15(1):e0227742. https://doi.org/10.1371/journal.pone.0227742

Other reviews

Tsafnat G, Glasziou P, Choong MK, Dunn A, Galgani F, Coiera E. Systematic review automation technologies. Syst Rev. 2014;3:74. https://doi.org/10.1186/2046-4053-3-74

Research studies

E JY, Saldanha IJ, Canner J, Schmid CH, Le JT, Li T. Adjudication rather than experience of data abstraction matters more in reducing errors in abstracting data in systematic reviews. Res Synth Methods. 2020;11(3):354-62. https://doi.org/10.1002/jrsm.1396

Ewald H, Klerings I, Wagner G, Heise TL, Dobrescu AI, Armijo-Olivo S, et al. Abbreviated and comprehensive literature searches led to identical or very similar effect estimates: a meta-epidemiological study. J Clin Epidemiol. 2020;128:1-12. https://doi.org/10.1016/j.jclinepi.2020.08.002

Gartlehner G, Affengruber L, Titscher V, Noel-Storr A, Dooley G, Ballarini N, et al. Single-reviewer abstract screening missed 13 percent of relevant studies: a crowd-based, randomized controlled trial. J Clin Epidemiol. 2020;121:20-8. https://doi.org/10.1016/j.jclinepi.2020.01.005

Huang X, Lin J, Demner-Fushman D. Evaluation of PICO as a knowledge representation for clinical questions. AMIA Annu Symp Proc. 2006;2006:359-63. http://www.ncbi.nlm.nih.gov/pmc/articles/pmc1839740/

Leng RI. A network analysis of the propagation of evidence regarding the effectiveness of fat-controlled diets in the secondary prevention of coronary heart disease (CHD): Selective citation in reviews. PLoS One. 2018;13(5):e0197716. https://doi.org/10.1371/journal.pone.0197716

Marshall IJ, Marshall R, Wallace BC, Brassey J, Thomas J. Rapid reviews may produce different results to systematic reviews: a meta-epidemiological study. J Clin Epidemiol. 2019;109:30-41. https://doi.org/10.1016/j.jclinepi.2018.12.015

Page MJ, Forbes A, Chau M, Green SE, McKenzie JE. Investigation of bias in meta-analyses due to selective inclusion of trial effect estimates: empirical study. BMJ Open. 2016;6(4):e011863. https://doi.org/10.1136/bmjopen-2016-011863

Scott AM, Forbes C, Clark J, Carter M, Glasziou P, Munn Z. Systematic review automation tools improve efficiency but lack of knowledge impedes their adoption: a survey. J Clin Epidemiol. 2021;138:80-94. https://doi.org/10.1016/j.jclinepi.2021.06.030

Urlings MJE, Duyx B, Swaen GMH, Bouter LM, Zeegers MP. Selective citation in scientific literature on the human health effects of bisphenol A. Res Integr Peer Rev. 2019;4:6. https://doi.org/10.1186/s41073-019-0065-7

Urlings MJE, Duyx B, Swaen GMH, Bouter LM, Zeegers MP. Citation bias and other determinants of citation in biomedical research: findings from six citation networks. J Clin Epidemiol. 2021;132:71-8. https://doi.org/10.1016/j.jclinepi.2020.11.019

Other references

Booth A, Clarke M, Dooley G, Ghersi D, Moher D, Petticrew M, et al. The nuts and bolts of PROSPERO: an international prospective register of systematic reviews. Syst Rev. 2012;1:2. https://doi.org/10.1186/2046-4053-1-2

Cumpston MS, McKenzie JE, Thomas J, Brennan SE. The use of 'PICO for synthesis' and methods for synthesis without meta-analysis: protocol for a survey of current practice in systematic reviews of health interventions. F1000Res. 2020;9:678. https://doi.org/10.12688/f1000research.24469.2

Lefebvre C, Glanville J, Briscoe S, Littlewood AM, C, Metzendorf M-I, Noel-Storr A, et al. Searching for and selecting studies. Cochrane Handbook for Systematic Reviews of Interventions version 6,2. 2021. https://training.cochrane.org/handbook/current/chapter-04

Oxman AD, Cook DJ, Guyatt GH. Users' guides to the medical literature. VI. How to use an overview. Evidence-Based Medicine Working Group. JAMA. 1994;272(17):1367-71. https://doi.org/10.1001/jama.272.17.1367

Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ. 2021;372:n160. https://doi.org/10.1136/bmj.n160

2.2b Consider whether unpublished results were considered.

Explanation

Many fair comparisons are never published, and outcomes are sometimes left out from those that are published. Those that are published are more likely to report favourable results. Consequently, reliance on published reports alone sometimes results in the beneficial effects of treatments being overestimated and the adverse effects being underestimated.

For example, among trials of antidepressant drugs submitted to the U.S. Federal Drug Administration (FDA) or the Swedish drug regulatory authority, efficacy trials reporting positive results and larger effect sizes were more likely to be published subsequently. A review of trials supporting new medicines approved by the FDA between 1998 and 2000 found that over half of all supporting trials for FDA-approved drugs remained unpublished more five or more years after approval *[Lee 2008 (RS)]*. Selective reporting of trial results was found for commonly marketed medicines.

Biased under-reporting of research is a major problem that is far from being solved. It is scientific and ethical malpractice and wastes research resources. Selective reporting is an important reason why fair comparisons of treatments should begin with protocols that are registered and searchable in registries such as clinicaltrials.gov. This can also help to reduce selective reporting of some outcomes but not others in published reports, depending on the nature and direction of the results.

Basis for this concept

Many registered trials (comparisons of treatments) are not published *[Chapman 2014 (SR), Hopewell 2009 (SR)]*. A systematic review of publication bias found five studies that examined the association between trial results and publication of registered trials *[Hopewell 2009 (SR)]*. The studies compared publication of trials with "positive findings" (that were "statistically significant", perceived to be important or striking, or indicating a desirable treatment effect) and with "negative findings" (that were not "statistically significant", perceived to be unimportant, or indicating an undesirable treatment effect or lack of effect). It found that trials with positive findings were nearly twice as likely to be published as studies with negative findings. Two studies that examined time-to-publication found that among the published trials, trials with positive findings tended to be published after four to five years compared to those with negative findings, which were published after six to eight years. A more recent systematic review found 20 studies and also found strong evidence of publication bias *[Dwan 2013 (SR)]*. A systematic review of comparisons between the results of trials found in "grey literature" (e.g., conference abstracts, research reports, book chapters, dissertations, policy documents, personal correspondence) compared to trials found in journals observed that published trials also tend to have larger effects than trials found in the grey literature *[Hopewell 2007 (SR)]*. When not recognised and addressed in systematic reviews, publication bias can sometimes result in overestimation of effects *[Schwab 2021 (SR)]*.

Discrepancies between the outcomes that researchers say they will measure in trials and what they report are common *[Fleming 2015 (SR), Jones 2015 (SR)]*. Outcome reporting bias occurs when researchers select for publication a subset of the original recorded outcomes based on knowledge of the results. Comparisons between trial protocols and reports of results have shown that outcomes are more likely to be reported when there is a "statistically significant" effect than when there is a "statistically nonsignificant effect" *[Dwan 2013 (SR)]*. When not recognised and not addressed in systematic reviews, outcome reporting bias can result in overestimation of treatment effects *[Kirkham 2018 , Kirkham 2010 (RS)]*.

Requiring researchers to register trials in databases such as clinicaltrials.gov has helped to address and reduce publication and reporting bias but has not eliminated the problem *[Baudard 2017 (SR), Dechartres*

2016 (RS), *Köhler 2015 (RS)*, *Manheimer 2002 (RS)*, *Papageorgiou 2018 (SR)]*. It is important that systematic review authors address these risks *[Kirkham 2018]*.

Implications

Be aware of the possibility of biased underreporting of fair comparisons and assess whether the authors of systematic reviews have addressed this risk.

References

Systematic reviews

Baudard M, Yavchitz A, Ravaud P, Perrodeau E, Boutron I. Impact of searching clinical trial registries in systematic reviews of pharmaceutical treatments: methodological systematic review and reanalysis of meta-analyses. BMJ. 2017;356:j448. https://doi.org/10.1136/bmj.j448

Chapman SJ, Shelton B, Mahmood H, Fitzgerald JE, Harrison EM, Bhangu A. Discontinuation and non-publication of surgical randomised controlled trials: observational study. BMJ. 2014;349:g6870. https://doi.org/10.1136/bmj.g6870

Dwan K, Gamble C, Williamson PR, Kirkham JJ. Systematic review of the empirical evidence of study publication bias and outcome reporting bias - an updated review. PLoS One. 2013;8(7):e66844. https://doi.org/10.1371/journal.pone.0066844

Fleming PS, Koletsi D, Dwan K, Pandis N. Outcome discrepancies and selective reporting: impacting the leading journals? PLoS One. 2015;10(5):e0127495. https://doi.org/10.1371/journal.pone.0127495

Hopewell S, Loudon K, Clarke MJ, Oxman AD, Dickersin K. Publication bias in clinical trials due to statistical significance or direction of trial results. Cochrane Database Syst Rev. 2009(1):MR000006. https://doi.org/10.1002/14651858.mr000006.pub3

Hopewell S, McDonald S, Clarke MJ, Egger M. Grey literature in meta-analyses of randomized trials of health care interventions. Cochrane Database Syst Rev. 2007(2). https://doi.org//10.1002/14651858.MR000010.pub3

Jones CW, Keil LG, Holland WC, Caughey MC, Platts-Mills TF. Comparison of registered and published outcomes in randomized controlled trials: a systematic review. BMC Med. 2015;13:282. https://doi.org/10.1186/s12916-015-0520-3

Papageorgiou SN, Xavier GM, Cobourne MT, Eliades T. Registered trials report less beneficial treatment effects than unregistered ones: a meta-epidemiological study in orthodontics. J Clin Epidemiol. 2018;100:44-52. https://doi.org/10.1016/j.jclinepi.2018.04.017

Schwab S, Kreiliger G, Held L. Assessing treatment effects and publication bias across different specialties in medicine: a meta-epidemiological study. BMJ Open. 2021;11(9):e045942. https://doi.org/10.1136/bmjopen-2020-045942

Research studies

Dechartres A, Ravaud P, Atal I, Riveros C, Boutron I. Association between trial registration and treatment effect estimates: a meta-epidemiological study. BMC Med. 2016;14(1):100. https://doi.org/10.1186/s12916-016-0639-x

Kirkham JJ, Dwan KM, Altman DG, Gamble C, Dodd S, Smyth R, et al. The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. BMJ. 2010;340:c365. https://doi.org/10.1136/bmj.c365

Köhler M, Haag S, Biester K, Brockhaus AC, McGauran N, Grouven U, et al. Information on new drugs at market entry: retrospective analysis of health technology assessment reports versus regulatory reports, journal publications, and registry reports. BMJ. 2015;350:h796. https://doi.org/10.1136/bmj.h796

Lee K, Bacchetti P, Sim I. Publication of clinical trials supporting successful new drug applications: a literature analysis. PLoS Med. 2008;5(9):e191. https://doi.org/10.1371/journal.pmed.0050191

Manheimer E, Anderson D. Survey of public information about ongoing clinical trials funded by industry: evaluation of completeness and accessibility. BMJ. 2002;325(7363):528-31. https://doi.org/10.1136/bmj.325.7363.528

Other references

Kirkham JJ, Altman DG, Chan AW, Gamble C, Dwan KM, Williamson PR. Outcome reporting bias in trials: a methodological approach for assessment and adjustment in systematic reviews. BMJ. 2018;362:k3802. https://doi.org/10.1136/bmj.k3802

2.2c Consider whether treatments were compared across studies.

Explanation

For many conditions (e.g., depression) there are more than two possible treatments (for example, different medicines, or types of psychotherapy). Only very rarely are all the possible treatments for a condition compared in a single study, so it may be necessary to consider indirect comparisons among treatments. For example, there may be comparisons of drug A with placebo and comparisons of drug B with placebo, but no studies that compare drug A with drug B directly. In this case, indirect comparisons among studies may be needed to inform a decision about whether to use drug A or drug B. However, there can be important differences between the studies examined in addition to the treatments they assessed, for example, differences in characteristics of the participants, or the way the comparisons were done, or in the outcome measures used. These differences can result in misleading estimates of treatment effects.

A systematic review of different doses of aspirin illustrates the problem with indirect comparisons *[Guyatt 2011b]*. The authors found five randomized trials that compared aspirin with placebo to prevent graft occlusion after coronary artery bypass surgery. Two trials tested medium-dose and three low-dose aspirin. Based on the indirect comparison, the relative risk reduction for medium- compared to low-dose aspirin was 0.74 (95% confidence interval 0.52 to 1.06; P = 0.10) suggesting the possibility of a larger effect with medium-dose aspirin. However, there are other characteristics of the trials that might be responsible for any differences found (or undetected differences that might exist). Compared with the low-dose trials, the patients included in the medium-dose trials may be different, interventions other than aspirin may have been differently administered, and outcomes may have been measured differently (e.g., dissimilar criteria for occlusion or different durations of follow-up). Differences in study methods and the risk of bias may also explain the results.

Basis for this concept

Indirect comparisons are non-randomized, even though they are based on two or more randomized trials. For indirect comparisons to be reliable, patient and other characteristics of the treatment comparisons must be similar across the trials included in the indirect comparison. As with other types of non-randomized studies, it is only possible to control for characteristics (confounders) that might modify the effects of treatments that are known, measured, and reported (see Concept 2.1a). Therefore, indirect comparisons can sometimes either overestimate or underestimate treatment effects *[Bucher 1997 , Song 2003 (SR)]*. Informal indirect comparisons – e.g. assuming that drug A is more effective than drug B simply because drug A had a larger effect compared to placebo than drub B – can be misleading and should be avoided *[Song 2009 (SR)]*. A systematic review of meta-analyses of randomized trials found that appropriately analysed indirect comparisons usually, but not always, agreed with those of direct comparisons *[Song 2003 (SR)]*. The reliability of the indirect comparisons depended on the risk of bias in the trials and the similarity of the trials.

However, there are often more than two treatment options for a condition and unreliable or no direct comparisons of all the treatments. When this is the case, indirect comparisons may provide the best available evidence to inform decisions. An increasing number of systematic reviews of multiple treatments for a condition use what is called “network meta-analysis” to evaluate the comparative effectiveness of multiple treatments. For each pair of treatments, these analyses combine effect estimates from direct and indirect comparisons. As with any systematic review, the reliability of estimates of treatment effects from network meta-analyses depends on the methods used to identify, select, critically appraise, and collect data from relevant studies (see Concept 2.2a). In addition, the reliability of effect estimates, and the ranking of treatments, depends on assessing the similarity of the included trials (apart from the treatments being compared and the consistency of direct and indirect effect estimates) *[Brignardello-Petersen 2018 , Jansen 2014 , Mills 2012 , Puhan 2014]*.

Network meta-analyses rely on the assumption that the different sets of studies included in the analysis are similar, on average, in all important factors that may affect the relative effects [*Chaimani 2021*], including characteristics of the participants, interventions, and outcome measures. This assumption cannot be tested statistically, but it is sometimes possible to adjust for potential confounders [*Efthimiou 2016 (SR)*, *Hutton 2015* , *Jansen 2013*]. Otherwise, it must be assessed conceptually, based on what is known about potential confounders and what information is available from the trials.

Direct and indirect evidence for a treatment comparison should be combined only when the effect estimates are similar [*Hutton 2015*]. Statistical tests can be used to assess whether differences in effect are greater than could be expected to occur by chance. However, the tests have limited ability ("power") to confirm that differences are larger than could be expected by chance. On the other hand, when multiple tests are undertaken, a few may indicate inconsistency simply by chance. A systematic review found 112 trial groups of trials that included both a direct and indirect comparison of two treatments [*Song 2011 (SR)*]. The direct and indirect comparisons were inconsistent 14% of the time, suggesting that inconsistency may be more common than in the previous systematic review noted above [*Song 2003 (SR)*]. However, assessments of inconsistency may differ depending on the test that is used, the effect measure used in the analysis, and how much variation there is in effect estimates from different studies [*Veroniki 2013 (SR)*].

Implications

Indirect comparisons are sometimes needed to inform treatment choices. In these circumstances, careful consideration should be given to differences between the studies besides the treatments that were compared.

References

Systematic reviews

Efthimiou O, Debray TP, van Valkenhoef G, Trelle S, Panayidou K, Moons KG, et al. GetReal in network meta-analysis: a review of the methodology. Res Synth Methods. 2016;7(3):236-63. https://doi.org/10.1002/jrsm.1195

Song F, Altman DG, Glenny AM, Deeks JJ. Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses. BMJ. 2003;326(7387):472. https://doi.org/10.1136/bmj.326.7387.472

Song F, Loke YK, Walsh T, Glenny AM, Eastwood AJ, Altman DG. Methodological problems in the use of indirect comparisons for evaluating healthcare interventions: survey of published systematic reviews. BMJ. 2009;338:b1147. https://doi.org/10.1136/bmj.b1147

Song F, Xiong T, Parekh-Bhurke S, Loke YK, Sutton AJ, Eastwood AJ, et al. Inconsistency between direct and indirect comparisons of competing interventions: meta-epidemiological study. BMJ. 2011;343:d4909. https://doi.org/10.1136/bmj.d4909

Research studies

Veroniki AA, Vasiliadis HS, Higgins JP, Salanti G. Evaluation of inconsistency in networks of interventions. Int J Epidemiol. 2013;42(1):332-45. https://doi.org/10.1093/ije/dys222

Other references

Brignardello-Petersen R, Bonner A, Alexander PE, Siemieniuk RA, Furukawa TA, Rochwerg B, et al. Advances in the GRADE approach to rate the certainty in estimates from a network meta-analysis. J Clin Epidemiol. 2018;93:36-44. https://doi.org/10.1016/j.jclinepi.2017.10.005

Bucher HC, Guyatt GH, Griffith LE, Walter SD. The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. J Clin Epidemiol. 1997;50(6):683-91. https://doi.org/10.1016/s0895-4356(97)00049-8

Chaimani A, Caldwell D, Li T, Higgins J, Salanti G. Undertaking network meta-analyses. Cochrane Handbook for Systematic Reviews of Interventions version 6,2. 2021. https://training.cochrane.org/handbook/current/chapter-11

Guyatt GH, Oxman AD, Kunz R, Woodcock J, Brozek J, Helfand M, et al. GRADE guidelines: 8. Rating the quality of evidence--indirectness. J Clin Epidemiol. 2011b;64(12):1303-10. https://doi.org/10.1016/j.jclinepi.2011.04.014

Hutton B, Salanti G, Caldwell DM, Chaimani A, Schmid CH, Cameron C, et al. The PRISMA extension statement for reporting of systematic reviews incorporating network meta-analyses of health care interventions: checklist and explanations. Ann Intern Med. 2015;162(11):777-84. https://doi.org/10.7326/m14-2385

Jansen JP, Naci H. Is network meta-analysis as valid as standard pairwise meta-analysis? It all depends on the distribution of effect modifiers. BMC Med. 2013;11:159. https://doi.org/10.1186/1741-7015-11-159

Jansen JP, Trikalinos T, Cappelleri JC, Daw J, Andes S, Eldessouki R, et al. Indirect treatment comparison/network meta-analysis study questionnaire to assess relevance and credibility to inform health care decision making: an ISPOR-AMCP-NPC Good Practice Task Force report. Value Health. 2014;17(2):157-73. https://doi.org/10.1016/j.jval.2014.01.004

Mills EJ, Ioannidis JP, Thorlund K, Schünemann HJ, Puhan MA, Guyatt GH. How to use an article reporting a multiple treatment comparison meta-analysis. JAMA. 2012;308(12):1246-53. https://doi.org/10.1001/2012.jama.11228

Puhan MA, Schünemann HJ, Murad MH, Li T, Brignardello-Petersen R, Singh JA, et al. A GRADE Working Group approach for rating the quality of treatment effect estimates from network meta-analysis. BMJ. 2014;349:g5630. https://doi.org/10.1136/bmj.g5630

2.2d Consider whether important assumptions were tested.

Explanation

Sometimes treatment claims are based on chains of evidence, or models. For example, the effects of using a diagnostic test may depend on how accurate the test is, assumptions about what will be done based on the test results, and evidence of the effects of what is done. Similarly, evidence of the effects of public health and health system policies sometimes comes from models that combine different types of studies and assumptions; and assumptions are sometimes made when fair comparisons are combined in systematic reviews. When treatment comparisons depend on assumptions, it is important to consider their basis and to test how sensitive the results are to plausible changes in the assumptions made. For example, a model used to compare the effects of using different diagnostic tests on outcomes that are important to patients might require assumptions about what actions doctors or patients will take, based on test results. If that is uncertain, it is important to consider whether changing the assumptions has a substantial impact on the estimated difference in outcomes important to patients.

During and prior to the Covid-19 pandemic there have been few randomized trials of public health measures used to control spread of infections, such as school closures *[Glasziou 2021]*. As a result, estimates of the effects of those interventions have frequently been based on models and non-randomized studies. The modelling studies make many different assumptions and often suggest different effects. For example, some modelling studies have suggested that school closures can reduce community transmission of the coronavirus, while others disagree *[Walsh 2021 (SR)]*. These models depend on many assumptions, and changes in these assumptions can change the results. Different models make different assumptions about per-contact transmission probabilities, how many parents go to work or work at home when schools are closed or opened, changes in contacts outside of home because of schools closing or opening, what other protective measures are in place, what happens during holidays, what proportion of infected people have symptoms, how long they are infected before they have symptoms and are tested, how long the symptoms last, contact tracing, how many people without symptoms are tested, the accuracy of testing, delays in getting test results, and compliance with and effects of isolation and quarantine. Because of all these assumptions and important uncertainty about many of them, the results of these modelling studies are very uncertain.

Early in the pandemic, some assumptions were empirically informed, such as how populations are distributed spatially. However, other assumptions were seemingly anecdotal, such as an assumption that children were twice as likely as adults to transmit the coronavirus. That assumption helped justify school closures. However, subsequent epidemiological studies suggested, if anything, children may be less likely to transmit the virus *[Reddy 2020]*. In addition, some models did not consider health consequences beyond deaths from coronavirus or how social and economic consequences might affect health. Models can be helpful when there is extreme uncertainty, but it is important to recognise their limitations and uncertainty.

Basis for this concept

Many different types of models are used to estimate treatment effects. One type is marginal structural models, which are increasingly used in analyses of routinely collected data. These models take account of confounders arising during follow-up when patients switch or stop treatments, as well as baseline differences. Like all non-randomized study designs, the underlying assumption is that all relevant confounders are known, measured, and correctly integrated in the analyses. A systematic review compared treatment effects found in marginal structural model studies with those found in randomized trials for mortality and other outcomes *[Ewald 2020 (SR)]*. The review found important differences, including effects going in the opposite direction for eight of the 19 included comparisons.

New medicines are normally approved for marketing based on the results of randomized trials. A systematic review of medicines that were approved for marketing without randomized trials found that the majority of models that were used to estimate effects were based on "historical controls" (how patients were treated in the past) without any adjustment for differences in patient population (see Concept 1.2e), with a high risk of bias *[Hatswell 2017 (SR)]*.

Modelling studies combine information from a variety of sources to compare treatments. Expert judgement is often used when there is limited or conflicting evidence about a variable or "parameter" included in a model. Systematic reviews of the use of expert judgements in modelling studies in health research and in health technology assessments found extensive use of expert judgement, but most modelling studies did not provide adequate details of how expert judgements were elicited *[Cadham 2021 (SR), Grigore 2013 (SR)]*. This makes it difficult to assess the reliability of those judgements and the findings of the modelling studies. Expert judgements may be misleading due to cognitive biases, overconfidence, and the choice of experts *[Morgan 2014]*. To reduce the risk of misleading judgements, there should be a protocol for selecting experts, helping them make systematic and transparent judgements, and combining (or not combining) judgements from different experts *[Morgan 2014 , Schunemann 2019]* (See Concept 1.4c).

When direct evidence is lacking, models can be used to link together evidence of the effects of screening (see Concept 1.3e) or diagnostic tests on outcomes that are important to people *[Petitti 2018]*. However, these models also can be misleading *[Koleva-Kolarova 2015 (SR)]*. Overall, the certainty of these models corresponds to the certainty of the weakest link in the chain of evidence *[Schünemann 2019]*.

Modelling is unavoidable in evaluations of the cost-effectiveness of treatments and decision analyses [*Buxton 1997*]. However, these models can be misleading. For example, a systematic review of models assessing the cost-effectiveness of antipsychotic medication for schizophrenia found 60 models *[Jin 2020 (SR)]*. The models varied greatly, and the quality of the models was generally low due to failure to capture the health and cost impact of adverse effects and input data from the best available source.

Challenges with modelling studies include choosing which technique to use (and not making an arbitrary or biased choice), avoiding arbitrary (or biased) ranges for variables (parameters) when examining the impact of uncertainty, and making details of the model available when that is in conflict with the "intellectual property" generated by a substantial investment in developing a model *[Caro 2012]*. The trustworthiness of a model depends on transparency and validation *[Eddy 2012]*. Unfortunately, both are often lacking, making it difficult to judge how much confidence can be placed in the findings of a model. Sensitivity analyses can be used to assess the uncertainty of a model from the assumptions that are made. Sources of uncertainty include uncertainty about the values or data used as input for each variable (parameter) in the model, uncertainty about the model (how the variables are combined), and uncertainty about how the model compares to other models using different methods). A systematic review of 406 cost-effectiveness analyses found that most analyses only addressed one of those sources of uncertainty (most often uncertainty about the variables) and that sensitivity analyses were often poorly reported *[Jain 2011 (SR)]*.

In summary, modelling studies can provide valuable information about the effects of treatments and treatment choices, but when they are used to assess the effects of treatments or to inform decisions, their reliability and uncertainty need to be carefully assessed and reported *[Briggs 2012 , Brozek 2021 , Egger 2017]*.

Implications

Whenever treatment comparisons depend on assumptions, consider whether the assumptions are well-founded and how sensitive the results are to plausible changes in the assumptions that are made.

References

Systematic reviews

Cadham CJ, Knoll M, Sánchez-Romero LM, Cummings KM, Douglas CE, Liber A, et al. The use of expert elicitation among computational modeling studies in health research: a systematic review. Med Decis Making. 2021:272989x211053794. https://doi.org/10.1177/0272989x211053794

Ewald H, Ioannidis JPA, Ladanie A, Mc Cord K, Bucher HC, Hemkens LG. Nonrandomized studies using causal-modeling may give different answers than RCTs: a meta-epidemiological study. J Clin Epidemiol. 2020;118:29-41. https://doi.org/10.1016/j.jclinepi.2019.10.012

Grigore B, Peters J, Hyde C, Stein K. Methods to elicit probability distributions from experts: a systematic review of reported practice in health technology assessment. Pharmacoeconomics. 2013;31(11):991-1003. https://doi.org/10.1007/s40273-013-0092-z

Hatswell AJ, Freemantle N, Baio G. Economic Evaluations of Pharmaceuticals Granted a Marketing Authorisation Without the Results of Randomised Trials: A Systematic Review and Taxonomy. Pharmacoeconomics. 2017;35(2):163-76. https://doi.org/10.1007/s40273-016-0460-6

Jain R, Grabner M, Onukwugha E. Sensitivity analysis in cost-effectiveness studies: from guidelines to practice. Pharmacoeconomics. 2011;29(4):297-314. https://doi.org/10.2165/11584630-000000000-00000

Jin H, Tappenden P, Robinson S, Achilla E, Aceituno D, Byford S. Systematic review of the methods of health economic models assessing antipsychotic medication for schizophrenia. PLoS One. 2020;15(7):e0234996. https://doi.org/10.1371/journal.pone.0234996

Koleva-Kolarova RG, Zhan Z, Greuter MJ, Feenstra TL, De Bock GH. Simulation models in population breast cancer screening: A systematic review. Breast. 2015;24(4):354-63. https://doi.org/10.1016/j.breast.2015.03.013

Walsh S, Chowdhury A, Braithwaite V, Russell S, Birch JM, Ward JL, et al. Do school closures and school reopenings affect community transmission of COVID-19? A systematic review of observational studies. BMJ Open. 2021;11(8):e053371. https://doi.org/10.1136/bmjopen-2021-053371

Other references

Briggs AH, Weinstein MC, Fenwick EA, Karnon J, Sculpher MJ, Paltiel AD. Model parameter estimation and uncertainty: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force--6. Value Health. 2012;15(6):835-42. https://doi.org/10.1016/j.jval.2012.04.014

Brozek JL, Canelo-Aybar C, Akl EA, Bowen JM, Bucher J, Chiu WA, et al. GRADE Guidelines 30: the GRADE approach to assessing the certainty of modeled evidence-An overview in the context of health decision-making. J Clin Epidemiol. 2021;129:138-50. https://doi.org/10.1016/j.jclinepi.2020.09.018

Buxton MJ, Drummond MF, Van Hout BA, Prince RL, Sheldon TA, Szucs T, et al. Modelling in economic evaluation: an unavoidable fact of life. Health Econ. 1997;6(3):217-27. https://doi.org/10.1002/(sici)1099-1050(199705)6:3%3C217::aid-hec267%3E3.0.co;2-w

Caro JJ, Briggs AH, Siebert U, Kuntz KM. Modeling good research practices--overview: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force--1. Value Health. 2012;15(6):796-803. https://doi.org/10.1016/j.jval.2012.06.012

Eddy DM, Hollingworth W, Caro JJ, Tsevat J, McDonald KM, Wong JB. Model transparency and validation: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force--7. Value Health. 2012;15(6):843-50. https://doi.org/10.1016/j.jval.2012.04.012

Egger M, Johnson L, Althaus C, Schöni A, Salanti G, Low N, et al. Developing WHO guidelines: Time to formally include evidence from mathematical modelling studies. F1000Res. 2017;6:1584. https://doi.org/10.12688/f1000research.12367.2

Glasziou PP, Michie S, Fretheim A. Public health measures for covid-19. BMJ. 2021;375:n2729. https://doi.org/10.1136/bmj.n2729

Morgan MG. Use (and abuse) of expert elicitation in support of decision making for public policy. Proc Natl Acad Sci U S A. 2014;111(20):7176-84. https://doi.org/10.1073/pnas.1319946111

Petitti DB, Lin JS, Owens DK, Croswell JM, Feuer EJ. Collaborative Modeling: Experience of the U.S. Preventive Services Task Force. Am J Prev Med. 2018;54(1s1):S53-s62. https://doi.org/10.1016/j.amepre.2017.07.003

Reddy S. How epidemiological models fooled us into trusting bad assumptions. Barrons. April 29, 2020. https://www.barrons.com/articles/the-danger-of-overreliance-on-epidemiological-models-51588179008

Schünemann HJ, Mustafa RA, Brozek J, Santesso N, Bossuyt PM, Steingart KR, et al. GRADE guidelines: 22. The GRADE approach for tests and strategies-from test accuracy to patient-important outcomes and recommendations. J Clin Epidemiol. 2019;111:69-82. https://doi.org/10.1016/j.jclinepi.2019.02.003

Schunemann HJ, Zhang Y, Oxman AD. Distinguishing opinion from evidence in guidelines. BMJ. 2019;366:l4606. https://doi.org/10.1136/bmj.l4606

2.3 Descriptions of effects should clearly reflect *the size of the effects*.

2.3a Be cautious of verbal descriptions alone of the size of effects.

Explanation

A treatment effect (a difference in outcomes in a comparison) is a numerical concept, but it may be difficult to understand quantitative information about the effects of treatments. Qualitative (descriptive) labels may be easier to understand and can be helpful. However, qualitative descriptions of effects may mean different things to different people, for example, saying that a treatment will 'slightly reduce', 'reduce', or 'greatly reduce' the likelihood of an undesirable outcome; or that a side effect is 'frequent' or 'rare'. In addition, verbal descriptions of treatments can be manipulative, for example, promising 'amazing results' or describing treatments as 'natural', implying that they are safe because of that.

Patients' perceptions of verbal descriptions of effects can affect their decisions. For example, a randomized comparison of verbal descriptors suggested by the European Union, such as "common" and "rare" compared to numerical descriptions found that those verbal descriptions were associated with overestimation of the likelihood of side effects *[Knapp 2004 (RS)]*. Patients shown verbal descriptions had more negative perceptions of the medicine than those shown numerical descriptions, and they were more likely to say that the information would affect their decision to take the medicine.

Basis for this concept

Verbal expressions of uncertainty or probability often mean different things to different people and some verbal expressions may be easier to understand than others *[Knapp 2004 (RS), Mazur 1991 (RS), Morgan 2014 , Trevena 2006 (SR), Visschers 2009 (SR), Wills 2003 (SR), Zipkin 2014 (SR)]*. Use of consistent language that has been tested can improve the understanding, usability, and usefulness of information about intervention effects *[Glenton 2010 (RS), Santesso 2015 (RS)]*.

Words may be easier to understand than numbers, and words used to express probabilities may be ordered consistently, but because their interpretation is highly variable, they may result in inappropriate perceptions and decisions *[Burkell 2004 (OR), Knapp 2004 (RS), Kong 1986 (RS), Lipkus 2007 (OR), Wills 2003 (SR)]*. Numbers are more accurate, but many people have poor numeracy skills and may have problems understanding effect estimates *[Lipkus 2007 (OR), Trevena 2006 (SR)]*. People differ in their preferences for words, numbers, or both *[Wills 2003 (SR)]*. Combinations of words and quantitative presentations are likely to have advantages over quantitative presentations alone as this can help to interpret and ensure understanding of numbers *[Lipkus 2007 (OR), Oxman 2020 (OR)]*. Carefully designed tables that summarise estimates of treatment effects from a systematic review are perceived as understandable and useful, and they can improve how quickly people find key information, understanding, accurate perceptions of effects, and choices *[Brandt 2017 (RS), Rosenbaum 2010a (RS), Rosenbaum 2010b (RS), Santesso 2015 (RS), Schwartz 2009 (RS)]*.

Implications

A verbal description of a treatment effect can be helpful, but it should be considered together with quantitative information about the size of the effect. Be wary of manipulative use of language in descriptions of treatment effects.

References

Systematic reviews

Trevena LJ, Davey HM, Barratt A, Butow P, Caldwell P. A systematic review on communicating with patients about evidence. J Eval Clin Pract. 2006;12(1):13-23. https://doi.org/10.1111/j.1365-2753.2005.00596.x

Visschers VH, Meertens RM, Passchier WW, de Vries NN. Probability information in risk communication: a review of the research literature. Risk Anal. 2009;29(2):267-87. https://doi.org/10.1111/j.1539-6924.2008.01137.x

Wills CE, Holmes-Rovner M. Patient comprehension of information for shared treatment decision making: state of the art and future directions. Patient Educ Couns. 2003;50(3):285-90. https://doi.org/10.1016/s0738-3991(03)00051-x

Zipkin DA, Umscheid CA, Keating NL, Allen E, Aung K, Beyth R, et al. Evidence-based risk communication: a systematic review. Ann Intern Med. 2014;161(4):270-80. https://doi.org/10.7326/m14-0295

Other reviews

Burkell J. What are the chances? Evaluating risk and benefit information in consumer health materials. J Med Libr Assoc. 2004;92(2):200-8. http://www.ncbi.nlm.nih.gov/pmc/articles/pmc385301/

Lipkus IM. Numeric, verbal, and visual formats of conveying health risks: suggested best practices and future recommendations. Med Decis Making. 2007;27(5):696-713. https://doi.org/10.1177/0272989x07307271

Oxman AD, Glenton C, Flottorp S, Lewin S, Rosenbaum S, Fretheim A. Development of a checklist for people communicating evidence-based information about the effects of healthcare interventions: a mixed methods study. BMJ Open. 2020;10(7):e036348. https://doi.org/10.1136/bmjopen-2019-036348

Research studies

Brandt L, Vandvik PO, Alonso-Coello P, Akl EA, Thornton J, Rigau D, et al. Multilayered and digitally structured presentation formats of trustworthy recommendations: a combined survey and randomised trial. BMJ Open. 2017;7(2):e011569. https://doi.org/10.1136/bmjopen-2016-011569

Glenton C, Santesso N, Rosenbaum S, Nilsen ES, Rader T, Ciapponi A, et al. Presenting the results of Cochrane Systematic Reviews to a consumer audience: a qualitative study. Med Decis Making. 2010;30(5):566-77. https://doi.org/10.1177/0272989x10375853

Knapp P, Raynor DK, Berry DC. Comparison of two methods of presenting risk information to patients about the side effects of medicines. Qual Saf Health Care. 2004;13(3):176-80. https://doi.org/10.1136/qhc.13.3.176

Kong A, Barnett GO, Mosteller F, Youtz C. How medical professionals evaluate expressions of probability. N Engl J Med. 1986;315(12):740-4. https://doi.org/10.1056/nejm198609183151206

Mazur DJ, Hickam DH. Patients' interpretations of probability terms. J Gen Intern Med. 1991;6(3):237-40. https://doi.org/10.1007/bf02598968

Rosenbaum SE, Glenton C, Nylund HK, Oxman AD. User testing and stakeholder feedback contributed to the development of understandable and useful Summary of Findings tables for Cochrane reviews. J Clin Epidemiol. 2010a;63(6):607-19. https://doi.org/10.1016/j.jclinepi.2009.12.013

Rosenbaum SE, Glenton C, Oxman AD. Summary-of-findings tables in Cochrane reviews improved understanding and rapid retrieval of key information. J Clin Epidemiol. 2010b;63(6):620-6. https://doi.org/10.1016/j.jclinepi.2009.12.014

Santesso N, Rader T, Nilsen ES, Glenton C, Rosenbaum S, Ciapponi A, et al. A summary to communicate evidence from systematic reviews to the public improved understanding and accessibility of information: a randomized controlled trial. J Clin Epidemiol. 2015;68(2):182-90. https://doi.org/10.1016/j.jclinepi.2014.04.009

Schwartz LM, Woloshin S, Welch HG. Using a drug facts box to communicate drug benefits and harms: two randomized trials. Ann Intern Med. 2009;150(8):516-27. https://doi.org/10.7326/0003-4819-150-8-200904210-00106

Other references

Morgan MG. Use (and abuse) of expert elicitation in support of decision making for public policy. Proc Natl Acad Sci U S A. 2014;111(20):7176-84. https://doi.org/10.1073/pnas.1319946111

2.3b Be cautious of relative effects of treatments alone.

Explanation

Relative effects are ratios, for example, the ratio of the probability of an outcome in one treatment group compared with that in a comparison group. They are insufficient for judging the importance of the difference (between the frequencies of the outcome). A relative effect may give the impression that a difference is more important than it actually is when the likelihood of the outcome is small to begin with. For example, if a treatment reduces the probability of getting an illness by 50% but also has harms, and the risk of getting the illness is 2 in 100, receiving the treatment may be worthwhile. If, however, the risk of getting the illness is 2 in 10,000, then receiving the treatment may not be worthwhile even though the *relative* effect is the same.

Absolute effects are differences, for example, the difference between the probability of an outcome in one treatment group compared with that in a comparison group. The absolute effect of a treatment is likely to vary for people with different baseline risks. Facemasks, for example, may have dramatically different effects depending on the baseline risk of infection *[Schünemann 2020]*. Facemasks reduce transmission of viruses, including coronavirus, but it is uncertain how effective they are for preventing Covid-19 infections *[Chu 2020 (SR), Glasziou 2021 , Talic 2021 (SR), Vestrheim 2020 (SR)]*. If we assume a 40% relative reduction in the number of new Covid-19 infections, it is possible to estimate the absolute effect for different baseline risks (see the table below). If the baseline risk is zero, it does not make a difference whether facemasks are used. The number of new infections is zero either way. If there is a low baseline risk, for example in the community when the incidence of Covid-19 is low (and not increasing), the difference is small (about eight fewer new infections if 10,000 people used facemasks for about two months). On the other hand, if the baseline risk is high, say for healthcare workers exposed to patients with Covid-19, the difference is much larger (about 700 fewer new infections per 10,000 people). In fact, the relative effect may also be larger for healthcare workers, if they use medical facemasks (rather than cloth masks), have training, and more often use facemasks correctly compared to people in the community. The absolute effect would also then be larger.

	Baseline risk[a]	**Risk with facemasks**[b]	**Difference**[c]	
No new infections	0	0	0	
Low risk	0.2%	0.12%	0.08%	(8 fewer per 10,000)
High risk	17.4%	10%	7%	(700 fewer per 10,000)

a) The low baseline risk corresponds to the number of new infections in eight weeks without facemasks if the two-week incidence is 50 per 100,000. The high risk is the assumed baseline risk from a systematic review *[Chu 2020 (SR)]*.

b) The risk with facemasks is based on reducing the baseline risk by 40% (the assumed relative risk reduction).

c) The difference between the baseline risk (without facemasks) and the risk with facemasks how many fewer new infections there would be with using facemasks compared to not using facemasks.

Basis for this concept

A relative effect may give readers the impression that a difference is more important than it actually is when the likelihood of the outcome is small to begin with. A systematic review found that showing people a relative risk reduction increased their willingness to get treatment, their willingness to advise treatment, and their willingness to pay to prevent the risk, compared to showing them the absolute effect *[Visschers 2009 (SR)]*. Another systematic review of randomized trials comparing people's responses to relative and absolute effects did not find a difference in understanding but found that relative effects were perceived to be larger and more persuasive *[Akl 2011b (SR)]*. A third systematic review found that presentations including risk differences were better than those including relative risk reductions for maximising accuracy and seemed less likely than presentations with relative risk reductions to influence decisions to accept a treatment *[Zipkin 2014 (SR)]*. Earlier systematic reviews had similar findings *[McGettigan 1999 (SR), Moxey 2003 (SR)]*.

Relative measures tend to be consistent across risk groups, whereas absolute measures do not *[Deeks 2002 (RS), Engels 2000 (RS), Furukawa 2002 (RS), Schmid 1998 (RS)]*. For this reason, meta-analyses tend to use a relative effect measure when estimating the average effect across studies. The risk difference can then be estimated by applying the relative effect to one or more relevant baseline risks *[Guyatt 2013a]*, as illustrated in the table above (if there is not a reason to expect different relative effects).

Implications

Always consider the absolute effects of treatments – that is, the difference in outcomes between the treatment groups being compared. Do not make a treatment decision based on relative effects alone.

References

Systematic reviews

Akl EA, Oxman AD, Herrin J, Vist GE, Terrenato I, Sperati F, et al. Using alternative statistical formats for presenting risks and risk reductions. Cochrane Database Syst Rev. 2011b(3):CD006776. https://doi.org/10.1002/14651858.cd006776.pub2

Chu DK, Akl EA, Duda S, Solo K, Yaacoub S, Schünemann HJ. Physical distancing, face masks, and eye protection to prevent person-to-person transmission of SARS-CoV-2 and COVID-19: a systematic review and meta-analysis. Lancet. 2020;395(10242):1973-87. https://doi.org/10.1016/S0140-6736(20)31142-9

McGettigan P, Sly K, O'Connell D, Hill S, Henry D. The effects of information framing on the practices of physicians. J Gen Intern Med. 1999;14(10):633-42. https://doi.org/10.1046/j.1525-1497.1999.09038.x

Moxey A, O'Connell D, McGettigan P, Henry D. Describing treatment effects to patients. J Gen Intern Med. 2003;18(11):948-59. https://doi.org/10.1046/j.1525-1497.2003.20928.x

Talic S, Shah S, Wild H, Gasevic D, Maharaj A, Ademi Z, et al. Effectiveness of public health measures in reducing the incidence of covid-19, SARS-CoV-2 transmission, and covid-19 mortality: systematic review and meta-analysis. BMJ. 2021;375:e068302. https://doi.org/10.1136/bmj-2021-068302

Vestrheim DF, Iversen BG, Flottorp SA, Denison EM-L, Oxman AD. Should individuals in the community without respiratory symptoms wear facemasks to reduce the spread of Covid-19?–Update 1. Oslo, Norway: Norwegian Institute of Public Health; 2020. https://hdl.handle.net/11250/2722757

Visschers VH, Meertens RM, Passchier WW, de Vries NN. Probability information in risk communication: a review of the research literature. Risk Anal. 2009;29(2):267-87. https://doi.org/10.1111/j.1539-6924.2008.01137.x

Zipkin DA, Umscheid CA, Keating NL, Allen E, Aung K, Beyth R, et al. Evidence-based risk communication: a systematic review. Ann Intern Med. 2014;161(4):270-80. https://doi.org/10.7326/m14-0295

Research studies

Deeks JJ. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. Stat Med. 2002;21(11):1575-600. https://doi.org/10.1002/sim.1188

Engels EA, Schmid CH, Terrin N, Olkin I, Lau J. Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. Stat Med. 2000;19(13):1707-28. https://doi.org/10.1002/1097-0258(20000715)19:13%3C1707::aid-sim491%3E3.0.co;2-p

Furukawa TA, Guyatt GH, Griffith LE. Can we individualize the 'number needed to treat'? An empirical study of summary effect measures in meta-analyses. Int J Epidemiol. 2002;31(1):72-6. https://doi.org/10.1093/ije/31.1.72

Schmid CH, Lau J, McIntosh MW, Cappelleri JC. An empirical study of the effect of the control rate as a predictor of treatment efficacy in meta-analysis of clinical trials. Stat Med. 1998;17(17):1923-42. https://doi.org/10.1002/(sici)1097-0258(19980915)17:17%3C1923::aid-sim874%3E3.0.co;2-6

Other references

Glasziou PP, Michie S, Fretheim A. Public health measures for covid-19. BMJ. 2021;375:n2729. https://doi.org/10.1136/bmj.n2729

Guyatt GH, Oxman AD, Santesso N, Helfand M, Vist G, Kunz R, et al. GRADE guidelines: 12. Preparing summary of findings tables-binary outcomes. J Clin Epidemiol. 2013a;66(2):158-72. https://doi.org/10.1016/j.jclinepi.2012.01.012

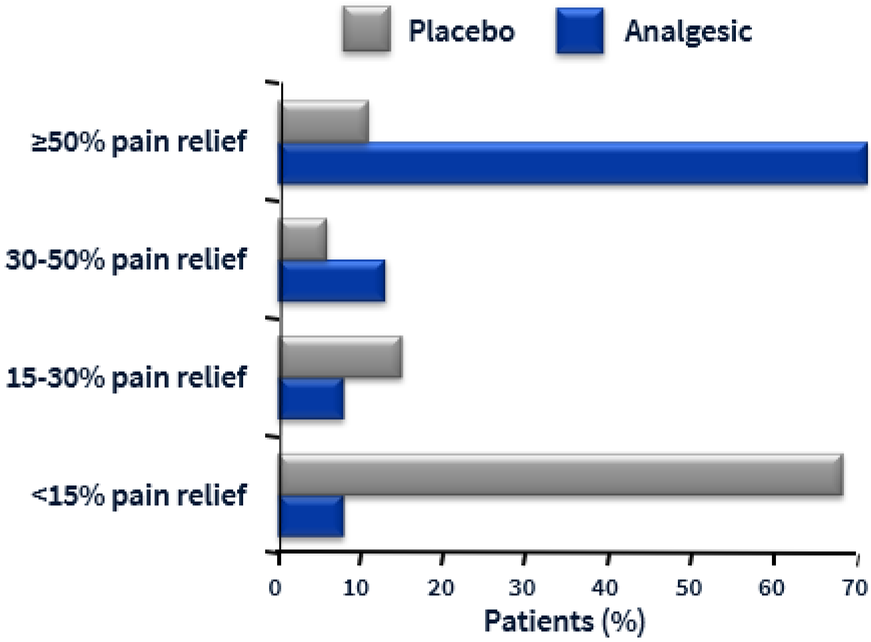
Schünemann HJ, Akl EA, Chou R, Chu DK, Loeb M, Lotfi T, et al. Use of facemasks during the COVID-19 pandemic. Lancet Respir Med. 2020;8(10):954-5. https://doi.org/10.1016/s2213-2600(20)30352-0

2.3c Be cautious of average differences between treatments.

Explanation

Average effects do not apply to everyone. For outcomes that are assessed using scales (for example, to measure weight, or pain) the difference between the average among people in one treatment group and the average among those in a comparison group may not make it clear how many people experienced a big enough change (for example, in weight or pain) for them to notice it, or that they would regard as important. In addition, many scales are difficult to interpret and are reported in ways that make them meaningless. This includes not reporting the lower and upper 'anchor', for example, whether a scale goes from 1 to 10 or 1 to 100; whether higher numbers are good or bad; and whether someone experiencing an improvement of, say, 5 on the scale would barely notice the difference, would consider it a meaningful improvement, or would consider it a large improvement.

For example, the average difference in pain relief is not only hard to interpret, but misleading. When asked what they would consider treatment success, patients with chronic pain specify a large reduction in pain intensity, by 50% or more *[Moore 2013 (OR)]*. Most people tend to respond to painkillers (or a placebo) in two ways. Some people experience a very good pain relief (50% or more), whereas others experience very little (less than 15%). So, the average pain relief does not reflect what most people experienced in randomized trials of painkillers (analgesics) compared to placebos *[Moore 2013 (OR)]*. In the illustration below, the average difference in pain relief is about 28%. A less misleading and easier to understand way of reporting those results would be the difference between the proportion of participants in the analgesic group and the placebo group who were treated successfully (with ≥50% pain relief). In the illustration below, about 60% more participants were treated successfully with the analgesic compared to placebo.



Basis for this concept

Even if the average difference between a treatment and "no treatment" or a comparison treatment is appreciably less than the smallest change that is important to people, treatment may have an important impact on many people *[Guyatt 1998]*. For example, for some quality-of-life questionnaires, it has been shown that the smallest change that is important to people on a seven-point scale is 0.5. Even if the mean difference between a treatment and a comparison treatment is much less than 0.5, the treatment may have important impacts (change greater than 0.5) on many patients.

Outcomes assessed using scales ("continuous outcomes"), such as pain or quality of life, are easily misinterpreted and it is often difficult to make sense of them, especially when different scales are used in different studies *[Guyatt 2013b , Mayer 2019 (OR)]*.

It is possible to convert continuous outcomes to yes/no outcomes (dichotomous outcomes). This makes it easier to interpret the results, and several methods for doing this have been validated by comparing the results of these conversions and dichotomous outcomes measured in the same trials *[da Costa 2012 (SR), Meister 2015 (SR)]*. However, these methods have several limitations *[Guyatt 2013b]*. They can sometimes be misleading when different studies have used different scales, and they may underestimate or overestimate effects when the comparison group's chance of achieving an important change was ≤20% or >60%, respectively *[da Costa 2012 (SR)]*. There are several other ways of presenting the effects of treatments that have been measured using a scale, all of which have limitations *[Guyatt 2013b]*. Therefore, using more than one presentation is likely to be both informative and, if the message is similar, reassuring. It can also reduce the risk of biased selection of which presentation to use when the messages are different. If the messages are different, and it is not clear which to believe, the treatment effect is less certain.

Implications

When outcomes are assessed using scales, it cannot be assumed that every individual in the treatment comparison groups experienced the average effect. Be wary of differences on scales that are not explained or easily understood.

References

Systematic reviews

da Costa BR, Rutjes AW, Johnston BC, Reichenbach S, Nüesch E, Tonia T, et al. Methods to convert continuous outcomes into odds ratios of treatment response and numbers needed to treat: meta-epidemiological study. Int J Epidemiol. 2012;41(5):1445-59. https://doi.org/10.1093/ije/dys124

Meister R, von Wolff A, Kriston L. Odds ratios of treatment response were well approximated from continuous rating scale scores for meta-analysis. J Clin Epidemiol. 2015;68(7):740-51. https://doi.org/10.1016/j.jclinepi.2015.02.006

Other reviews

Mayer M. Continuous outcome measures: conundrums and conversions contributing to clinical application. BMJ Evid Based Med. 2019;24(4):133-6. https://doi.org/10.1136/bmjebm-2018-111136

Moore RA, Straube S, Aldington D. Pain measures and cut-offs - 'no worse than mild pain' as a simple, universal outcome. Anaesthesia. 2013;68(4):400-12. https://doi.org/10.1111/anae.12148

Other references

Guyatt GH, Juniper EF, Walter SD, Griffith LE, Goldstein RS. Interpreting treatment effects in randomised trials. BMJ. 1998;316(7132):690-3. https://doi.org/10.1136/bmj.316.7132.690

Guyatt GH, Thorlund K, Oxman AD, Walter SD, Patrick D, Furukawa TA, et al. GRADE guidelines: 13. Preparing summary of findings tables and evidence profiles-continuous outcomes. J Clin Epidemiol. 2013b;66(2):173-83. https://doi.org/10.1016/j.jclinepi.2012.08.001

2.3d Be cautious of lack of evidence being interpreted as evidence of "no difference".

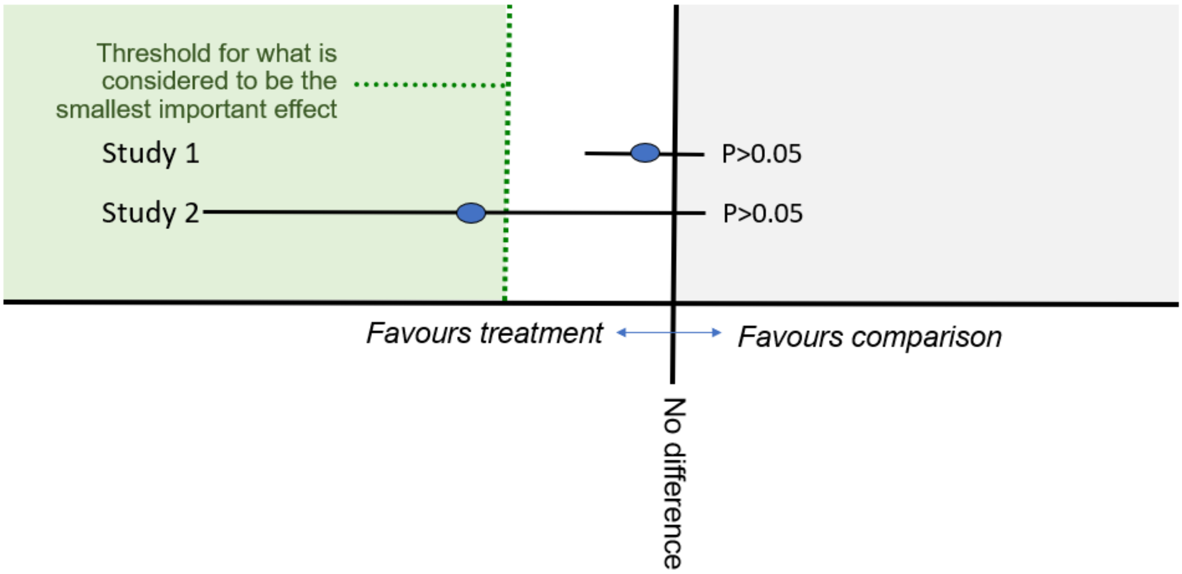
Explanation

Systematic reviews sometimes conclude that there is "no difference" between the treatments compared. However, studies can never show that there is "no difference" ("no effect"). They can only rule out, with specific degrees of confidence, differences of a specific size.

Misinterpreting "statistically non-significant" results and failing to recognise uncertainty in estimates of effect can sometimes impede further research to reduce the uncertainty and result in delays in the uptake of effective treatments. For example, a systematic review of randomized trials of thrombolytic therapy (medicine that prevents blood clots from growing) given to patients after an acute heart attack found a 22% relative reduction in mortality that was highly unlikely to have occurred by chance alone *[Yusuf 1985 (SR)]*. But only five of the 24 trials had shown a "statistically significant" effect (P<0.05). The lack of "statistical significance" of most of the individual trials and misinterpretation of those results led to a long delay before the value of thrombolytic therapy was appreciated.

Basis for this concept

By convention, a 5% probability that the results observed in a treatment comparison could have occurred by the play of chance (P>0.05) is considered "not significant" *[Altman 1995]*. Trials with "statistically non-significant" results are commonly referred to as "negative". But this is misleading. Often those studies are not big enough to either rule in or rule out an important difference (effect) *[Freiman 1978 (RS)]*. This is illustrated in the figure below.



The blue dots in this figure indicate the estimated effect for each study. The horizontal lines indicate the 95% confidence intervals. The dotted green vertical line indicates the smallest effect considered to be important. The results for Study 1 in the figure indicate that an important impact is highly unlikely. It clearly rules out the likelihood of an effect that large or larger. On the other hand, the results for Study 2 are inconclusive. It clearly does not rule out an important effect. The results for both studies are "statistically nonsignificant" (P>0.05), but the interpretation of the two studies should be quite different. The first study was big enough to rule out an important difference. The second study was not. It is inconclusive, not "negative".

A survey of systematic reviews published in 2001-2002 found unqualified claims of "no difference" or "no effect" in 21% of review abstracts (summaries) *[Alderson 2003 (RS)]*. In 2017, such claims were found in 6%

to 8% of systematic reviews *[Marson Smith 2021 (RS)]*. This may indicate greater awareness of the problem. However, the survey found 71 examples of misleading interpretations. These included, for example, “evidence for no effect”, “does not affect”, and “found no beneficial or harmful effects”. This suggests that there is still a problem with misinterpreting lack of evidence as “no difference”. A survey of press releases and associated media coverage in 2010 found misleading claims of “equivalence” in 7% of the abstracts of randomized trials that were the basis for the press release *[Yavchitz 2012 (RS)]*. Those misinterpretations were reflected in the press releases and related news reports. A survey of abstracts of randomized trials published in four high-profile journals in 2016-2017 found that 54% of the authors concluded that there was no treatment benefit, 12% that there was “no significant benefit”, and 13% that there was “no significant difference *[Gates 2019 (RS)]*. Only 3% referred to uncertainty when drawing conclusions. The authors of that survey concluded: “Despite many years of warnings, inappropriate interpretations of [randomized trial] results are widespread in the most prestigious medical journals.”

Considering the precision of effect estimates when making judgements about the certainty of the evidence, and not reporting effects as “significant” or “non-significant” can reduce the chances of being misled *[Altman 1995]*.

Implications

Don’t be misled by statements of “no difference” between treatments (“no effect”). Consider instead the degree to which it is possible to confidently rule out a difference of a specified size.

References

Systematic reviews

Yusuf S, Collins R, Peto R, Furberg C, Stampfer MJ, Goldhaber SZ, et al. Intravenous and intracoronary fibrinolytic therapy in acute myocardial infarction: overview of results on mortality, reinfarction and side-effects from 33 randomized controlled trials. Eur Heart J. 1985;6(7):556-85. https://doi.org/10.1093/oxfordjournals.eurheartj.a061905

Research studies

Alderson P, Chalmers I. Survey of claims of no effect in abstracts of Cochrane reviews. BMJ. 2003;326(7387):475. https://doi.org/10.1136/bmj.326.7387.475

Freiman JA, Chalmers TC, Smith H, Jr., Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71 "negative" trials. N Engl J Med. 1978;299(13):690-4. https://doi.org/10.1056/nejm197809282991304

Gates S, Ealing E. Reporting and interpretation of results from clinical trials that did not claim a treatment difference: survey of four general medical journals. BMJ Open. 2019;9(9):e024785. https://doi.org/10.1136/bmjopen-2018-024785

Marson Smith PR, Ware L, Adams C, Chalmers I. Claims of 'no difference' or 'no effect' in Cochrane and other systematic reviews. BMJ Evid Based Med. 2021;26(3):118-20. https://doi.org/10.1136/bmjebm-2019-111257

Yavchitz A, Boutron I, Bafeta A, Marroun I, Charles P, Mantz J, et al. Misrepresentation of randomized controlled trials in press releases and news coverage: a cohort study. PLoS Med. 2012;9(9):e1001308. https://doi.org/10.1371/journal.pmed.1001308

Other references

Altman DG, Bland JM. Absence of evidence is not evidence of absence. BMJ. 1995;311(7003):485. https://doi.org/10.1136/bmj.311.7003.485

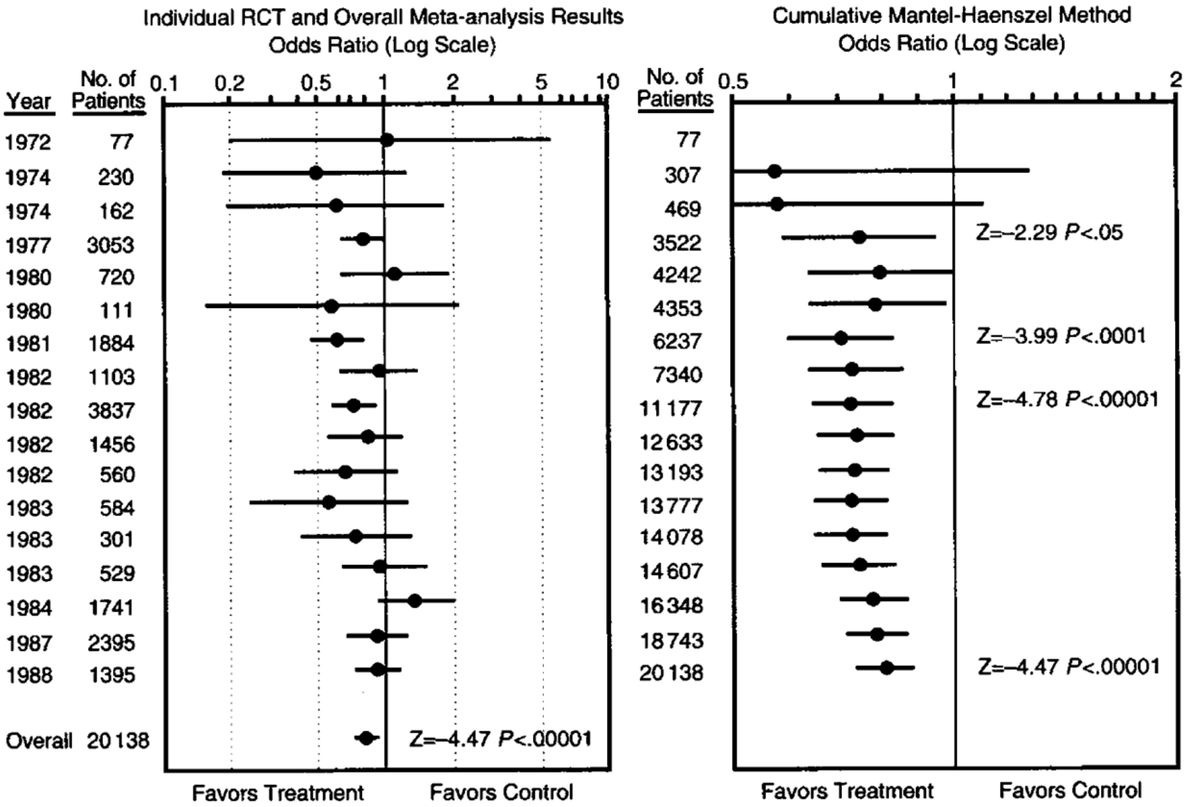
2.4 Descriptions of effects should clearly reflect *the risk of being misled by the play of chance*.

2.4a Be cautious of small studies.

Explanation

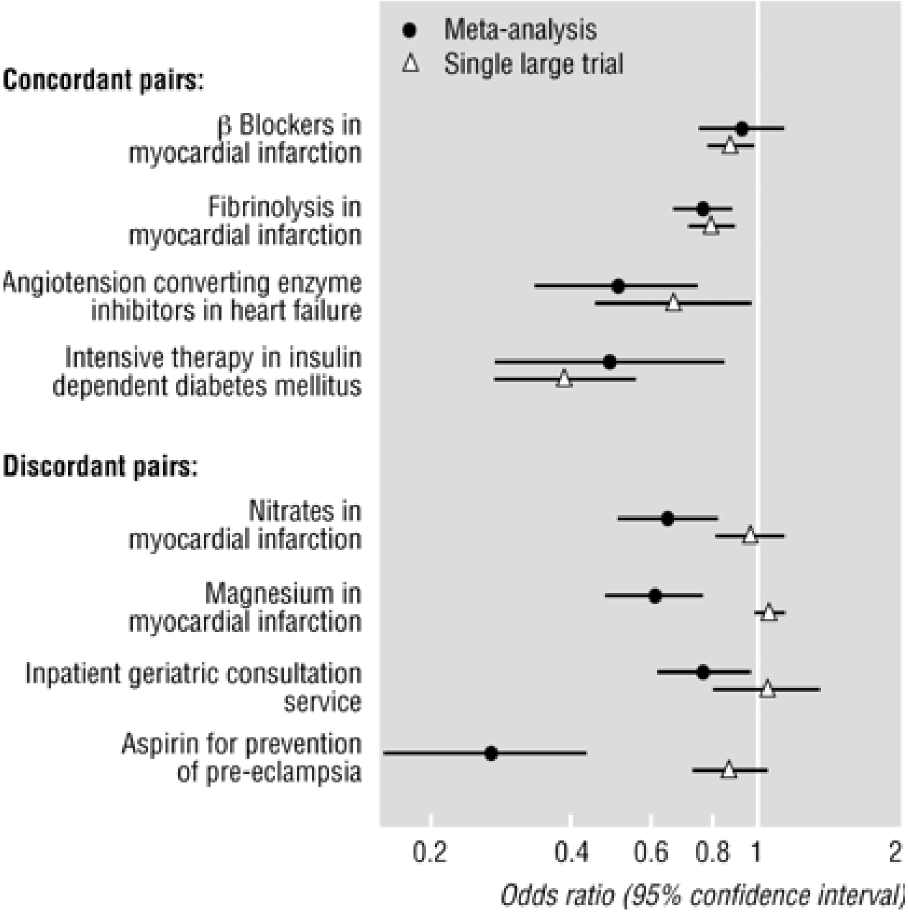
When there are few outcome events, differences in outcome frequencies between the treatment comparison groups may easily have occurred by chance and may mistakenly be attributed to differences in the effects of the treatments, or the lack of a difference.

For example, by 1977 there were at least four randomized trials that compared the number of deaths in patients given a beta-blocker to patients given a placebo. Beta-blockers are medicines that work by blocking the effects of epinephrine (also known as adrenaline). There was a small number of deaths in each study and the results appeared to be inconsistent, as can be seen in the figure on the left below *[Antman 1992 (SR)]*. The results of individual studies continued to vary up until 1988. However, as can be seen in the figure on the right below, if the results of the available studies were combined, the overall estimate (across studies) changed very little after 1977. It simply became more precise. This is indicated by the horizontal lines, which show the confidence intervals for each effect estimate.



In the example above, the variation in effect estimates may have occurred largely by chance alone. The overall effect estimate across the small studies was consistent with the results of a large randomized trial with a low risk of bias published in 1986 *[Egger 1997]*. However, effect estimates from small studies may overestimate actual effects. There are several possible reasons for this. Compared to large studies, small studies may be more prone to publication bias and reporting bias, may have a higher risk of bias because of the design of the studies. Small studies also may include more highly selected participants and may implement treatments more uniformly.

For example, in some countries, intravenous (IV) magnesium was administered to heart attack patients to limit damage to the heart muscle, prevent serious arrhythmias and reduce the risk of death. A controversy erupted in 1995, when a large well-designed trial with 58,050 participants did not demonstrate any beneficial effect to IV magnesium, contradicting earlier meta-analyses of the smaller trials. The figure below shows four examples where the results of small trials were consistent with the results of a single large trial (concordant pairs) and four examples where they were not consistent (discordant pairs), including IV magnesium for acute heart attacks *[Egger 1997]*.



It is difficult to predict when or why effect estimates from small studies will differ from effect estimates from large studies with a low risk of bias or to be certain about the reasons for differences. However, systematic reviews should consider the risk of small studies being biased towards larger effects and consider potential reasons for bias in effect estimates from small studies. A systematic review published in 2007 included 26 randomized trials that compared IV magnesium to an inactive substance (placebo) *[Li 2007 (SR)]*. IV magnesium reduced the incidence of serious arrhythmias, but also increased the incidence of profound hypotension, bradycardia and flushing. The apparent large effect of magnesium on reducing the number of deaths may have reflected various biases in smaller trials.

Basis for this concept

A systematic review of 93 meta-analyses found that effect estimates differed within meta-analyses based on their size, with larger effect estimates seen in small to moderately sized trials compared to the largest trials *[Dechartres 2013 (SR)]*. Another systematic review found that smaller studies reported larger effects in 19% of the 5,534 meta-analyses with ≥10 studies *[Schwab 2021 (SR)]*. Only 4% of those meta-analyses showed evidence of publication bias. The extent to which small studies reported larger effects varied across medical specialities. Other systematic reviews have also found that small studies sometimes overestimate treatment effects *[Ioannidis 2007 (SR), Lin 2020 (SR), Nüesch 2010 (SR)]*.

There are several reasons why small studies may overestimate treatment effects *[Egger 1997 , Schwab 2021 (SR)]*. One reason is reporting: small studies may be more prone to publication bias and selective outcome reporting (reporting bias). Another reason is that small studies may have a higher risk of bias due to their design and implementation compared to large studies. Also, large studies may include more diverse patients and implementation of treatments (making them more "pragmatic") compared to small studies (which may be more "explanatory"). It may be difficult to detect when small studies overestimate treatment effects [Ioannidis 2007 (SR)], and to detect the reasons for them overestimating or potentially overestimating effects.

Implications

Be cautious about relying on the results of treatment comparisons with few outcome events. The results of such comparisons can be misleading.

References

Systematic reviews

Antman EM, Lau J, Kupelnick B, Mosteller F, Chalmers TC. A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts. Treatments for myocardial infarction. JAMA. 1992;268(2):240-8. https://doi.org/10.1001/jama.1992.03490020088036

Dechartres A, Trinquart L, Boutron I, Ravaud P. Influence of trial sample size on treatment effect estimates: meta-epidemiological study. BMJ. 2013;346:f2304. https://www.bmj.com/content/bmj/346/bmj.f2304.full.pdf

Ioannidis JP, Trikalinos TA. The appropriateness of asymmetry tests for publication bias in meta-analyses: a large survey. CMAJ. 2007;176(8):1091-6. https://doi.org/10.1503/cmaj.060410

Li J, Zhang Q, Zhang M, Egger M. Intravenous magnesium for acute myocardial infarction. Cochrane Database Syst Rev. 2007;2007(2):Cd002755. https://doi.org/10.1002/14651858.cd002755.pub2

Lin L, Shi L, Chu H, Murad MH. The magnitude of small-study effects in the Cochrane Database of Systematic Reviews: an empirical study of nearly 30 000 meta-analyses. BMJ Evid Based Med. 2020;25(1):27-32. https://doi.org/10.1136/bmjebm-2019-111191

Nüesch E, Trelle S, Reichenbach S, Rutjes AW, Tschannen B, Altman DG, et al. Small study effects in meta-analyses of osteoarthritis trials: meta-epidemiological study. BMJ. 2010;341:c3515. https://doi.org/10.1136/bmj.c3515

Schwab S, Kreiliger G, Held L. Assessing treatment effects and publication bias across different specialties in medicine: a meta-epidemiological study. BMJ Open. 2021;11(9):e045942. https://doi.org/10.1136/bmjopen-2020-045942

Other references

Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ. 1997;315(7109):629-34. https://doi.org/10.1136/bmj.315.7109.629

2.4b Be cautious of results for a selected group of people within a study.

Explanation

Average effects do not apply to everyone. However, comparisons of treatments often report results for selected groups of participants to assess whether the effect of a treatment is different for different types of people (e.g., men and women or different age groups). These analyses are often poorly planned and reported. Most differential effects suggested by these "subgroup" analyses are likely to be due to the play of chance and are unlikely to reflect true treatment differences.

For example, in 1983 the authors of a paper that presented 146 subgroup analyses of the Beta Blocker Heart Attack trial, found that the results were normally distributed – a pattern that would be expected if the variation in results was simply due to the play of chance *[Oxman 2012b]*. Roughly 2.5% of the subgroup analyses had results that statistically were "significantly" worse and 2.5% had results that were "significantly" better. Five years later the International Study of Infarct Survival 2 (ISIS-2) trial found that aspirin reduced mortality after heart attack overall (P<0.00001) but increased mortality by a small amount in patients born under Gemini and Libra astrological signs. The authors included this subgroup analysis in their report to illustrate the likelihood of misleading subgroup analyses. Six years after that, the DICE (Don't Ignore Chance Effects) collaborators in their meta-analysis of trials of DICE therapy (rolling dice) for acute stroke found that red dice are deadly, based on a predefined subgroup analysis by colour of dice. All these findings illustrate the important message that chance influences the results of treatment comparisons and systematic reviews. Unfortunately, researchers, health professionals, patients and the public continue to be misled by subgroup analyses.

Basis for this concept

Reviews of published randomized trials and protocols have found that subgroup analyses are commonly reported (38-87% of the time), and that appropriate statistical analyses (tests of interaction) are not used 38-91% of the time *[Oxman 2012b]*. In addition, planned subgroup analyses are commonly not reported (48-69% of the time) and 43-91% of randomized trials report subgroup analyses that were not planned. When subgroup analyses are reported, authors claim differences in 25-69% of cases, and these results are commonly featured prominently (15-45% of the time).

A systematic review of randomized trials published in core journals in 2007 found that 44% of the trials reported subgroup analyses *[Sun 2012 (SR)]*. The review authors assessed the credibility of the subgroup claims using explicit criteria. They found that the credibility of most of subgroup claims, including strong claims, was usually low. Subsequent systematic reviews have found that inadequate specification and reporting of subgroup analyses remain problematic in protocols and reports of randomized controlled trials *[Fan 2019 (SR), Gabler 2016 (SR), Kasenda 2014 (RS), Wallach 2017 (SR)]*. Justifications or rationales for subgroup analyses were only rarely provided in trial protocols and reports, and large discrepancies were found between planning of subgroup analyses in protocols and their reporting in publications of randomized trials. A systematic review of subgroup analyses based on sex found that "statistically significant" sex-treatment interactions were only slightly more frequent than would be expected by chance *[Wallach 2016 (SR)]*.

Several different checklists have been developed that can help to assess the credibility of claims about subgroup effects *[Gil-Sierra 2020 , Oxman 2002 , Oxman 1992 , Sun 2014]*.

Implications

Findings based on results for subgroups of people within treatment comparisons may be misleading.

References

Systematic reviews

Fan J, Song F, Bachmann MO. Justification and reporting of subgroup analyses were lacking or inadequate in randomized controlled trials. J Clin Epidemiol. 2019;108:17-25. https://doi.org/10.1016/j.jclinepi.2018.12.009

Gabler NB, Duan N, Raneses E, Suttner L, Ciarametaro M, Cooney E, et al. No improvement in the reporting of clinical trial subgroup effects in high-impact general medical journals. Trials. 2016;17(1):320. https://doi.org/10.1186/s13063-016-1447-5

Sun X, Briel M, Busse JW, You JJ, Akl EA, Mejza F, et al. Credibility of claims of subgroup effects in randomised controlled trials: systematic review. BMJ. 2012;344:e1553. https://doi.org/10.1136/bmj.e1553

Wallach JD, Sullivan PG, Trepanowski JF, Sainani KL, Steyerberg EW, Ioannidis JP. Evaluation of evidence of statistical support and corroboration of subgroup claims in randomized clinical trials. JAMA Intern Med. 2017;177(4):554-60. https://doi.org/10.1001/jamainternmed.2016.9125

Wallach JD, Sullivan PG, Trepanowski JF, Steyerberg EW, Ioannidis JP. Sex based subgroup differences in randomized controlled trials: empirical evidence from Cochrane meta-analyses. BMJ. 2016;355:i5826. https://doi.org/10.1136/bmj.i5826

Research studies

Kasenda B, Schandelmaier S, Sun X, von Elm E, You J, Blümle A, et al. Subgroup analyses in randomised controlled trials: cohort study on trial protocols and journal publications. BMJ. 2014;349:g4539. https://doi.org/10.1136/bmj.g4539

Other references

Gil-Sierra MD, Fénix-Caballero S, Abdel Kader-Martin L, Fraga-Fuentes MD, Sánchez-Hidalgo M, Alarcón de la Lastra-Romero C, et al. Checklist for clinical applicability of subgroup analysis. J Clin Pharm Ther. 2020;45(3):530-8. https://doi.org/10.1111/jcpt.13102

Oxman AD. Subgroup analyses. BMJ. 2012b;344:e2022. https://doi.org/10.1136/bmj.e2022

Oxman AD, Guyatt G. When to believe a subgroup analysis. In: Guyatt G, Rennie D, editors. Users' Guide to the Medical Literature A Manual for Evidence-Based Clinical Practice. Chicago: AMA Press; 2002. p. 553-65.

Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. Ann Intern Med. 1992;116(1):78-84. https://doi.org/10.7326/0003-4819-116-1-78

Sun X, Ioannidis JP, Agoritsas T, Alba AC, Guyatt G. How to use a subgroup analysis: users' guide to the medical literature. JAMA. 2014;311(4):405-11. https://doi.org/10.1001/jama.2013.285063

2.4c Be cautious of p-values.

Explanation

The observed difference in outcomes is the best estimate of how relatively effective and safe treatments are (or would be, if the comparison were made in many more people). However, because of the play of chance, the true difference may be larger or smaller than this. The confidence interval is the range within which the true difference is likely to lie, after considering the play of chance. Although a confidence interval (margin of error) is more informative than a p-value, often only the latter is reported. P-values are often misinterpreted to mean that treatments have or do not have important effects.

For example, George Siontis and John Ioannidis reviewed 51 articles that reported "statistically significant tiny effects" published in four high profile journals *[Siontis 2011 (SR)]*. Even minimal bias in those studies could explain the observed "effects". Yet, more than half (28) of the articles did not express any concern about the size or uncertainty of the estimate of the observed effect. Despite the low p-values reported in these articles, the results often excluded effects that would be large enough to be important. Interpretation of small effects based on p-values alone is likely to be misleading.

Basis for this concept

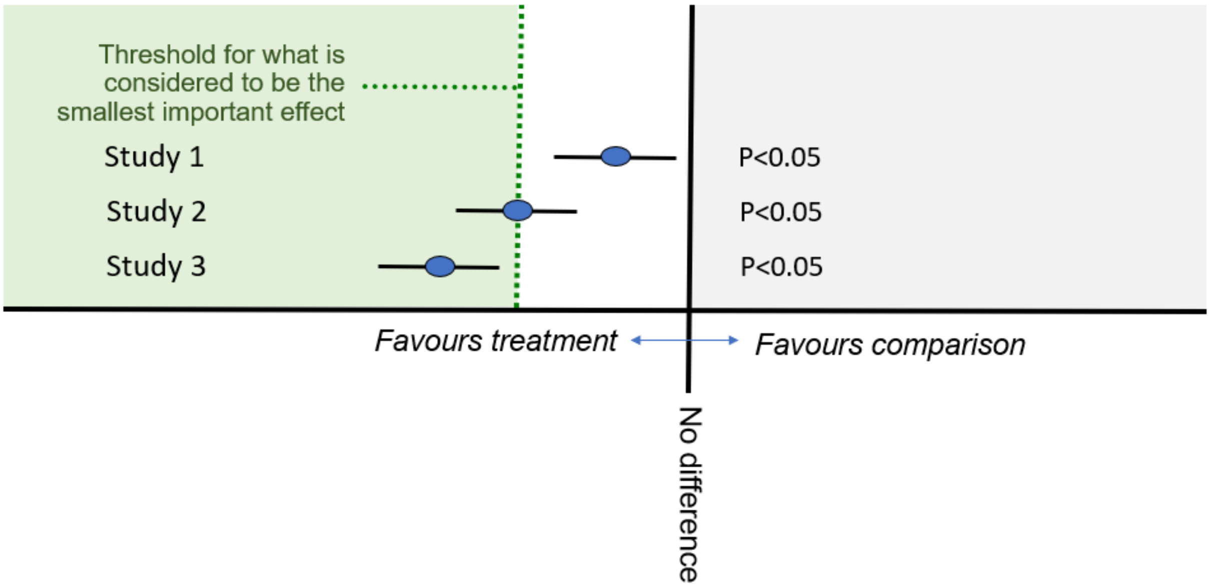
P-values, or "significance" levels, measure the probability of observing a result as extreme or more extreme than the actual result, simply by chance, if, in reality, there is no treatment difference. The smaller the p-value the less likely it is that there is no difference. Hundreds of warnings of the limitations of p-values and significance testing have been published *[Stang 2017 (SR)]*. From the 1970s to 2014, the proportion of abstracts (summaries of studies) with significance testing without any confidence intervals decreased from close to 100% to below 25%. However, the proportion of abstracts reporting only confidence intervals (and not p-values) in the top medical journals was only 22%. Another systematic review of abstracts indexed in MEDLINE found that more abstracts and articles reported p-values over time between 1990 and 2015 *[Chavalarias 2016 (SR)]*. Almost all abstracts and articles with p-values reported "statistically significant" results. Confidence intervals were only reported in about 2% of abstracts.

Despite all the warnings about p-values and significance testing, use and misinterpretation of p-values continues to be a problem. In 2016, the American Statistical Association released a policy statement on statistical significance and p-values, which included this warning *[Wasserstein 2016]*: ''The widespread use of 'statistical significance' (generally interpreted as '$p \leq 0.05$') as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process.''

A systematic review of abstracts describing the results of cancer randomized trials with p-values between 0.01 and 0.10 found that trials commonly failed to convey uncertainty when describing results of "marginal statistical significance" *[Rubinstein 2019 (SR)]*. The results were often conveyed as definitively demonstrating that the null hypothesis (no difference) was false. This is likely associated with a discrete threshold for "statistical significance" (generally 0.05).

Another systematic review of surgical randomized trials found that outcomes reported in the abstract had three times the odds of being "statistically significant" compared to the corresponding full text *[Assem 2017 (SR)]*. Biased reporting of outcomes in abstracts based on p-values being below an arbitrary threshold has been found in other studies *[Boutron 2010 (SR), Chavalarias 2016 (SR), Ginsel 2015 (SR), Gøtzsche 2006 (SR)]*. This problem is like problems with publication bias and selective outcome reporting (see Concept 2.2b).

P-values can be misinterpreted in several ways *[Goodman 2008 , Greenland 2016]*. Perhaps most importantly, "statistical significance" may be confused with importance, and the cut-off for considering a result as statistically significant (generally $p \leq 0.05$) is arbitrary (see Concept 2.4d). People often assume that a low p-value indicates an important effect. However, a low p-value may or may not indicate an important effect, as illustrated in the figure below. All three studies have a p-value less than 0.05, indicating that it is unlikely that the observed treatment difference could have occurred simply by chance. But Study 1 indicates that it is unlikely that the difference was important, Study 2 indicates it is uncertain whether there was an important difference, and Study 3 indicates it is likely there was an important difference.



The blue dots in the figure above indicate the observed treatment effect and the horizontal lines indicate the confidence interval for each effect estimate. The figure illustrates why confidence intervals are more informative than p-values, as well as why the results of treatment comparisons should be interpreted in relation to thresholds for what is considered to be an important effect, not in relation to no difference.

Another problem with p-values is that people may assume that a p-value is the probability that there is no treatment difference and that a high p-value indicates a high probability that there is not a difference *[Sterne 2001]*. However, p-values indicate the probability of a "type I error" (assuming there is a difference when in fact there is not). They do not indicate the probability of a "type II error" (assuming there is not a difference when in fact there is). Many studies are too small to rule out an important difference (see Concept 2.3d).

Furthermore, people may assume that a low p-value indicates the likelihood that the observed treatment effect is the "true" effect. However, p-values only indicate the probability of wrongly assuming there is a difference when the observed difference could have occurred simply by chance ("random error"). It does not indicate anything about the risk of bias (systematic errors) because of how studies are designed, analysed, or reported (see Concepts 2.1a-2.1g).

P-values are used for testing the "null hypothesis" (that there is not a difference). A low p-value indicates that the null hypothesis can be rejected, with respect to random error. But hypothesis testing is unhelpful for people deciding whether to use a treatment. Hypothesis testing implies that there is a simple yes or no answer (there is or is not an effect) and it does not convey any information about the size of the effect *[Gardner 1986]*. Estimation of the size of the effect – for example, how big the difference is – and the confidence interval for that estimate is much more informative and less likely to mislead people.

However, it should be noted that confidence intervals are also sometimes misinterpreted *[Greenland 2016]*. In addition, 95 % confidence intervals correspond to a 0.05 cut-off for p-values. Thus, they have some of the same shortcomings as p-values. Nonetheless, confidence intervals are preferable to "significance" tests and p-values because they shift the focus away from the null hypothesis, toward the range of effect estimates compatible with the data. Provided they are interpreted carefully, they can also shift the focus from any difference greater than zero to effects that are large enough to be important *[Zeng 2021]*.

Implications

Understanding a confidence interval may be necessary to understand the reliability of estimates of treatment effects. Whenever possible, consider confidence intervals when assessing estimates of treatment effects. Do not be misled by p-values.

References

Systematic reviews

Assem Y, Adie S, Tang J, Harris IA. The over-representation of significant p values in abstracts compared to corresponding full texts: A systematic review of surgical randomized trials. Contemp Clin Trials Commun. 2017;7:194-9. https://doi.org/10.1016/j.conctc.2017.07.007

Boutron I, Dutton S, Ravaud P, Altman DG. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. JAMA. 2010;303(20):2058-64. https://doi.org/10.1001/jama.2010.651

Chavalarias D, Wallach JD, Li AH, Ioannidis JP. Evolution of Reporting P Values in the Biomedical Literature, 1990-2015. JAMA. 2016;315(11):1141-8. https://doi.org/10.1001/jama.2016.1952

Ginsel B, Aggarwal A, Xuan W, Harris I. The distribution of probability values in medical abstracts: an observational study. BMC Res Notes. 2015;8:721. https://doi.org/10.1186/s13104-015-1691-x

Gøtzsche PC. Believability of relative risks and odds ratios in abstracts: cross sectional study. BMJ. 2006;333(7561):231-4. https://doi.org/10.1136/bmj.38895.410451.79

Rubinstein SM, Sigworth EA, Etemad S, Martin RL, Chen Q, Warner JL. Indication of measures of uncertainty for statistical significance in abstracts of published oncology trials: a systematic review and meta-analysis. JAMA Netw Open. 2019;2(12):e1917530. https://doi.org/10.1001/jamanetworkopen.2019.17530

Siontis GC, Ioannidis JP. Risk factors and interventions with statistically significant tiny effects. Int J Epidemiol. 2011;40(5):1292-307. https://doi.org/10.1093/ije/dyr099

Stang A, Deckert M, Poole C, Rothman KJ. Statistical inference in abstracts of major medical and epidemiology journals 1975-2014: a systematic review. Eur J Epidemiol. 2017;32(1):21-9. https://doi.org/10.1007/s10654-016-0211-1

Other references

Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. BMJ. 1986;292(6522):746-50. https://doi.org/10.1136/bmj.292.6522.746

Goodman S. A dirty dozen: twelve p-value misconceptions. Semin Hematol. 2008;45(3):135-40. https://doi.org/10.1053/j.seminhematol.2008.04.003

Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol. 2016;31(4):337-50. https://doi.org/10.1007/s10654-016-0149-3

Sterne JA, Davey Smith G. Sifting the evidence-what's wrong with significance tests? BMJ. 2001;322(7280):226-31. https://doi.org/10.1093/ptj/81.8.1464

Wasserstein RL, Lazar NA. The ASA Statement on p-Values: Context, Process, and Purpose. Am Stat. 2016;70(2):129-33. https://doi.org/10.1080/00031305.2016.1154108

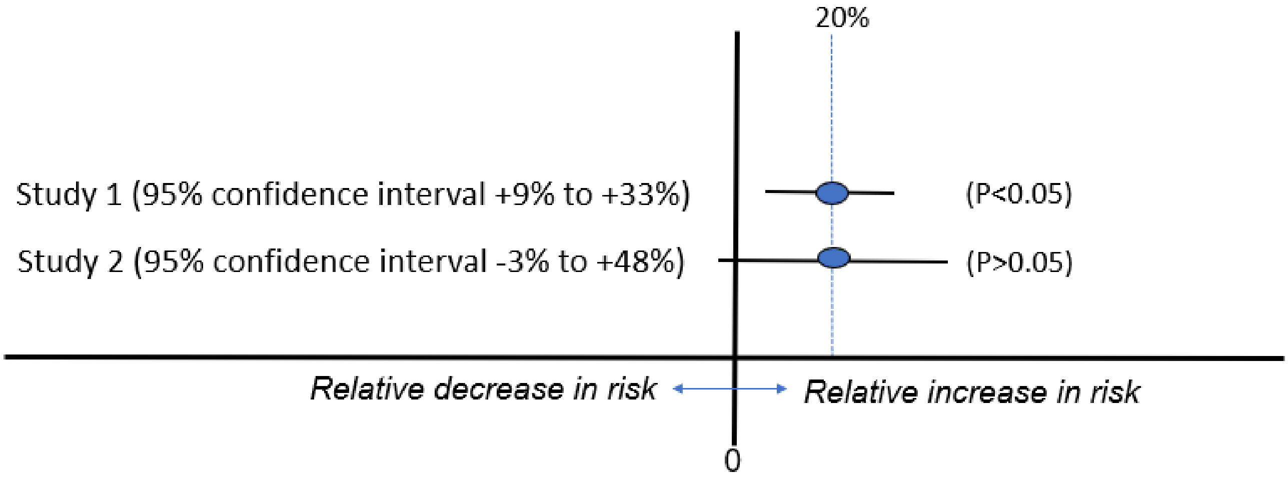
Zeng L, Brignardello-Petersen R, Hultcrantz M, Siemieniuk RAC, Santesso N, Traversy G, et al. GRADE guidelines 32: GRADE offers guidance on choosing targets of GRADE certainty of evidence ratings. J Clin Epidemiol. 2021;137:163-75. https://doi.org/10.1016/j.jclinepi.2021.03.026

2.4d Be cautious of results reported as “statistically significant” or “non-significant”.

Explanation

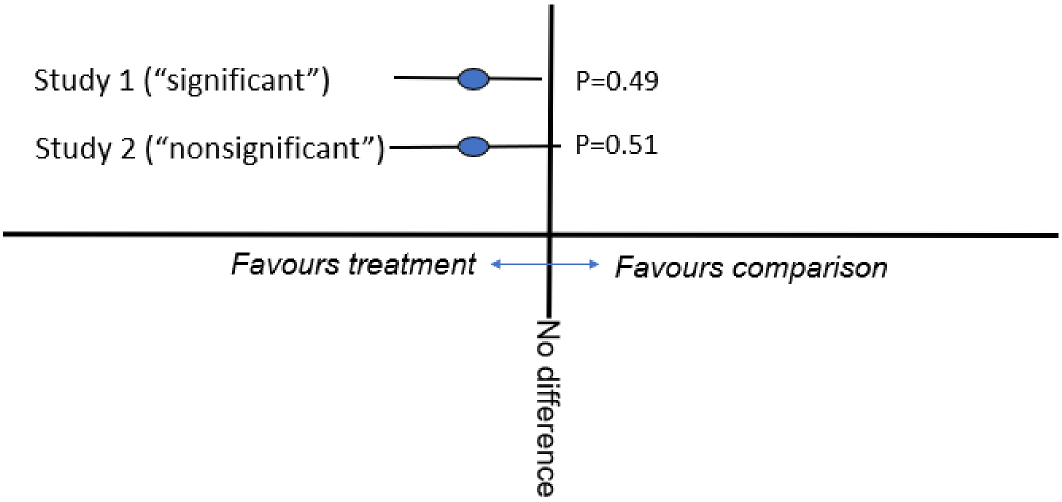
“Statistical significance may be confused with “importance”. The cut-off for considering a result as statistically significant is arbitrary, and statistically non-significant results can be either informative (showing that it is very unlikely that a treatment has an important effect) or inconclusive (showing that the relative effects of the treatments compared are uncertain).

For example, two studies of a possible adverse effect of anti-inflammatory drugs on the risk of heart rhythm abnormalities (atrial fibrillation) were reported as having had “statistically nonsignificant” results *[Schmidt 2014]*. The authors of one of the articles concluded that exposure to the drugs was “not associated” with an increased risk and that the results stood in contrast to those from an earlier study with a “statistically significant” result. However, the effect estimates were the same for the two studies: a risk ratio of 1.2 (that is, a 20% relative increase). The earlier study was simply more precise, as indicated by the narrower confidence interval in the figure below. Concluding that the results of the second study showed “no association” was misleading, considering that the confidence interval ranged from a 3% decrease in risk to a 48% increase. It is also misleading to conclude that the results were in contrast with the earlier study that had an identical observed effect. Yet, misleading interpretations like this, which are based on an arbitrary cut-off for “statistical significance” are common.



Basis for this concept

The arbitrariness of a cut-off for “statistical significance is illustrated in the figure below. The results of the two studies are almost identical. Yet, Study 1 is “statistically significant” and Study 2 is “statistically nonsignificant”.



In addition to being arbitrary, "significant" and "important" are synonyms, and statistical significance is often confused with importance, especially when "significant" is not prefaced by "statistically". Statistical significance does not convey any information about the size of the effect. A "statistically significant" effect may or may not be important. Similarly, an observed effect that is "statistically nonsignificant" may or may not be important, and the results may or may not rule out an important effect (see Concept 2.3d).

Systematic reviews have found that the reporting and interpretation of randomized trials with "statistically nonsignificant" findings was frequently inconsistent with the results and biased *[Boutron 2010 (SR)]*, with some authors supporting treatments despite evidence that they might be ineffective or harmful *[Hewitt 2008 (RS)]*, while over half inappropriately interpreted "statistically nonsignificant" results as indicating no effect *[Freiman 1978 (RS), Gates 2019 (RS)]*. On the other hand, reports of findings that were marginally "statistically significant (p-values between 0.01 and 0.10) commonly failed to convey uncertainty when describing the results and often conveyed them as definitively demonstrating an effect" *[Rubinstein 2019 (SR)]*. In addition, results that are "statistically significant" are more likely to be reported in abstracts compared to the corresponding full text *[Assem 2017 (SR), Boutron 2010 (SR), Chavalarias 2016 (SR), Ginsel 2015 (SR), Gøtzsche 2006 (SR)]*, whereas, "statistically nonsignificant results may not be reported at all" (see Concept 2.2b).

Implications

Claims that results were 'significant' or 'non-significant' usually mean that they were 'statistically significant' or 'statistically non-significant'. This is not the same as 'important' or 'not important'. Do not be misled by such claims.

References

Systematic reviews

Assem Y, Adie S, Tang J, Harris IA. The over-representation of significant p values in abstracts compared to corresponding full texts: A systematic review of surgical randomized trials. Contemp Clin Trials Commun. 2017;7:194-9. https://doi.org/10.1016/j.conctc.2017.07.007

Boutron I, Dutton S, Ravaud P, Altman DG. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. JAMA. 2010;303(20):2058-64. https://doi.org/10.1001/jama.2010.651

Chavalarias D, Wallach JD, Li AH, Ioannidis JP. Evolution of Reporting P Values in the Biomedical Literature, 1990-2015. JAMA. 2016;315(11):1141-8. https://doi.org/10.1001/jama.2016.1952

Ginsel B, Aggarwal A, Xuan W, Harris I. The distribution of probability values in medical abstracts: an observational study. BMC Res Notes. 2015;8:721. https://doi.org/10.1186/s13104-015-1691-x

Gøtzsche PC. Believability of relative risks and odds ratios in abstracts: cross sectional study. BMJ. 2006;333(7561):231-4. https://doi.org/10.1136/bmj.38895.410451.79

Rubinstein SM, Sigworth EA, Etemad S, Martin RL, Chen Q, Warner JL. Indication of measures of uncertainty for statistical significance in abstracts of published oncology trials: a systematic review and meta-analysis. JAMA Netw Open. 2019;2(12):e1917530. https://doi.org/10.1001/jamanetworkopen.2019.17530

Research studies

Freiman JA, Chalmers TC, Smith H, Jr., Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. Survey of 71 "negative" trials. N Engl J Med. 1978;299(13):690-4. https://doi.org/10.1056/nejm197809282991304

Gates S, Ealing E. Reporting and interpretation of results from clinical trials that did not claim a treatment difference: survey of four general medical journals. BMJ Open. 2019;9(9):e024785. https://doi.org/10.1136/bmjopen-2018-024785

Hewitt CE, Mitchell N, Torgerson DJ. Listen to the data when results are not significant. BMJ. 2008;336(7634):23-5. https://doi.org/10.1136/bmj.39379.359560.ad

Other references

Schmidt M, Rothman KJ. Mistaken inference caused by reliance on and misinterpretation of a significance test. Int J Cardiol. 2014;177(3):1089-90. https://doi.org/10.1016/j.ijcard.2014.09.205

3. Choices

What to do depends on judgements about a problem, the relevance of the available evidence, and the balance of expected benefits, harms, and costs.

3.1 Evidence should be relevant.

3.1a Be clear about what the problem or goal is and what the options are.

Explanation

Good decisions depend on correctly identifying the problems and considering an appropriate set of options to address the problems. For personal health choices, this means starting with a correct diagnosis (or assessment of risk) and then identifying the treatments that are available. For public health and health system policy decisions, this means describing the problem correctly and identifying the policy options relevant for that problem. Changing how a problem is framed can lead to different options for addressing it.

The following passage from Archie Cochrane's autobiography is a striking illustration of the importance of correctly identifying the problem and considering appropriate options *[Cochrane 1989]*. He recalls an event when he was a doctor in a German prisoner of war camp. "*The Germans dumped a young Soviet prisoner in my ward late one night. The ward was full, so I put him in my room as he was moribund and screaming and I did not want to wake the ward. I examined him. He had obvious gross bilateral cavitation and a severe pleural rub. I thought the latter was the cause of the pain and the screaming. I had no morphia, just aspirin, which had no effect. I felt desperate. I knew very little Russian then and there was no one in the ward who did. I finally instinctively sat down on the bed and took him in my arms, and the screaming stopped almost at once. He died peacefully in my arms a few hours later. It was not the pleurisy that caused the screaming but loneliness. It was a wonderful education about the care of the dying. I was ashamed of my misdiagnosis and kept the story secret.*"

Basis for this concept

Failure to correctly diagnose health problems can result in inappropriate treatment and unnecessary suffering. For example, studies suggest that from 20% to 70% of people with asthma remain undiagnosed and therefore untreated *[Aaron 2018 (OR)]*. At the same time, 30% to 35% of people diagnosed with asthma do not have asthma and therefore are treated inappropriately. Other examples of conditions that are frequently misdiagnosed include chronic obstructive pulmonary disease (COPD) *[Diab 2018 (OR)]* and heart failure *[Wong 2021 (SR)]*. Other examples of undiagnosed and undertreated conditions include hypertension and HIV *[Glasziou 2017 (OR)]*. Examples of overdiagnosed and overtreated conditions include cancers, bipolar disorder, depression, attention deficit hyperactivity disorder, diabetes, allergic reactions, and infections *[Jenniskens 2017 (SR)]*. Estimates of diagnostic errors (missed, wrong or delayed diagnoses) vary, but may affect between 10% and 15% of hospital admissions and patients with common diseases attending outpatient clinics *[Graber 2013 (OR)]*. A systematic review found that about 0.7% of adults admitted to hospital in the U.S. experienced a diagnostic error that causes them harm *[Gunderson 2020 (SR)]*.

In addition to diagnostic errors, non-medical problems may sometimes be mistakenly treated as medical problems *[Conrad 2010 (RS), Moynihan 2006]*. Similarly, inadequate clarification of public health or health system problems and failing to identify appropriate options for addressing the problem can result in misguided efforts and wasted resources *[Lavis 2009a , Lavis 2009b]*.

Common cognitive biases that can result in diagnostic errors include *[Blumenthal-Barby 2015 (SR), Scott 2020]*:

- Premature closure – narrow rapid focus on single or a few features to support a diagnosis without considering other alternatives
- Anchoring bias – clinging to an initial diagnosis
- Confirmation bias – selectively searching for evidence to support an initial or favoured diagnosis
- Availability bias – overestimating the probability of a diagnosis based on how easily it is recalled

Both underuse and overuse of treatments are common and costly around the world, in part because appropriate options have not been considered or used *[Brownlee 2017 (OR), Elshaug 2017 (OR), Glasziou 2017 (OR), Saini 2017 (OR)]*.

Implications

Make sure you are considering the correct diagnosis or problem, and appropriate options for addressing it.

References

Systematic reviews

Blumenthal-Barby JS, Krieger H. Cognitive biases and heuristics in medical decision making: a critical review using a systematic search strategy. Med Decis Making. 2015;35(4):539-57. https://doi.org/10.1177/0272989x14547740

Gunderson CG, Bilan VP, Holleck JL, Nickerson P, Cherry BM, Chui P, et al. Prevalence of harmful diagnostic errors in hospitalised adults: a systematic review and meta-analysis. BMJ Qual Saf. 2020;29(12):1008-18. https://doi.org/10.1136/bmjqs-2019-010822

Jenniskens K, de Groot JAH, Reitsma JB, Moons KGM, Hooft L, Naaktgeboren CA. Overdiagnosis across medical disciplines: a scoping review. BMJ Open. 2017;7(12):e018448. https://doi.org/10.1136/bmjopen-2017-018448

Wong CW, Tafuro J, Azam Z, Satchithananda D, Duckett S, Barker D, et al. Misdiagnosis of heart failure: a systematic review of the literature. J Card Fail. 2021;27(9):925-33. https://doi.org/10.1016/j.cardfail.2021.05.014

Other reviews

Aaron SD, Boulet LP, Reddel HK, Gershon AS. Underdiagnosis and overdiagnosis of asthma. Am J Respir Crit Care Med. 2018;198(8):1012-20. https://doi.org/10.1164/rccm.201804-0682ci

Brownlee S, Chalkidou K, Doust J, Elshaug AG, Glasziou P, Heath I, et al. Evidence for overuse of medical services around the world. Lancet. 2017;390(10090):156-68. https://doi.org/10.1016/S0140-6736(16)32585-5

Diab N, Gershon AS, Sin DD, Tan WC, Bourbeau J, Boulet LP, et al. Underdiagnosis and overdiagnosis of chronic obstructive pulmonary disease. Am J Respir Crit Care Med. 2018;198(9):1130-9. https://doi.org/10.1164/rccm.201804-0621ci

Elshaug AG, Rosenthal MB, Lavis JN, Brownlee S, Schmidt H, Nagpal S, et al. Levers for addressing medical underuse and overuse: achieving high-value health care. Lancet. 2017;390(10090):191-202. https://doi.org/10.1016/s0140-6736(16)32586-7

Glasziou P, Straus S, Brownlee S, Trevena L, Dans L, Guyatt G, et al. Evidence for underuse of effective medical services around the world. Lancet. 2017;390(10090):169-77. https://doi.org/10.1016/s0140-6736(16)30946-1

Graber ML. The incidence of diagnostic error in medicine. BMJ Qual Saf. 2013;22 Suppl 2(Suppl 2):ii21-ii7. https://doi.org/10.1136/bmjqs-2012-001615

Saini V, Garcia-Armesto S, Klemperer D, Paris V, Elshaug AG, Brownlee S, et al. Drivers of poor medical care. Lancet. 2017;390(10090):178-90. https://doi.org/10.1016/s0140-6736(16)30947-3

Scott IA, Crock C. Diagnostic error: incidence, impacts, causes and preventive strategies. Med J Aust. 2020;213(7):302-5.e2. https://doi.org/10.5694/mja2.50771

Research studies

Conrad P, Mackie T, Mehrotra A. Estimating the costs of medicalization. Soc Sci Med. 2010;70(12):1943-7. https://doi.org/10.1016/j.socscimed.2010.02.019

Other references

Cochrane AL, Blythe M. One Man's Medicine. An autobiography of Professor Archie Cochrane. London: The British Medical Journal; 1989.

Lavis JN, Wilson MG, Oxman AD, Grimshaw J, Lewin S, Fretheim A. SUPPORT Tools for evidence-informed health Policymaking (STP) 5: Using research evidence to frame options to address a problem. Health Res Policy Syst. 2009a;7 Suppl 1:S5. https://doi.org/10.1186/1478-4505-7-s1-s5

Lavis JN, Wilson MG, Oxman AD, Lewin S, Fretheim A. SUPPORT Tools for evidence-informed health Policymaking (STP) 4: Using research evidence to clarify a problem. Health Res Policy Syst. 2009b;7 Suppl 1:S4. https://doi.org/10.1186/1478-4505-7-s1-s4

Moynihan R, Henry D. The fight against disease mongering: generating knowledge for action. PLoS Med. 2006;3(4):e191. https://doi.org/10.1371/journal.pmed.0030191

3.1b Consider the relevance of the outcomes measured in the research.

Explanation

A fair comparison may not include all outcomes – short- and long-term – that are important. Patients, professionals, and researchers may have different views about which outcomes are important. For example, studies often measure outcomes, such as heart rhythm irregularities, as surrogates for important outcomes, like death after heart attack. The effects of treatments on surrogate outcomes often do not provide a reliable indication of the effects on outcomes that are important. Similarly, short-term effects may not reflect long-term effects.

Despite dozens of randomized trials since the introduction of the first oral agent for treating type 2 diabetes, it has remained uncertain if any of those medicines favourably affects outcomes that are important to people, including morbidity, mortality, and quality of life *[Montori 2007]*. A key reason for this is that the trials have focused on glucose control measured with laboratory tests rather than on outcomes that are important to people with diabetes. Unfortunately, those laboratory tests (HbA) are not a reliable indicator of outcomes that are important to people with type 2 diabetes.

It is sometimes important to consider outcomes that are important to other people besides the person being treated. For example, the use of antibiotics may increase antibiotic resistance, and not being vaccinated for Covid-19 or not avoiding contact with other people may increase the risk of infection for others. Similarly, when decisions are made for a group of people rather than for individuals, the outcomes that are important to anyone who is affected should be considered.

Basis for this concept

A systematic review found 436 registered randomized trials that enrolled patients with diabetes *[Gandhi 2008 (SR)]*. Primary outcomes were patient-important outcomes in only 78 (18%) of the trials. One reason for trials measuring surrogate outcomes rather than patient-important outcomes is the preference of researchers and funding agencies to obtain results faster, with fewer patients and at lower costs. A major downside of this is that the results do not provide information about benefits that patients would consider important, given the paucity of validation of surrogate outcomes in diabetes and other conditions *[Bucher 1999]*.

A systematic review of trials using surrogate outcomes compared to trials using patient relevant outcomes found that surrogate outcomes reported larger treatment effects than trials reporting patient relevant outcomes *[Ciani 2013 (SR)]*. This finding was not explained by differences in the risk of bias or characteristics of the two groups of trials. In the absence of patient relevant outcomes, it is important to consider whether surrogate outcomes have been validated and uncertainty about whether surrogate outcomes predict important benefits and harms.

Implications

Always consider the possibility that important outcomes may not have been addressed in fair comparisons. Avoid being misled by surrogate outcomes.

References

Systematic reviews

Ciani O, Buyse M, Garside R, Pavey T, Stein K, Sterne JA, et al. Comparison of treatment effect sizes associated with surrogate and final patient relevant outcomes in randomised controlled trials: meta-epidemiological study. BMJ. 2013;346:f457. https://doi.org/10.1136/bmj.f457

Gandhi GY, Murad MH, Fujiyoshi A, Mullan RJ, Flynn DN, Elamin MB, et al. Patient-important outcomes in registered diabetes trials. JAMA. 2008;299(21):2543-9. https://doi.org/10.1001/jama.299.21.2543

Other references

Bucher HC, Guyatt GH, Cook DJ, Holbrook A, McAlister FA. Users' guides to the medical literature: XIX. Applying clinical trial results. A. How to use an article measuring the effect of an intervention on surrogate end points. Evidence-Based Medicine Working Group. JAMA. 1999;282(8):771-8. https://doi.org/10.1001/jama.282.8.771

Montori VM, Gandhi GY, Guyatt GH. Patient-important outcomes in diabetes--time for consensus. Lancet. 2007;370(9593):1104-6. https://doi.org/10.1016/s0140-6736(07)61489-5

3.1c Consider the relevance of fair comparisons in laboratories, animals, or highly selected people.

Explanation

Studies that only include animals, or only a selected minority of people, may not provide results that are relevant to most people.

Here are some examples of misleading extrapolation from animals or a selected minority of people, found in news reports *[Haneef 2015 (RS)]*:

Quote from news reports	**Basis for the quote**
"Researchers have shown that contact lenses laced with medicines are an effective way of treating glaucoma patients."	A study that showed the effect only in rabbit eyes.
"It could treat phobias and perhaps even post-traumatic stress disorders."	A before-after study in 15 healthy volunteers without any phobia.
"Broccoli slows arthritis".	A study in mice of a sulphoraphane compound present in cruciferous vegetables, including broccoli.
"The results of the trial – the first in humans – could offer hope to one in five people who are resistant to statins. It could also be offered to patients who suffer ill-effects from the drugs, or those whose cholesterol remains high even after statins are prescribed."	A study in healthy volunteers with high cholesterol levels who had received no lipid-lowering treatment in the past 30 days and were not statin resistant.
"Everyone should have at least 10-15 minutes of exposure to the sun every day to ensure that vitamin D levels are adequate."	A study in rats that assessed dietary vitamin D deficiency leading to elevated tyrosine nitration in the brain, which may promote cognitive decline.

Basis for this concept

It has been estimated that 11% of agents tested in humans are ultimately licensed, and only 5% of high-impact basic science discoveries claiming practical relevance are successfully translated into approved agents within a decade *[Henderson 2013 (SR)]*. Testing so many agents is potentially harmful to individuals in trials, and wastes resources. Animal studies are used to screen drugs and other treatments prior to testing in humans. A reason that so many animal studies fail to predict effectiveness or safety in humans is the use of treatments, animal models, or outcome assessments that are poorly matched to people – for example, using an acute disease model in animals to represent a chronic disease in humans. Another is when the pathophysiology underlying the disease in humans is not the same as in animals. A third reason is poorly designed and conducted animal studies. Several systematic reviews have documented major shortcomings of animal studies that limit their usefulness, including being too small, being badly reported and poorly summarised and interpreted in systematic reviews, being inconsistent, having a high risk of bias, and using animal models that cannot be generalised to humans *[Avey 2016 (SR), Bahadoran 2020 , Grüter 2020 (SR), Korevaar 2011 (SR), Kringe 2020 (SR), Lamontagne 2010 (SR), Moja 2014 (SR), Mueller 2014 (SR), Roberts 2002 (SR), Xiao 2021 (SR)]*.

Reviews that have compared the results of animal studies to studies in humans have found success rates that range from 0% to 100% *[Leenaars 2019 (SR)]*. This wide range suggests that the potential of animal studies to predict successful treatments in humans is unpredictable.

Implications

Results of systematic reviews of studies in animals, or highly selected groups of people, may be misleading.

References

Systematic reviews

Avey MT, Moher D, Sullivan KJ, Fergusson D, Griffin G, Grimshaw JM, et al. Thedevil is in the details: incomplete reporting in preclinical animal research. PLoS One. 2016;11(11):e0166733. https://doi.org/10.1371/journal.pone.0166733

Grüter BE, Croci D, Schöpf S, Nevzati E, d'Allonzo D, Lattmann J, et al. Systematic review and meta-analysis of methodological quality in in vivo animal studies of subarachnoid hemorrhage. Transl Stroke Res. 2020;11(6):1175-84. https://doi.org/10.1007/s12975-020-00801-4

Henderson VC, Kimmelman J, Fergusson D, Grimshaw JM, Hackam DG. Threats to validity in the design and conduct of preclinical efficacy studies: a systematic review of guidelines for in vivo animal experiments. PLoS Med. 2013;10(7):e1001489. https://doi.org/10.1371/journal.pmed.1001489

Korevaar DA, Hooft L, ter Riet G. Systematic reviews and meta-analyses of preclinical studies: publication bias in laboratory animal experiments. Lab Anim. 2011;45(4):225-30. https://doi.org/10.1258/la.2011.010121

Kringe L, Sena ES, Motschall E, Bahor Z, Wang Q, Herrmann AM, et al. Quality and validity of large animal experiments in stroke: a systematic review. J Cereb Blood Flow Metab. 2020;40(11):2152-64. https://doi.org/10.1177/0271678x20931062

Lamontagne F, Briel M, Duffett M, Fox-Robichaud A, Cook DJ, Guyatt G, et al. Systematic review of reviews including animal studies addressing therapeutic interventions for sepsis. Crit Care Med. 2010;38(12):2401-8. https://doi.org/10.1097/ccm.0b013e3181fa0468

Leenaars CHC, Kouwenaar C, Stafleu FR, Bleich A, Ritskes-Hoitinga M, De Vries RBM, et al. Animal to human translation: a systematic scoping review of reported concordance rates. J Transl Med. 2019;17(1):223. https://doi.org/10.1186/s12967-019-1976-2

Moja L, Pecoraro V, Ciccolallo L, Dall'Olmo L, Virgili G, Garattini S. Flaws in animal studies exploring statins and impact on meta-analysis. Eur J Clin Invest. 2014;44(6):597-612. https://doi.org/10.1111/eci.12264

Mueller KF, Briel M, Strech D, Meerpohl JJ, Lang B, Motschall E, et al. Dissemination bias in systematic reviews of animal research: a systematic review. PLoS One. 2014;9(12):e116016. https://doi.org/10.1371/journal.pone.0116016

Roberts I, Kwan I, Evans P, Haig S. Does animal experimentation inform human healthcare? Observations from a systematic review of international animal experiments on fluid resuscitation. BMJ. 2002;324(7335):474-6. https://doi.org/10.1136/bmj.324.7335.474

Xiao LY, Li Z, Du YZ, Shi HY, Yang SQ, Zhang YX, et al. Acupuncture for hypertension in animal models: a systematic review and meta-analysis. Evid Based Complement Alternat Med. 2021;2021:8171636. https://doi.org/10.1155/2021/8171636

Research studies

Haneef R, Lazarus C, Ravaud P, Yavchitz A, Boutron I. Interpretation of results of studies evaluating an intervention highlighted in Google Health News: a cross-sectional study of news. PLoS One. 2015;10(10):e0140889. https://doi.org/10.1371/journal.pone.0140889

Other references

Bahadoran Z, Mirmiran P, Kashfi K, Ghasemi A. Importance of systematic reviews and meta-analyses of animal studies: challenges for animal-to-human translation. J Am Assoc Lab Anim Sci. 2020;59(5):469-77. https://doi.org/10.30802/aalas-jaalas-19-000139

3.1d Consider the relevance of the treatments that were compared.

Explanation

A fair comparison of the effects of a surgical procedure done in a specialised hospital or delivered by an experienced surgeon may not provide a reliable estimate of its effects and safety in other settings, or in the hands of less experienced surgeons.

For example, results from a large randomized trial showed that endarterectomy (surgical removal of part of the inner lining of an artery) for asymptomatic carotid stenosis (narrowing of the large arteries on each side of the neck) reduced the five-year absolute risk of stroke by about 5% *[Rothwell 2005]*. However, the trial only accepted surgeons with a good safety record, rejecting 40% of applicants and subsequently barring those who had adverse operative outcomes in the trial from further participation. The benefit from surgery was largely attributable to the low operative risk. Operative mortality was eight-fold higher outside of the trial and the risk of stroke and death was about three-fold higher.

Similarly, comparing a new drug to a drug or dose that is not commonly used (and which may be less effective or safe than those in common use) would not provide a relevant estimate of how the new drug compares to what is commonly done.

For example, in randomized trials of atypical antipsychotics for schizophrenia, haloperidol (one of the most frequently prescribed "typical" antipsychotics worldwide) was used as the comparison treatment *[Hugenholtz 2006 (SR)]*. However, the trials used haloperidol in doses that were higher than that recommended. In a meta-analysis of 52 randomized trials that controlled for the higher-than-recommended dose of comparator drugs, differences in effectiveness and overall tolerability between typical and atypical antipsychotics disappeared, suggesting that the perceived benefits of atypical antipsychotics were due to excessive doses of the comparison treatments, such as haloperidol *[Geddes 2000 (SR)]*.

Basis for this concept

Characteristics of treatments, including the duration, dose or intensity, mode of delivery, and skill of the person delivering the treatment can influence the effectiveness of a treatment. Unfortunately, characteristics such as these are poorly described. This can make it difficult to judge the relevance of treatments compared in studies to other contexts. For example, a review of randomized trials in oncology found that only 11% of 262 trials of cancer chemotherapy provided complete details of the trial treatments *[Duff 2010 (SR)]*. The completeness of treatment descriptions is often worse for non-pharmacological treatments. A review of 80 randomized trials and reviews found that 67% of descriptions of drug treatments were adequate compared with only 29% of non-pharmacological treatments *[Glasziou 2008 (SR)]*. Another review of 137 randomized trials of non-pharmacological treatments found that only 39% of treatments were adequately described *[Hoffmann 2013 (SR)]*.

Inadequate descriptions of treatments led to the development of the Template for Intervention Description and Replication (TIDieR) checklist *[Hoffmann 2014 , Hoffmann 2017]*. TIDieR has helped to improve descriptions of treatments, but further improvements are needed. An overview of 56 reviews that used the TIDieR checklist to evaluate the adequacy of treatment descriptions found that the names of treatments and reasons for using them were generally reported adequately *[Dijkers 2021 (SR)]*. However, only between 25% and 75% adequately reported other characteristics of treatments and as few as 10% of studies adequately reported modifications of treatments. Comparison treatments were reported less well than the treatment that was the focus of the comparison.

Another challenge with making judgements about the relevance of treatments is the choice of the comparison treatment, as illustrated by the example of treatments for schizophrenia presented above. This is often a problem for pharmacological treatments. Many drug trials are funded by industry and pharmaceutical companies typically choose to compare their drugs to a placebo rather than to another drug *[Dunn 2013 (SR), Lathyris 2010 (SR)]*. When there is more than one effective treatment available, people often need to decide which treatment to use, not whether to use a particular treatment or a placebo. Consequently, when direct comparisons of treatments are not available, as is often the case, indirect comparisons (across studies) must be used (see Concept 2.2c).

A related challenge for judgements about the relevance of pharmacological treatments is the assumption that drugs within a class are interchangeable *[Furberg 1999 , Furberg 2003 , McAlister 1999 , Mills 2014]*. Pharmaceuticals are categorised as members of existing "drug classes". The U.S. Food and Drug Administration (FDA) uses class labelling when "all products within a class are assumed to be closely related in chemical structure, pharmacology, therapeutic activity, and adverse reactions". However, this assumption can be dangerous. In addition to drugs within a class not being directly compared, new drugs are often approved based on randomized trials that measure surrogate outcomes rather than outcomes that are important to people (see Concept 3.1b). But there can be important differences in both beneficial and harmful effects of drugs within the same class. It should not be assumed that drugs within a class are interchangeable in the absence of reliable evidence of comparable benefits and long-term safety.

Implications

Be aware that treatments available to you may be sufficiently different from those in the research studies that the results may not apply to you.

References

Systematic reviews

Dijkers MP. Overview of Reviews Using the Template for Intervention Description and Replication (TIDieR) as a Measure of Trial Intervention Reporting Quality. Arch Phys Med Rehabil. 2021;102(8):1623-32. https://doi.org/10.1016/j.apmr.2020.09.397

Duff JM, Leather H, Walden EO, LaPlant KD, George TJ, Jr. Adequacy of published oncology randomized controlled trials to provide therapeutic details needed for clinical application. J Natl Cancer Inst. 2010;102(10):702-5. https://doi.org/10.1093/jnci/djq117

Dunn AG, Mandl KD, Coiera E, Bourgeois FT. The effects of industry sponsorship on comparator selection in trial registrations for neuropsychiatric conditions in children. PLoS One. 2013;8(12):e84951. https://doi.org/10.1371/journal.pone.0084951

Geddes J, Freemantle N, Harrison P, Bebbington P. Atypical antipsychotics in the treatment of schizophrenia: systematic overview and meta-regression analysis. BMJ. 2000;321(7273):1371-6. https://doi.org/10.1136/bmj.321.7273.1371

Glasziou P, Meats E, Heneghan C, Shepperd S. What is missing from descriptions of treatment in trials and reviews? BMJ. 2008;336(7659):1472-4. https://doi.org/10.1136/bmj.39590.732037.47

Hoffmann TC, Erueti C, Glasziou PP. Poor description of non-pharmacological interventions: analysis of consecutive sample of randomised trials. BMJ. 2013;347:f3755. https://doi.org/10.1136/bmj.f3755

Hugenholtz GW, Heerdink ER, Stolker JJ, Meijer WE, Egberts AC, Nolen WA. Haloperidol dose when used as active comparator in randomized controlled trials with atypical antipsychotics in schizophrenia: comparison with officially recommended doses. J Clin Psychiatry. 2006;67(6):897-903. https://doi.org/10.4088/jcp.v67n0606

Lathyris DN, Patsopoulos NA, Salanti G, Ioannidis JP. Industry sponsorship and selection of comparators in randomized clinical trials. Eur J Clin Invest. 2010;40(2):172-82. https://doi.org/10.1111/j.1365-2362.2009.02240.x

Other references

Furberg CD, Herrington DM, Psaty BM. Are drugs within a class interchangeable? Lancet. 1999;354(9185):1202-4. https://doi.org/10.1016/s0140-6736(99)03190-6

Furberg CD, Psaty BM. Should evidence-based proof of drug efficacy be extrapolated to a "class of agents"? Circulation. 2003;108(21):2608-10. https://doi.org/10.1161/01.cir.0000090572.51900.92

Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. BMJ. 2014;348:g1687. https://doi.org/10.1136/bmj.g1687

Hoffmann TC, Oxman AD, Ioannidis JP, Moher D, Lasserson TJ, Tovey DI, et al. Enhancing the usability of systematic reviews by improving the consideration and description of interventions. BMJ. 2017;358:j2998. https://doi.org/10.1136/bmj.j2998

McAlister FA, Laupacis A, Wells GA, Sackett DL. Users' Guides to the Medical Literature: XIX. Applying clinical trial results B. Guidelines for determining whether a drug is exerting (more than) a class effect. JAMA. 1999;282(14):1371-7. https://doi.org/10.1001/jama.282.14.1371

Mills EJ, Gardner D, Thorlund K, Briel M, Bryan S, Hutton B, et al. A users' guide to understanding therapeutic substitutions. J Clin Epidemiol. 2014;67(3):305-13. https://doi.org/10.1016/j.jclinepi.2013.09.008

Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?". Lancet. 2005;365(9453):82-93. https://doi.org/10.1016/s0140-6736(04)17670-8

3.1e Consider the relevance of the circumstances in which the treatments were compared.

Explanation

Some treatment comparisons are designed to find out if a treatment can work under ideal circumstances, for example, with people who are most likely to benefit and most likely to comply, and with highly trained practitioners who deliver the treatment exactly as intended. These comparisons, which are sometimes called 'explanatory' or 'efficacy' studies, may not reflect what happens under usual circumstances.

The North American Symptomatic Carotid Endarterectomy Trial is an example of an "explanatory randomized trial" *[Barnett 1991 (RS), Thorpe 2009 , Zwarenstein 2008]*. It demonstrated that a lower risk of stroke was experienced by highly selected patients with severe narrowing of the large arteries on either side of the neck carotid stenosis. Those patients who had a high risk of stroke and were most likely to respond to surgical removal of the inner layer of those arteries (endarterectomy). Participating patients were closely followed and were operated on by skilled surgeons in academic and specialist hospitals who adhered to a strict protocol. The trial showed that endarterectomy reduced the risk of stroke under those ideal circumstances, but it did not provide a reliable estimate of the beneficial and harmful effects of endarterectomy under more typical circumstances.

A randomized trial of a short course of acupuncture to reduce pain in patients with persistent non-specific low-back pain is an example of a "pragmatic trial" *[Thomas 2006 (RS), Zwarenstein 2008]*. It was conducted in general practice and private acupuncture clinics in the UK and enrolled anyone aged 18-65 with non-specific low back pain of 4-52 weeks duration. The acupuncturists determined the content and number of treatments. The results of this trial were directly relevant to general practitioners and patients.

Basis for this concept

"Explanatory randomized trials" seek to answer the question "Can this treatment work under ideal conditions?" "Pragmatic randomized trials seek to answer the question "Does this treatment work under usual conditions?" Explanatory and pragmatic trials can differ in several ways, including eligibility criteria (restricted to patients who are most likely to respond to the treatment versus all patients with the condition of interest), expertise of the health professional, the comparison treatment, efforts to ensure compliance with the assigned treatment, and how the results are analysed *[Loudon 2015 , Thorpe 2009 , Zwarenstein 2008]*. Trials are not necessarily purely explanatory or pragmatic. They can be explanatory in some ways and pragmatic in other ways.

Only a small proportion of published trials are explicitly described as pragmatic. An analysis in 2002 identified 95 randomized trials that were identified as pragmatic in the title or abstract *[Vallvé 2003]*. A simple search in PubMed (15 December 2021) found 154,829 articles indexed as a "randomized controlled trial" and of those 2,362 (< 2%) were indexed as a "pragmatic clinical trial". Another indication of the scarcity of pragmatic trials can be deduced from services that use skilled critical appraisers to filter clinical publications, initially on methodological quality, and then by experienced practitioners in a particular field to screen for relevance to practice and noteworthiness. For Family Medicine, one of these services sends out just 30 or 40 article abstracts per year, in monthly batches. Of these, perhaps half are randomized trials, which means that this system has retained as valid, useful, and noteworthy only 15 or 20 of the 1,000 or more randomized trials published each year that are relevant to this field *[Zwarenstein 2006]*. Although these analyses may underestimate the proportion of randomized trials that are mostly pragmatic, they suggest that randomized trials are mostly explanatory. Reasons for this include financial interests of pharmaceutical and medical device manufacturers that sponsor trials, regulations for licensing pharmaceuticals, and a lack of public funding for large pragmatic trials *[Tunis 2003 , Zwarenstein 2009]*. Other ways of estimating the proportion of randomized trials that are pragmatic suggest that in some contexts there may be a higher proportion of pragmatic trials. For example, an analysis of randomized trials used in a guideline for psychosocial treatments for psychoses found that about one-third of the studies were pragmatic *[Gastaldon 2019 (SR)]*. Another analysis of randomized trials of surgery published in 2010 and 2011 classified about half of the trials as pragmatic, although the trial reports did not provide information about several aspects of the trials, such as the expertise of the surgeons *[Blencowe 2015 (SR)]*.

Implications

Be aware that the results of studies with the aim of finding out if a treatment can work may overestimate the benefits of a treatment under more usual circumstances.

References

Systematic reviews

Blencowe NS, Boddy AP, Harris A, Hanna T, Whiting P, Cook JA, et al. Systematic review of intervention design and delivery in pragmatic and explanatory surgical randomized clinical trials. Br J Surg. 2015;102(9):1037-47. https://doi.org/10.1002/bjs.9808

Gastaldon C, Mosler F, Toner S, Tedeschi F, Bird VJ, Barbui C, et al. Are trials of psychological and psychosocial interventions for schizophrenia and psychosis included in the NICE guidelines pragmatic? A systematic review. PLoS One. 2019;14(9):e0222891. https://doi.org/10.1371/journal.pone.0222891

Other reviews

Vallvé C. [A critical review of the pragmatic clinical trial]. Med Clin (Barc). 2003;121(10):384-8. https://doi.org/10.1016/s0025-7753(03)73957-8

Research studies

Barnett HJM, Taylor DW, Haynes RB, Sackett DL, Peerless SJ, Ferguson GG, et al. Beneficial effect of carotid endarterectomy in symptomatic patients with high-grade carotid stenosis. N Engl J Med. 1991;325(7):445-53. https://doi.org/10.1056/nejm199108153250701

Thomas KJ, MacPherson H, Thorpe L, Brazier J, Fitter M, Campbell MJ, et al. Randomised controlled trial of a short course of traditional acupuncture compared with usual care for persistent non-specific low back pain. BMJ. 2006;333(7569):623. https://doi.org/10.1136/bmj.38878.907361.7c

Other references

Loudon K, Treweek S, Sullivan F, Donnan P, Thorpe KE, Zwarenstein M. The PRECIS-2 tool: designing trials that are fit for purpose. BMJ. 2015;350:h2147. https://doi.org/10.1136/bmj.h2147

Thorpe KE, Zwarenstein M, Oxman AD, Treweek S, Furberg CD, Altman DG, et al. A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers. CMAJ. 2009;180(10):E47-57. https://doi.org/10.1503/cmaj.090523

Tunis SR, Stryer DB, Clancy CM. Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. JAMA. 2003;290(12):1624-32. https://doi.org/10.1001/jama.290.12.1624

Zwarenstein M, Oxman A, Pragmatic Trials in Health Care S. Why are so few randomized trials useful, and what can we do about it? J Clin Epidemiol. 2006;59(11):1125-6. https://doi.org/10.1016/j.jclinepi.2006.05.010

Zwarenstein M, Treweek S. What kind of randomized trials do we need? J Clin Epidemiol. 2009;62(5):461-3. https://doi.org/10.1016/j.jclinepi.2009.01.011

Zwarenstein M, Treweek S, Gagnier JJ, Altman DG, Tunis S, Haynes B, et al. Improving the reporting of pragmatic trials: an extension of the CONSORT statement. BMJ. 2008;337:a2390. https://doi.org/10.1136/bmj.a2390

3.2 Expected advantages should outweigh expected disadvantages.

3.2a Weigh the benefits and savings against the harms and costs of acting or not.

Explanation

Individuals, clinicians, and policymakers deciding about whether to use a treatment should consider the potential benefits and the potential harms, costs and other advantages and disadvantages of the treatment. When a decision affects many people, it is important to consider the distribution of the advantages and disadvantages, i.e., who will benefit, who will be harmed, who will achieve savings, and who will bear the costs.

When the advantages of a treatment clearly outweigh the disadvantages, deciding what to do is relatively easy. For example, for patients who have had a heart attack, stroke or transient ischemic attack, the advantages of low-dose aspirin compared to not taking aspirin (reduced deaths, heart attacks, and strokes) are substantially more than the disadvantages (increased serious gastrointestinal bleeds, and minimal inconvenience and cost) *[Vandvik 2012]*. Most people in this situation would choose to take aspirin. On the other hand, when the advantages and disadvantages are closely balanced, deciding what to do can be difficult. For example, for someone 50 years or older without symptomatic cardiovascular disease, aspirin only slightly reduces deaths if taken over 10 years, and a reduction in heart attacks is closely balanced with an increase in serous gastrointestinal bleeds. Some people in this situation would choose to take aspirin, and some would not.

Basis for this concept

Treatments have both advantages and disadvantages. Often, people tend to exaggerate the advantages of treatments and ignore or downplay their disadvantages (see Concept 1.1a). For some treatments, the advantages clearly outweigh the disadvantages. However, the advantages and disadvantages are often closely balanced, and need to be carefully weighed. UpToDate is a widely used electronic medical textbook that includes thousands of recommendations based on the best available evidence. Strong recommendations are ones for which the authors were confident that the desirable consequences clearly outweigh the undesirable consequences *[Andrews 2013a , Andrews 2013b , Guyatt 2008a]*. Weak or conditional recommendations are ones for which the balance of desirable and undesirable consequences between alternatives is close or uncertain. A review of more than 9,400 recommendations in UpToDate found that less than one-third (31%) of the recommendations were strong *[Agoritsas 2017 (RS)]*. Most (69%) of the recommendations were weak or conditional. A review of 456 recommendations in 43 guidelines found a higher proportion of strong recommendations (63%) *[Alexander 2014 (RS)]*. However, more than half (56%) of the strong recommendations were based on evidence warranting very low confidence in the effect estimates and another 23% on evidence warranting low confidence. A critical review of those recommendations determined that 46% warranted a conditional, rather than a strong recommendation *[Alexander 2016 (RS)]*, suggesting that 17% of the recommendations were strong.

Cost-effectiveness analyses can help to inform judgements about whether the net benefits of a treatment (the difference between the benefits and the harms) is worth the cost. Cost-effectiveness analyses are especially helpful for health insurance schemes when deciding which treatments should be paid for. Because these analyses use models and depend on assumptions, the results are often uncertain (see Concept 2.2d). Nonetheless, sensitivity analyses can help to reveal important uncertainties and the models can help to inform decisions. However, most published analyses report cost-effectiveness ratios below thresholds commonly used to decide whether a treatment is "cost-effective", and industry funded analyses are more likely to report ratios below those thresholds than other analyses *[Bell 2006 (SR)]*. In addition, there is no evidence of an agreed public threshold *[Harris 2008 (RS)]*. Willingness to pay for a treatment is related to the severity of the condition being treated, the importance of the treatment effect, confidence in the evidence, and total cost to the government or other payer, as well as the estimated cost-effectiveness. Equity, acceptability, and feasibility may also influence decisions *[Alonso-Coello 2016]*.

Implications

Always consider the balance between advantages and disadvantages of treatments.

References

Systematic reviews

Bell CM, Urbach DR, Ray JG, Bayoumi A, Rosen AB, Greenberg D, et al. Bias in published cost effectiveness studies: systematic review. BMJ. 2006;332(7543):699-703. https://doi.org/10.1136/bmj.38737.607558.80

Research studies

Agoritsas T, Merglen A, Heen AF, Kristiansen A, Neumann I, Brito JP, et al. UpToDate adherence to GRADE criteria for strong recommendations: an analytical survey. BMJ Open. 2017;7(11):e018593. https://doi.org/10.1136/bmjopen-2017-018593

Alexander PE, Bero L, Montori VM, Brito JP, Stoltzfus R, Djulbegovic B, et al. World Health Organization recommendations are often strong based on low confidence in effect estimates. J Clin Epidemiol. 2014;67(6):629-34. https://doi.org/10.1016/j.jclinepi.2013.09.020

Alexander PE, Brito JP, Neumann I, Gionfriddo MR, Bero L, Djulbegovic B, et al. World Health Organization strong recommendations based on low-quality evidence (study quality) are frequent and often inconsistent with GRADE guidance. J Clin Epidemiol. 2016;72:98-106. https://doi.org/10.1016/j.jclinepi.2014.10.011

Harris AH, Hill SR, Chin G, Li JJ, Walkom E. The role of value for money in public insurance coverage decisions for drugs in Australia: a retrospective analysis 1994-2004. Med Decis Making. 2008;28(5):713-22. https://doi.org/10.1177/0272989x08315247

Other references

Alonso-Coello P, Schunemann HJ, Moberg J, Brignardello-Petersen R, Akl EA, Davoli M, et al. GRADE Evidence to Decision (EtD) frameworks: a systematic and transparent approach to making well informed healthcare choices. 1: Introduction. BMJ. 2016;353:i2016. https://doi.org/10.1136/bmj.i2016

Andrews J, Guyatt G, Oxman AD, Alderson P, Dahm P, Falck-Ytter Y, et al. GRADE guidelines: 14. Going from evidence to recommendations: the significance and presentation of recommendations. J Clin Epidemiol. 2013a;66(7):719-25. https://doi.org/10.1016/j.jclinepi.2012.03.013

Andrews JC, Schunemann HJ, Oxman AD, Pottie K, Meerpohl JJ, Coello PA, et al. GRADE guidelines: 15. Going from evidence to recommendation-determinants of a recommendation's direction and strength. J Clin Epidemiol. 2013b;66(7):726-35. https://doi.org/10.1016/j.jclinepi.2013.02.003

Guyatt GH, Oxman AD, Kunz R, Falck-Ytter Y, Vist GE, Liberati A, et al. Going from evidence to recommendations. BMJ. 2008a;336(7652):1049-51. https://doi.org/10.1136/bmj.39493.646875.ae

Vandvik PO, Lincoff AM, Gore JM, Gutterman DD, Sonnenberg FA, Alonso-Coello P, et al. Primary and secondary prevention of cardiovascular disease: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. Chest. 2012;141(2 Suppl):e637S-e68S. https://doi.org/10.1378/chest.11-2306

3.2b Consider the baseline risk or severity of the symptoms when estimating the size of expected effects.

Explanation

The balance between the benefits and harms of treatments often depends on the baseline risk (the likelihood of an individual experiencing an undesirable event), or on the severity of the symptoms. The balance between the advantages and disadvantages of a treatment is more likely to favour the use of a treatment by people with a higher baseline risk, or more severe symptoms. For example, consider patients who have had a heart attack, stroke, or transient ischemic attack, or have a high probability of dying or having another cardiovascular event in the next **five years** (see table below). Because they have a high baseline risk, aspirin has a large absolute effect (risk difference), despite the relative effect being small to moderate, and the benefits substantially outweigh the harms for someone in this situation *[Vandvik 2012]*.

Probability of an event in the next 5 years for someone with a high baseline risk

Outcome	**Relative risk reduction** (95% confidence interval)	**Risk without aspirin in the next 5 years**	**Risk difference** (95% confidence interval)
Deaths	10% (1% to 18%)	133 per 1,000	13 fewer per 1,000 (1 to 24 fewer)
Strokes	19% (8% to 29%)	135 per 1,000	26 fewer per 1,000 (11 to 39 fewer)
Heart attacks	31% (20% to 40%)	117 per 1,000	37 fewer per 1,000 (23 to 47 fewer)
Serious gastrointestinal bleeds	169% increase (25% to 476%)	15 per 1,000	25 more per 1,000 (4 to 71 more)

On the other hand, for someone 60 years old without symptomatic cardiovascular disease who has a low risk of having a cardiovascular event or a gastrointestinal bleed, aspirin has little if any beneficial effect on deaths and strokes. The probability of having a heart attack (27 per 1,000 in the next **10 years**) is much lower than it is for someone who has had a cardiovascular event and has a high risk (117 per 1,000 in the next **five years**). The relative effect is also slightly lower. The absolute effect is six fewer heart attacks per 1,000 people who take aspirin for 10 years (see table below), compared to 37 fewer per 1,000 people who take aspirin for just five years. The relative risk increase, the baseline risk without aspirin, and the risk difference for having a serious gastrointestinal bleed are also less for someone who has not had a cardiovascular event and has a low risk of bleeding. Consequently, the benefits and harms of low-dose aspirin are closely balanced for someone in this situation.

Probability of an event in the next 10 years for someone with a low baseline risk

Outcome	**Relative risk reduction** (95% confidence interval)	**Risk without aspirin in the next 10 years**	**Risk difference** (95% confidence interval)
Heart attacks	23% (14% to 31%)	27 per 1,000	6 fewer per 1,000 (4 to 8 fewer)
Serious gastrointestinal bleeds	54% increase (30% to 82%)	8 per 1,000	4 more per 1,000 (2 to 7 more)

Basis for this concept

Relative measures tend to be consistent across risk groups, but aren't always *[Deeks 2002 (RS), Engels 2000 (RS), Furukawa 2002 (RS), Schmid 1998 (RS)]*, as illustrated in the above example (see Concept 2.3b). The risk difference can be estimated by applying the relative effect to one or more relevant baseline risks, as illustrated in the tables above. Generally, the benefits of a treatment are less for someone with a low risk of an outcome compared to someone with a high risk. On the other hand, the risk of adverse effects is often the same (although it was not in the example above). Therefore, the benefits and harms of a treatment tend to be more closely balanced for people with a low risk of the condition being treated than for someone with a high risk. The same is true for people with more severe symptoms (e.g., pain or depression) compared to people with less severe symptoms.

Unfortunately, someone's baseline risk is often uncertain. Studies that estimate someone's risk or prognosis and systematic reviews of those studies have been scarce and often of poor quality, but both the quantity and quality of this research has been increasing *[Collins 2014 (SR), Debray 2017 , Matino 2017 (SR), Peat 2014 , Riley 2016 , Skoetz 2019 (SR)]*. Uncertainty in baseline risk estimates and its impact on confidence in absolute estimates of treatment effects are often not considered by guideline developers or systematic review authors *[Iorio 2015 , Spencer 2012]*, as illustrated in the tables above. However, important uncertainty about someone's baseline risk can reduce confidence in absolute effect estimates and increase uncertainty about the balance between the benefits and harms of a treatment *[Iorio 2015 , Spencer 2012]*.

Implications

When making decisions about treatments, consider the estimated baseline risk or the severity of symptoms.

References

Systematic reviews

Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol. 2014;14:40. https://doi.org/10.1186/1471-2288-14-40

Matino D, Chai-Adisaksopha C, Iorio A. Systematic reviews of prognosis studies: a critical appraisal of five core clinical journals. Diagn Progn Res. 2017;1:9. https://doi.org/10.1186/s41512-017-0008-z

Skoetz N, Goldkuhle M, Weigl A, Dwan K, Labonté V, Dahm P, et al. Methodological review showed correct absolute effect size estimates for time-to-event outcomes in less than one-third of cancer-related systematic reviews. J Clin Epidemiol. 2019;108:1-9. https://doi.org/10.1016/j.jclinepi.2018.12.006

Research studies

Deeks JJ. Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. Stat Med. 2002;21(11):1575-600. https://doi.org/10.1002/sim.1188

Engels EA, Schmid CH, Terrin N, Olkin I, Lau J. Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. Stat Med. 2000;19(13):1707-28. https://doi.org/10.1002/1097-0258(20000715)19:13%3C1707::aid-sim491%3E3.0.co;2-p

Furukawa TA, Guyatt GH, Griffith LE. Can we individualize the 'number needed to treat'? An empirical study of summary effect measures in meta-analyses. Int J Epidemiol. 2002;31(1):72-6. https://doi.org/10.1093/ije/31.1.72

Schmid CH, Lau J, McIntosh MW, Cappelleri JC. An empirical study of the effect of the control rate as a predictor of treatment efficacy in meta-analysis of clinical trials. Stat Med. 1998;17(17):1923-42. https://doi.org/10.1002/(sici)1097-0258(19980915)17:17%3C1923::aid-sim874%3E3.0.co;2-6

Other references

Debray TP, Damen JA, Snell KI, Ensor J, Hooft L, Reitsma JB, et al. A guide to systematic review and meta-analysis of prediction model performance. BMJ. 2017;356:i6460. https://doi.org/10.1136/bmj.i6460

Iorio A, Spencer FA, Falavigna M, Alba C, Lang E, Burnand B, et al. Use of GRADE for assessment of evidence about prognosis: rating confidence in estimates of event rates in broad categories of patients. BMJ. 2015;350:h870. https://doi.org/10.1136/bmj.h870

Peat G, Riley RD, Croft P, Morley KI, Kyzas PA, Moons KG, et al. Improving the transparency of prognosis research: the role of reporting, data sharing, registration, and protocols. PLoS Med. 2014;11(7):e1001671. https://doi.org/10.1371/journal.pmed.1001671

Riley RD, Ensor J, Snell KI, Debray TP, Altman DG, Moons KG, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ. 2016;353:i3140. https://doi.org/10.1136/bmj.i3140

Spencer FA, Iorio A, You J, Murad MH, Schünemann HJ, Vandvik PO, et al. Uncertainties in baseline risk estimates and confidence in treatment effects. BMJ. 2012;345:e7401. https://doi.org/10.1136/bmj.e7401

Vandvik PO, Lincoff AM, Gore JM, Gutterman DD, Sonnenberg FA, Alonso-Coello P, et al. Primary and secondary prevention of cardiovascular disease: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. Chest. 2012;141(2 Suppl):e637S-e68S. https://doi.org/10.1378/chest.11-2306

3.2c Consider how important each advantage and disadvantage is when weighing the pros and cons and making choices.

Explanation

Estimates of benefits and harms depend on how much weight people give to treatment advantages and disadvantages. Different people may value outcomes differently and sometimes make different choices because of this. In addition, people usually place more value on outcomes that happen soon than on outcomes that happen years into the future. In other words, the further into the future an outcome (for example, reducing the chance of heart disease or cancer after many years) the more people tend to "discount" its value or importance. The balance between the advantages and disadvantages of treatments may also depend on how much costs and events in the future are discounted.

Consider the example of aspirin to prevent cardiovascular disease in someone 60 years old with a low risk. The main advantage is a reduced risk of having a heart attack. The main disadvantage is an increased risk of having a serious gastrointestinal bleed, as shown in the table below *[Vandvik 2012]*.

Probability of an event in the next 10 years for someone with a low baseline risk

Outcome	**Relative risk reduction** (95% confidence interval)	**Risk without aspirin**	**Risk difference** (95% confidence interval)
Heart attacks	23% (14% to 31%)	27 per 1,000	6 fewer per 1,000 (4 to 8 fewer)
Serious gastrointestinal bleeds	54% increase (30% to 82%)	8 per 1,000	4 more per 1,000 (2 to 7 more)

Although aspirin costs very little, for someone with very little money, this might be another important disadvantage. There is also minimal inconvenience – taking a pill every day for 10 years – but for some people this might be enough of a bother to be another disadvantage. Someone who is more averse to having a heart attack than having a serious gastrointestinal bleed and who is not concerned about the cost or the bother, might choose to take aspirin. On the other hand, someone who is more averse to having a serious gastrointestinal bleed and less averse to having a heart attack, might choose not to take aspirin, especially if they were concerned about the cost or the bother.

Basis for this concept

People vary greatly in the importance they attribute to outcomes. A systematic review of studies that assessed how much people value potential benefits and harms of aspirin and other antithrombotic therapy found 48 studies *[MacLean 2012 (SR)]*. There were inconsistencies across studies and variation within studies. The authors concluded that, on average, a stroke was two or three times worse than a gastrointestinal bleed and a heart attack was between being about the same and two times worse than a gastrointestinal bleed. Those estimates of the relative importance of these outcomes were very uncertain, and not everyone is "average". Other systematic reviews have found similar variation in how much people value potential benefits and harms of treatments, as well as important limitations in studies that have measured people's values *[Etxeandia-Ikobaltzeta 2020 (SR), González-González 2021 (SR), Guerra 2019 (SR), Hansson 2021 (SR), Heen 2021 (SR), Malde 2021 (SR), Mathioudakis 2019 (SR), Muñoz-Velandia 2019 (SR), O'Reilly 2021 (SR), Pillay 2021 (SR), Valli 2019 (SR), Vernooij 2018 (SR), Zeng 2021 (SR), Zhang 2018 (SR)]*.

It is important to consider how long a condition lasts as well as how severe it is. For example, most people consider a severe stroke as being much worse than a heart attack or a gastrointestinal bleed. Some people even consider having a severe stroke to be worse than dying. In addition, disability following a stroke may last for years, whereas most people are able to return to a normal life shortly after having a nonfatal heart attack or gastrointestinal bleed. Because of differences in both the severity and duration of different outcomes, it can be misleading when researchers group together outcomes, such as "cardiovascular events", which can include heart attacks, strokes, deaths, and other outcomes (with different degrees of severity, duration, and occurrence) *[Cordoba 2010 (SR), Freemantle 2003 (SR), Lim 2008 (SR), McGrath 2011 (RS)]*.

Although a majority of people prefer to make decisions together with a health professional, some people prefer to delegate decisions to a health professional *[Chewning 2012 (SR)]*. Unfortunately, health professionals' perceptions of their patients' desire to be involved in decisions are often inaccurate *[Cox 2007 (RS)]*. They may be more likely to underestimate the extent to which patients prefer to be involved in decisions. Regardless of who decides, decisions should be consistent with a patient's values. Decision aids can help patients to clarify their values and may help them to make choices that are more consistent with their values compared to choices made without decision aids *[Stacey 2014 (SR)]*. There is some evidence that patients choose more conservative approaches when they become better informed *[Walsh 2014 (SR)]*.

Economists use quality-adjusted life-years (QALYs) as a measure that captures both the severity and duration of a condition and allows for comparisons across different conditions. However, QALYs reflect, at best, average values. The values attached to different conditions are often uncertain and individuals can have very different values. Thus, although QALYs can help, if used judiciously, to inform decisions about how healthcare resources are spent, they are unlikely to be helpful for patients and clinicians making decisions for individuals *[Franklin 2019 (RS), Rand 2021 (SR)]*.

When decisions are made for a group of people rather than for individuals, it is important to consider how much the people affected by the decision value the benefits and harms, whether there is important uncertainty about this, and whether there is important variability in how much people value the benefits and harms *[Moberg 2018]*.

Implications

Consider how important each treatment advantage and disadvantage is when choosing a treatment.

References

Systematic reviews

Chewning B, Bylund CL, Shah B, Arora NK, Gueguen JA, Makoul G. Patient preferences for shared decisions: a systematic review. Patient Educ Couns. 2012;86(1):9-18. https://doi.org/10.1016/j.pec.2011.02.004

Cordoba G, Schwartz L, Woloshin S, Bae H, Gøtzsche PC. Definition, reporting, and interpretation of composite outcomes in clinical trials: systematic review. BMJ. 2010;341:c3920. https://doi.org/10.1136/bmj.c3920

Etxeandia-Ikobaltzeta I, Zhang Y, Brundisini F, Florez ID, Wiercioch W, Nieuwlaat R, et al. Patient values and preferences regarding VTE disease: a systematic review to inform American Society of Hematology guidelines. Blood Adv. 2020;4(5):953-68. https://doi.org/10.1182/bloodadvances.2019000462

Freemantle N, Calvert M, Wood J, Eastaugh J, Griffin C. Composite outcomes in randomized trials: greater precision but with greater uncertainty? JAMA. 2003;289(19):2554-9. https://doi.org/10.1001/jama.289.19.2554

González-González JG, Díaz González-Colmenero A, Millán-Alanís JM, Lytvyn L, Solis RC, Mustafa RA, et al. Values, preferences and burden of treatment for the initiation of GLP-1 receptor agonists and SGLT-2 inhibitors in adult patients with type 2 diabetes: a systematic review. BMJ Open. 2021;11(7):e049130. https://doi.org/10.1136/bmjopen-2021-049130

Guerra RL, Castaneda L, de Albuquerque RCR, Ferreira CBT, Corrêa FM, Fernandes RRA, et al. Patient preferences for breast cancer treatment interventions: a systematic review of discrete choice experiments. Patient. 2019;12(6):559-69. https://doi.org/10.1007/s40271-019-00375-w

Hansson E, Sandman L, Davidson T. A systematic review of direct preference measurements in health states treated with plastic surgery. J Plast Surg Hand Surg. 2021:1-11. https://doi.org/10.1080/2000656x.2021.1953039

Heen AF, Lytvyn L, Shapiro M, Guyatt GH, Siemieniuk RAC, Zhang Y, et al. Patient values and preferences on valve replacement for aortic stenosis: a systematic review. Heart. 2021;107(16):1289-95. https://doi.org/10.1136/heartjnl-2020-318334

Lim E, Brown A, Helmy A, Mussa S, Altman DG. Composite outcomes in cardiovascular research: a survey of randomized trials. Ann Intern Med. 2008;149(9):612-7. https://doi.org/10.7326/0003-4819-149-9-200811040-00004

MacLean S, Mulla S, Akl EA, Jankowski M, Vandvik PO, Ebrahim S, et al. Patient values and preferences in decision making for antithrombotic therapy: a systematic review: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. Chest. 2012;141(2 Suppl):e1S-e23S. https://doi.org/10.1378/chest.11-2290

Malde S, Umbach R, Wheeler JR, Lytvyn L, Cornu JN, Gacci M, et al. A systematic review of patients' values, preferences, and expectations for the diagnosis and treatment of male lower urinary tract symptoms. Eur Urol. 2021;79(6):796-809. https://doi.org/10.1016/j.eururo.2020.12.019

Mathioudakis AG, Salakari M, Pylkkanen L, Saz-Parkinson Z, Bramesfeld A, Deandrea S, et al. Systematic review on women's values and preferences concerning breast cancer screening and diagnostic services. Psychooncology. 2019;28(5):939-47. https://doi.org/10.1002/pon.5041

Muñoz-Velandia O, Guyatt G, Devji T, Zhang Y, Li SA, Alexander PE, et al. Patient values and preferences regarding continuous subcutaneous insulin infusion and artificial pancreas in adults with type 1 diabetes: a systematic review of quantitative and qualitative data. Diabetes Technol Ther. 2019;21(4):183-200. https://doi.org/10.1089/dia.2018.0346

O'Reilly R, Yokoyama S, Boyle J, Kwong JC, McGeer A, To T, et al. The impact of acute pneumococcal disease on health state utility values: a systematic review. Qual Life Res. 2021. https://doi.org/10.1007/s11136-021-02941-y

Pillay J, Wingert A, MacGregor T, Gates M, Vandermeer B, Hartling L. Screening for chlamydia and/or gonorrhea in primary health care: systematic reviews on effectiveness and patient preferences. Syst Rev. 2021;10(1):118. https://doi.org/10.1186/s13643-021-01658-w

Rand LZ, Kesselheim AS. Controversy over using quality-adjusted life-years in cost-effectiveness analyses: a systematic literature review. Health Aff. 2021;40(9):1402-10. https://doi.org/10.1377/hlthaff.2021.00343

Stacey D, Légaré F, Col NF, Bennett CL, Barry MJ, Eden KB, et al. Decision aids for people facing health treatment or screening decisions. Cochrane Database Syst Rev. 2014(1):Cd001431. https://doi.org/10.1002/14651858.cd001431.pub4

Valli C, Rabassa M, Johnston BC, Kuijpers R, Prokop-Dorner A, Zajac J, et al. Health-related values and preferences regarding meat consumption: a mixed-methods systematic review. Ann Intern Med. 2019;171(10):742-55. https://doi.org/10.7326/m19-1326

Vernooij RWM, Lytvyn L, Pardo-Hernandez H, Albarqouni L, Canelo-Aybar C, Campbell K, et al. Values and preferences of men for undergoing prostate-specific antigen screening for prostate cancer: a systematic review. BMJ Open. 2018;8(9):e025470. https://doi.org/10.1136/bmjopen-2018-025470

Walsh T, Barr PJ, Thompson R, Ozanne E, O'Neill C, Elwyn G. Undetermined impact of patient decision support interventions on healthcare costs and savings: systematic review. BMJ. 2014;348:g188. https://doi.org/10.1136/bmj.g188

Zeng L, Lytvyn L, Wang X, Kithulegoda N, Agterberg S, Shergill Y, et al. Values and preferences towards medical cannabis among people living with chronic pain: a mixed-methods systematic review. BMJ Open. 2021;11(9):e050831. https://doi.org/10.1136/bmjopen-2021-050831

Zhang Y, Morgan RL, Alonso-Coello P, Wiercioch W, Bała MM, Jaeschke RR, et al. A systematic review of how patients value COPD outcomes. Eur Respir J. 2018;52(1). https://doi.org/10.1183/13993003.00222-2018

Research studies

Cox K, Britten N, Hooper R, White P. Patients' involvement in decisions about medicines: GPs' perceptions of their preferences. Br J Gen Pract. 2007;57(543):777-84. http://www.ncbi.nlm.nih.gov/pmc/articles/pmc2151809/

Franklin EF, Nichols HM, Charap E, Buzaglo JS, Zaleta AK, House L. Perspectives of patients with cancer on the quality-adjusted life year as a measure of value in healthcare. Value Health. 2019;22(4):474-81. https://doi.org/10.1016/j.jval.2018.09.2844

McGrath E, O'Conghaile A, Eikelboom JW, Dinneen SF, Oczkowski C, O'Donnell MJ. Validity of composite outcomes in meta-analyses of stroke prevention trials: the case of aspirin. Cerebrovasc Dis. 2011;32(1):22-7. https://doi.org/10.1159/000324629

Other references

Moberg J, Oxman AD, Rosenbaum S, Schunemann HJ, Guyatt G, Flottorp S, et al. The GRADE Evidence to Decision (EtD) framework for health system and public health decisions. Health Res Policy Syst. 2018;16(1):45. https://doi.org/10.1186/s12961-018-0320-2

Vandvik PO, Lincoff AM, Gore JM, Gutterman DD, Sonnenberg FA, Alonso-Coello P, et al. Primary and secondary prevention of cardiovascular disease: Antithrombotic Therapy and Prevention of Thrombosis, 9th ed: American College of Chest Physicians Evidence-Based Clinical Practice Guidelines. Chest. 2012;141(2 Suppl):e637S-e68S. https://doi.org/10.1378/chest.11-2306

3.2d Consider how certain you can be about each advantage and disadvantage.

Explanation

The certainty of the evidence (the extent to which the research provides a good indication of the likely effects of treatments) can affect peoples' treatment choices. For example, someone might decide not to use or to pay for a treatment if the certainty of the evidence is low or very low. How 'certain' the evidence is depends on the fairness of the comparisons, the risk of being misled by the play of chance, and how directly relevant the evidence is. Systematic reviews provide the best basis for these judgements and based on these judgements, should report an assessment of the certainty of the evidence. Unexplained inconsistencies in effect estimates from different studies can also affect the certainty of the evidence.

The use of hydroxychloroquine (HCQ) and chloroquine (CQ) to treat Covid-19 illustrates the importance of considering the certainty of the evidence when making decisions about treatments. On March 28, 2020, the US Food and Drug Administration (FDA) issued a letter granting an Emergency Use Authorization for use of HCQ and HQ for treating Covid-19 *[Thomson 2020]*, and the use of HCQ and HQ surged *[Vaduganathan 2020]*. The letter did not describe the evidence underlying the decision. It stated that the authorisation was supported by recommendations "for treatment of hospitalized COVID-19 patients in several countries, and a number of national guidelines" based on "limited in-vitro and anecdotal clinical data in case series". By June, controlled trials had shown that the FDA guidelines had been misleading – no beneficial effects on morbidity or mortality had been detected. On June 15, the FDA revoked the Emergency Use Authorization. A systematic review published in April 2021 included 14 unpublished trials (1,308 patients) and 14 publications/preprints (9,011 patients) *[Axfors 2021 (SR)]*. It found that HCQ increased deaths in Covid-19 patients, and no benefit of chloroquine had been demonstrated.

Basis for this concept

The certainty of evidence is low for many decisions about treatments. Of 9,451 recommendations in UpToDate, a widely used digital medical textbook, about half were based on low-certainty evidence (see table below) *[Agoritsas 2017 (RS)]*. Most (92%) of the recommendations based on low-certainty evidence were weak recommendations. Weak or conditional recommendations apply to most, but not all patients *[Andrews 2013a]*. Decisions depend on the preferences of patients more than when there is a strong recommendation and require more effort by health professionals to ensure that decisions reflect patients' values (see Concept 3.2c).

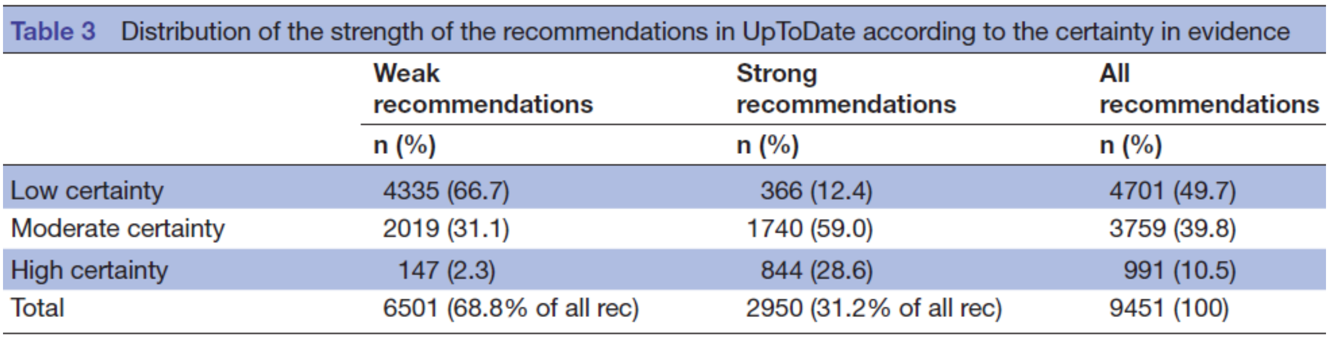


Table 3 Distribution of the strength of the recommendations in UpToDate according to the certainty in evidence

	Weak recommendations	Strong recommendations	All recommendations
	n (%)	n (%)	n (%)
Low certainty	4335 (66.7)	366 (12.4)	4701 (49.7)
Moderate certainty	2019 (31.1)	1740 (59.0)	3759 (39.8)
High certainty	147 (2.3)	844 (28.6)	991 (10.5)
Total	6501 (68.8% of all rec)	2950 (31.2% of all rec)	9451 (100)

Sometimes it is appropriate to make a strong recommendation despite low-certainty evidence *[Andrews 2013b]*. That is, there are some treatment decisions where nearly everyone would make the same choice, despite the uncertainty. That was the case for about 2% of the UpToDate recommendations. Reasons for this include low-certainty evidence that suggests:

- a possible benefit – and high-certainty evidence of harm or high cost

- two treatments may be equivalent – and there is high-certainty evidence of less harm for one of the treatments, or there is high-certainty evidence of equivalence and low-certainty evidence suggests harm for one of the treatments
- a possibility of catastrophic harm – and high-certainty evidence of modest benefits

The reason for about one-third of UpToDate's strong recommendations based on low-certainty evidence was a life-threatening or catastrophic situation when low-certainty evidence suggests benefit. However, as illustrated by the hydroxychloroquine example above, such decisions can sometimes be deadly.

When there is moderate- or high-certainty evidence, different people will nonetheless sometimes make different choices. In UpToDate, 54% of recommendations based on moderate-certainty evidence and 15% of recommendations based on high certainty were weak or conditional recommendations. Often this is because of differences in the relative importance of desirable and undesirable outcomes (see Concept 3.2c). In addition, even when there is high-certainty evidence, there is almost always some uncertainty about who will benefit, who will not, and who will be harmed (see Concept 1.1d). People vary in terms of how risk averse or risk taking they are in relation to the desirable and undesirable effects. Lower comfort with uncertainty has been found to be associated with overutilization of diagnostic tests, but there is sparse evidence of the effects of uncertainty or attitudes towards uncertainty on health professionals' decisions about treatments *[Saposnik 2016 (SR), Tubbs 2006 (SR)]*.

Similarly, a variety of research has addressed how people respond to and deal with uncertainty generally, but relatively little has focused specifically on uncertainty about the effects of treatments. How patients respond to health professional expressions of uncertainty varies *[McGovern 2017 (SR)]*. This may depend on how the uncertainty is communicated, but few studies have investigated this. Although there are recommendations on how to orally communicate uncertainty, most of these lack an evidence base *[Medendorp 2021 (SR)]*.

Uncertainty of the effects of treatments is often inadequately reported in news reports, including uncertainty due to the play of chance (imprecision), the risk of bias, unexplained inconsistencies in effect estimates from different studies, and extrapolation (indirectness of the evidence) *[Oxman 2022 (SR)]*. A systematic review of the effects of uncertainty in public science communication found that most findings of negative effects (such as reduced credibility and beliefs) were from experiments that operationalised uncertainty as disagreement or conflict in science ("consensus uncertainty") *[Gustafson 2020 (SR)]*. Consensus uncertainty was not found to have positive effects. In contrast, uncertainty in the form of quantified error ranges and probabilities ("technical uncertainty") had positive effects, if any, and not negative effects.

Few studies have investigated the impacts of communicating the certainty or quality of the evidence. Two online experiments compared presenting the effect of face shields on reducing the risk of Covid-19 with and without a message that the certainty or quality of the evidence was low *[Schneider 2021 (RS)]*. Participants who were told that the certainty of the evidence was low rated the evidence less trustworthy and rated facemasks as subjectively less effective. When there is a public health emergency, it may be appropriate to persuade people to change their behaviour – for example, to wear facemasks – despite important uncertainties about the potential benefits and harms. However, when there are important uncertainties, they should be acknowledged. Not disclosing uncertainties distorts what is known, inhibits research to reduce important uncertainties, and can undermine public trust in health authorities *[Oxman 2022]*.

Several cognitive biases can affect decisions by both health professionals and patients when there is uncertainty *[Blumenthal-Barby 2015 (SR), Kahneman 2017 , Saposnik 2016 (SR), Tversky 1974 (OR)]*. However, most studies of cognitive biases in healthcare decision making are based on hypothetical scenarios *[Blumenthal-Barby 2015 (SR)]*. The extent to which these biases affect actual decisions is uncertain. Evaluations of interventions to counter cognitive biases suggest that these interventions may be helpful

[Ludolph 2018 (SR)]. Interventions that have been evaluated include cognitive strategies, primarily aimed at improving people's critical thinking skills, and communication strategies, such as providing graphical information in addition to statistical information. The cognitive biases that have most often been targeted in evaluations of these strategies in the context of health-related judgements are "optimism bias" (being overly optimistic) [Chalmers 2006], "framing effects" (choosing among options based on whether they are presented with positive or negative connotations, e.g., as a loss or as a gain) [Akl 2011a (SR)], and base-rate neglect (paying too much attention to numerators and insufficient attention to denominators) [Ludolph 2018 (SR)]. "Relative risk bias" (a stronger inclination to choose a treatment when presented a relative effect than when presented an absolute effect), which is similar to base-rate neglect, has also often been targeted [Akl 2011b (SR)].

More broadly, tolerance of uncertainty has been found to be associated with emotional well-being [Strout 2018 (SR)]. However, the certainty of this evidence is low. Intolerance of uncertainty, on the other hand, may cause anxiety [Rosser 2019 (SR)]. Studies have found a strong association between intolerance of uncertainty and both anxiety and worry in young people [Osmanağaoğlu 2018 (SR)].

Studies have investigated choices made after laboratory-induced stress versus a nonstress condition. A systematic review of those studies found that overall, stress conditions led to decisions that were more disadvantageous, more reward seeking, and more risk taking than nonstress conditions [Starcke 2016 (SR)]. A variety of strategies have been evaluated to help patients and their families manage uncertainty [Zhang 2020 (SR)]. On average, these strategies had small to moderate beneficial effects for both patients and their family members. However, the certainty of this evidence is low. Uncertainty is a ubiquitous concern in health professional education, with students experiencing different forms of uncertainty at many stages of their training. However, strategies that directly support learning around uncertainty are taught infrequently [Moffett 2021 (SR)].

Implications

Consider the certainty of the evidence when choosing treatments.

References

Systematic reviews

Akl EA, Oxman AD, Herrin J, Vist GE, Terrenato I, Sperati F, et al. Framing of health information messages. Cochrane Database Syst Rev. 2011a(12):CD006777. https://doi.org/10.1002/14651858.cd006777.pub2

Akl EA, Oxman AD, Herrin J, Vist GE, Terrenato I, Sperati F, et al. Using alternative statistical formats for presenting risks and risk reductions. Cochrane Database Syst Rev. 2011b(3):CD006776. https://doi.org/10.1002/14651858.cd006776.pub2

Axfors C, Schmitt AM, Janiaud P, Van't Hooft J, Abd-Elsalam S, Abdo EF, et al. Mortality outcomes with hydroxychloroquine and chloroquine in COVID-19 from an international collaborative meta-analysis of randomized trials. Nat Commun. 2021;12(1):2349. https://doi.org/10.1038/s41467-021-22446-z

Blumenthal-Barby JS, Krieger H. Cognitive biases and heuristics in medical decision making: a critical review using a systematic search strategy. Med Decis Making. 2015;35(4):539-57. https://doi.org/10.1177/0272989x14547740

Gustafson A, Rice RE. A review of the effects of uncertainty in public science communication. Public Underst Sci. 2020;29(6):614-33. https://doi.org/10.1177/0963662520942122

Ludolph R, Schulz PJ. Debiasing Health-Related Judgments and Decision Making: A Systematic Review. Med Decis Making. 2018;38(1):3-13. https://doi.org/10.1177/0272989x17716672

McGovern R, Harmon D. Patient response to physician expressions of uncertainty: a systematic review. Ir J Med Sci. 2017;186(4):1061-5. https://doi.org/10.1007/s11845-017-1592-1

Medendorp NM, Stiggelbout AM, Aalfs CM, Han PKJ, Smets EMA, Hillen MA. A scoping review of practice recommendations for clinicians' communication of uncertainty. Health Expect. 2021;24(4):1025-43. https://doi.org/10.1111/hex.13255

Moffett J, Hammond J, Murphy P, Pawlikowska T. The ubiquity of uncertainty: a scoping review on how undergraduate health professions' students engage with uncertainty. Adv Health Sci Educ Theory Pract. 2021;26(3):913-58. https://doi.org/10.1007/s10459-021-10028-z

Osmanağaoğlu N, Creswell C, Dodd HF. Intolerance of Uncertainty, anxiety, and worry in children and adolescents: A meta-analysis. J Affect Disord. 2018;225:80-90. https://doi.org/10.1016/j.jad.2017.07.035

Oxman M, Larun L, Gaxiola GP, Alsaid D, Qasim A, Rose CJ, et al. Quality of information in news media reports about the effects of health interventions: systematic review and meta-analyses. F1000Res. 2022;10:433. https://doi.org/10.12688/f1000research.52894.2

Rosser BA. Intolerance of uncertainty as a transdiagnostic mechanism of psychological difficulties: A systematic review of evidence pertaining to causality and temporal precedence. Cognit Ther Res. 2019;43(2):438-63. https://doi.org/10.1007/s10608-018-9964-z

Saposnik G, Redelmeier D, Ruff CC, Tobler PN. Cognitive biases associated with medical decisions: a systematic review. BMC Med Inform Decis Mak. 2016;16(1):138. https://doi.org/10.1186/s12911-016-0377-1

Starcke K, Brand M. Effects of stress on decisions under uncertainty: A meta-analysis. Psychol Bull. 2016;142(9):909-33. https://doi.org/10.1037/bul0000060

Strout TD, Hillen M, Gutheil C, Anderson E, Hutchinson R, Ward H, et al. Tolerance of uncertainty: A systematic review of health and healthcare-related outcomes. Patient Educ Couns. 2018;101(9):1518-37. https://doi.org/10.1016/j.pec.2018.03.030

Tubbs EP, Elrod JA, Flum DR. Risk taking and tolerance of uncertainty: implications for surgeons. J Surg Res. 2006;131(1):1-6. https://doi.org/10.1016/j.jss.2005.06.010

Zhang Y, Kwekkeboom K, Kim KS, Loring S, Wieben AM. Systematic Review and Meta-analysis of Psychosocial Uncertainty Management Interventions. Nurs Res. 2020;69(1):3-12. https://doi.org/10.1097/nnr.0000000000000368

Other reviews

Tversky A, Kahneman D. Judgment under uncertainty: heuristics and biases. science. Science. 1974;185(4157):1124-31. http://www.jstor.org.ezproxy.uio.no/stable/1738360

Research studies

Agoritsas T, Merglen A, Heen AF, Kristiansen A, Neumann I, Brito JP, et al. UpToDate adherence to GRADE criteria for strong recommendations: an analytical survey. BMJ Open. 2017;7(11):e018593. https://doi.org/10.1136/bmjopen-2017-018593

Schneider CR, Freeman ALJ, Spiegelhalter D, van der Linden S. The effects of quality of evidence communication on perception of public health information about Covid-19: two randomised controlled trials. PLoS One. 2021;16(11):e0259048. https://doi.org/10.1371/journal.pone.0259048

Vaduganathan M, van Meijgaard J, Mehra MR, Joseph J, O'Donnell CJ, Warraich HJ. Prescription fill patterns for commonly used drugs during the Covid-19 pandemic in the United States. JAMA. 2020;323(24):2524-6.

Other references

Andrews J, Guyatt G, Oxman AD, Alderson P, Dahm P, Falck-Ytter Y, et al. GRADE guidelines: 14. Going from evidence to recommendations: the significance and presentation of recommendations. J Clin Epidemiol. 2013a;66(7):719-25. https://doi.org/10.1016/j.jclinepi.2012.03.013

Andrews JC, Schunemann HJ, Oxman AD, Pottie K, Meerpohl JJ, Coello PA, et al. GRADE guidelines: 15. Going from evidence to recommendation-determinants of a recommendation's direction and strength. J Clin Epidemiol. 2013b;66(7):726-35. https://doi.org/10.1016/j.jclinepi.2013.02.003

Chalmers I, Matthews R. What are the implications of optimism bias in clinical research? Lancet. 2006;367(9509):449-50. https://doi.org/10.1016/s0140-6736(06)68153-1

Kahneman D. Thinking, Fast and Slow. New York: Farrar, Straus and Giroux; 2017.

Oxman AD, Fretheim A, Lewin S, Flottorp S, Glenton C, Helleve A, et al. Health communication in and out of public health emergencies: to persuade or to inform? Health Res Policy Syst. 2022;20:28. https://doi.org/10.1186/s12961-022-00828-z

Thomson K, Nachlis H. Emergency Use Authorizations during the Covid-19 pandemic: lessons from hydroxychloroquine for vaccine authorization and approval. JAMA. 2020;324(13):1282-3. https://doi.org/10.1001/jama.2020.16253

3.2e Consider the need for further fair comparisons.

Explanation

There is always some uncertainty about the effects of treatments. If that uncertainty affects decisions that are important to people, the uncertainty should be reduced by further fair comparisons whenever possible. Individuals should consider participating in those fair comparisons when they are uncertain about which alternative to choose because of uncertainty about the effects of the alternatives. Participating in a fair comparison is a good hedging strategy when there is important uncertainty about effects. Moreover, people in fair comparisons sometimes fare better than people outside of fair comparisons. In addition, the results of fair comparisons can help to generate reliable information on which to base future decisions.

Willingness to contribute to the collective good and to help others is commonly thought to be the key motivating factor for participation in randomized trials. However, although willingness to help others might incline people towards participation, participation may be conditional, to some extent, on expectations of personal benefit. For example, a study interviewed people about their motivation to participate in a trial of surgery compared to medical management of gastroesophageal reflux (heartburn and regurgitation caused by stomach contents regurgitating into the oesophagus (tube connecting the mouth and stomach) *[McCann 2010 (RS)]*. It found that people invited to participate viewed:

- recruitment appointments as an opportunity for learning and review,
- participation as potentially offering access or faster access to surgery, and
- participation as offering careful monitoring.

Participants reported that being inclined to help others predisposed them towards trial participation, but considerations of the implications of trial participation for them personally also influenced decisions about participation. For the people who agreed to be randomized, trial participation seemed to be a win-win situation – one in which they could both help others and benefit (or at least not be harmed) personally.

Basis for this concept

A systematic review of factors that affect decisions to participate in randomized trials found 29 studies of experiences of being invited to participate in a trial and of choosing whether to participate *[Houghton 2020 (SR)]*. People were less likely to participate if they were discouraged by other people, felt they had nothing to gain, perceived participation as burdensome, felt they had something to lose, or if there was ineffective trial communication. Conversely, they were more likely to participate if they were encouraged by other people, felt they had something to gain, felt they could help others, felt they had nothing to lose, and if there was effective trial communication. The possible benefits of taking part were key to the decision. Individuals were influenced by the chance of improvement to their health. In addition, many welcomed the opportunity to participate for altruistic reasons or to make a difference by contributing to science.

Some people consider trial participants to be “guinea pigs” *[Sackett 2005]*. Concerns that participants in trials are being “sacrificed” originated from, and are perpetuated by, the examination of single trials or very selective collections of them. Reports of abuse of trial participants are sufficiently publicised that they cause some to question whether randomized trials generally do more harm than good to their participants. However, individual cases and selective reviews are not the best ways to address questions about the benefits and harms of participating in randomized trials.

Several reviews have assessed whether it is beneficial or harmful to participate in randomized trials *[Braunholtz 2001 (SR), Fernandes 2014 (SR), Gross 2006 (SR), Nijjar 2017 (SR), Peppercorn 2004 (SR), Stiller 1994 (SR)]*. Some have compared patients who were treated within trials with those treated outside the trials, regardless of differences between the treatments or between the participants and non-participants. They suggest that participants in trials sometimes have better outcomes than patients outside of trials, and do not have worse outcomes. But it is uncertain whether the results reflect the effects of participating in a randomized trial (trial effects), differences in the treatments in and outside of the trials (treatment effects), or differences between participants and non-participants. A systematic review of studies that compared outcomes in participants who participated in randomized trials with comparable non-participants who received the same or similar treatment found that, on average, participants in randomized trials had similar outcomes to comparable patients who received the same or similar treatments outside the trials *[Vist 2008 (SR), Vist 2005]*. A systematic review of studies that compared patients treated by health professionals or institutions that take part in research found that there may be greater adherence to guidelines and more use of evidence by health professionals and institutions that take part in trials *[Clarke 2011 (SR)]*. However, the consequences for patient health were uncertain.

A common reason for not participating in randomized trials is a strong preference for (or against) one of the treatments being compared *[McCann 2010 (RS)]*. In addition to personal considerations about the pros and cons of participating in a randomized trial, people should only participate in trials if:

- the trial protocol has been registered and made publicly available (see Concept 2.2b),
- the protocol refers to a systematic review showing that the trial is justified (see Concept 2.2a), and
- you receive written assurance that the full study results will be published and sent to all participants who indicate that they wish to receive them (see Concept 2.2b).

In addition, to increase the value of research and reduce waste, new randomized trials should address the needs of users of research (patients, health professionals, and policymakers) and be informed by systematic reviews of existing research *[Chalmers 2014]*.

Implications

Consider advocating for and participating in fair comparisons of treatments when there are important uncertainties about the effects of the treatments.

References

Systematic reviews

Braunholtz DA, Edwards SJ, Lilford RJ. Are randomized clinical trials good for us (in the short term)? Evidence for a "trial effect". J Clin Epidemiol. 2001;54(3):217-24. https://doi.org/10.1016/s0895-4356(00)00305-x

Clarke M, Loudon K. Effects on patients of their healthcare practitioner's or institution's participation in clinical trials: a systematic review. Trials. 2011;12:16. https://doi.org/10.1186/1745-6215-12-16

Fernandes N, Bryant D, Griffith L, El-Rabbany M, Fernandes NM, Kean C, et al. Outcomes for patients with the same disease treated inside and outside of randomized trials: a systematic review and meta-analysis. CMAJ. 2014;186(16):E596-609. https://doi.org/10.1503/cmaj.131693

Gross CP, Krumholz HM, Van Wye G, Emanuel EJ, Wendler D. Does random treatment assignment cause harm to research participants? PLoS Med. 2006;3(6):e188. https://doi.org/10.1371/journal.pmed.0030188

Houghton C, Dowling M, Meskell P, Hunter A, Gardner H, Conway A, et al. Factors that impact on recruitment to randomised trials in health care: a qualitative evidence synthesis. Cochrane Database Syst Rev. 2020;10(10):Mr000045. https://doi.org/10.1002/14651858.mr000045.pub2

Nijjar SK, D'Amico MI, Wimalaweera NA, Cooper N, Zamora J, Khan KS. Participation in clinical trials improves outcomes in women's health: a systematic review and meta-analysis. BjJOG. 2017;124(6):863-71. https://doi.org/10.1111/1471-0528.14528

Peppercorn JM, Weeks JC, Cook EF, Joffe S. Comparison of outcomes in cancer patients treated within and outside clinical trials: conceptual framework and structured review. Lancet. 2004;363(9405):263-70. https://doi.org/10.1016/s0140-6736(03)15383-4

Stiller CA. Centralised treatment, entry to trials and survival. Br J Cancer. 1994;70(2):352-62. https://doi.org/10.1038/bjc.1994.306

Vist GE, Bryant D, Somerville L, Birminghem T, Oxman AD. Outcomes of patients who participate in randomized controlled trials compared to similar patients receiving similar interventions who do not participate. Cochrane Database Syst Rev. 2008(3):MR000009. https://doi.org/10.1002/14651858.mr000009.pub4

Vist GE, Hagen KB, Devereaux PJ, Bryant D, Kristoffersen DT, Oxman AD. Systematic review to determine whether participation in a trial influences outcome. BMJ. 2005;330(7501):1175. https://doi.org/10.1136/bmj.330.7501.1175

Research studies

McCann SK, Campbell MK, Entwistle VA. Reasons for participating in randomised controlled trials: conditional altruism and considerations for self. Trials. 2010;11:31. https://doi.org/10.1186/1745-6215-11-31

Other references

Chalmers I, Bracken MB, Djulbegovic B, Garattini S, Grant J, Gülmezoglu AM, et al. How to increase value and reduce waste when research priorities are set. Lancet. 2014;383(9912):156-65. https://doi.org/10.1016/s0140-6736(13)62229-1

Sackett DL. Participants in research. BMJ. 2005;330(7501):1164. https://doi.org/10.1136/bmj.330.7501.1164

Glossary[†]

Absolute effects	Absolute effects are differences between outcomes in the groups being compared. For example, if 10% (10 per 100) experience an outcome in one of the treatment comparison groups and 5% (5 per 100) experience that outcome in the other group, the absolute effect is 10% - 5% = a 5% difference.
Accuracy	The ability of an outcome measure or diagnostic test to distinguish between people with a health condition and people without it. Diagnostic test accuracy is the proportion of people with a correct diagnosis out of all the people tested.
Allocation	Allocation is the assignment of participants in comparisons of treatments to the different treatment comparison groups.
Allocation bias	Bias resulting from the way participants in a study have been allocated to treatment comparison groups. (Also called selection bias.)
Association or correlation	Association or correlation is a relationship between two attributes, such as using a treatment and experiencing an outcome.
Attrition bias	Systematic differences between treatment comparison groups in withdrawals or exclusions of participants from the results of a study.
Average difference	The average difference is used to express treatment differences for continuous outcomes, such as weight, blood pressure or pain assessed using a scale. It is the difference between the average value for an outcome measure (for example kilograms) in one group and that in a comparison group.
Baseline risk	Baseline risk is an estimate of the likelihood that an individual or group will experience a health problem before a treatment is used.
Bias	A systematic error that may affect the results of a study because of weaknesses in its design, analysis, or reporting.
Blinding	In treatment comparisons, blinding is an action intended to prevent study participants (the people receiving and providing care) or the researchers (or others measuring outcomes) from knowing which participants received which treatment.
Case-control study	A case-control study is a type of non-randomized study comparing the characteristics of people with a particular health condition (cases) with the characteristics of people without that condition (controls), to find what may have caused the problem.
Certainty of the evidence	The certainty of the evidence is an assessment of how good an indication a systematic review provides of the likely effect of a treatment. Judgements about the certainty of the evidence take into account factors that reduce the certainty (risk of bias, inconsistency, indirectness, imprecision, and publication bias) and factors that increase the certainty.
Chance	In the context of comparisons of treatments, chance is responsible for differences between comparison groups that are not due to treatment effects or bias. The play of chance (random error) can lead to incorrect conclusions about treatment effects if too few outcomes occur in studies.
Co-intervention	Treatment, in addition to the treatment being studied, that could impact the outcome of interest.

[†] See GET-IT Glossary for additional plain language definitions and explanations of health research terms.

Cohort study	A cohort study is a type of non-randomized study in which defined groups of people (cohort) are followed up over time to explore the effects of treatments or other factors that may affect health outcomes.
Confidence interval	A confidence interval is a statistical measure of a range within which there is a high probability (usually 95%) that the true value lies. Wide intervals indicate lower confidence; narrow intervals greater confidence.
Confounders	In treatment comparisons, confounders are any factors other than the treatments being compared which may affect the health outcomes being measured.
Contamination	Contamination is the inadvertent application of a treatment allocated to people in one comparison group to people in another comparison group in treatment comparisons.
Continuous outcomes	Continuous outcomes are outcomes (such as measures of pain, weight or depression) which are measured on scales with a potentially infinite number of possible values within a given range (e.g. between “no pain” and “worst pain you have ever experienced”).
Data	Information gathered in studies to help address research questions, such as assessing treatment effects.
Diagnostic test	A procedure used to detect the presence or absence of a health condition.
Dichotomous outcomes	Dichotomous outcomes are yes/no outcomes that people either experience or do not experience.
Effect estimate	An effect estimate is a statistical measure indicating the most likely size of a treatment effect.
Eligibility criteria	Characteristics used to decide whether people are eligible to participate in a study and should be invited to participate.
Evidence	Facts (actual or asserted) intended for use in support of a conclusion.
Explanatory study	An explanatory study (sometimes called an ‘efficacy’ study) is designed to assess the effects of a treatment given in ideal circumstances, in contrast to a ‘pragmatic’ study.
Fair comparison	Fair comparisons of treatments are comparisons designed to minimize the risk of systematic errors (biases) and random errors (resulting from the play of chance).
Follow-up	In treatment comparisons, assessment of study participants after treatment, or the length of time that participants are observed after being allocated to a treatment comparison group.
Incidence	The number of new occurrences of something in a population over a particular period of time.
Indirect comparison	A direct comparison is a head-to-head comparison of treatments within a study. If there are no direct comparisons of the treatments of interest, indirect comparisons – comparisons of people in one study to people in another study.
Intention-to-treat analysis	Analyses based on the outcomes in all the study participants allocated to each of the treatment comparison groups. Intention-to-treat analyses are analyses that include data from all the participants assigned unbiasedly to the treatment comparison groups, whether or not they received the treatment to which they were assigned, even if they never started the treatment, or switched to a different one during the study. Intention-to-treat analyses prevent bias caused by disruption of the baseline equivalence established by random allocation.
Measurement bias	In treatment comparisons, measurement error refers to bias resulting from systematic differences in how outcomes are measured in treatment comparison groups in a study. (Also called detection bias and observer bias.) For outcome measures, measurement error refers to systematic or random error that is not attributable to true changes in the outcome.

Meta-analysis	Statistical combination of estimates derived from two or more similar studies, to give an overall effect estimate.
Model	A representation of the relationship between components of a system. Causal models represent causal relationships in a system, population, or individual.
Nocebo effect	An undesirable effect presumed to act psychologically through suggestion that is or could be caused by information or a treatment believed to be otherwise inactive.
Non-randomized study	A study that does not use random allocation to assign participants to treatment comparison groups.
Outcome	An outcome is a potential benefit or harm of a treatment assessed in a treatment comparison. An outcome measure is how the outcome is assessed in a study.
P-value	A p-value is the probability of observing a result, as extreme or more extreme than the actual result, simply by chance, if in reality there is no treatment difference.
Performance bias	Bias resulting from differences in the care provided to the participants in a study, other than the treatments being compared.
Placebo	A placebo is a treatment that does not contain active ingredients, which has been designed to be indistinguishable from the active treatment being compared with it.
Placebo effect	A measurable, observable, or felt improvement in health or behaviour not attributable to the treatment administered.
Pragmatic study	A pragmatic study (sometimes called an 'effectiveness' study) is designed to assess the effects of a treatment given in the circumstances of everyday practice.
Precision	The extent to which errors resulting from the play of chance affect the results of a study or an outcome assessment are likely to have occurred.
Probability	Probability is the chance or risk of something, such as an outcome, occurring. See Risk.
Propensity score	The probability of being assigned to a particular treatment given a set of observed baseline characteristics.
Protocol	A document providing detailed plans for a study.
Randomized trial	Randomized trials are treatment comparisons in which two or more treatments, possibly including a placebo or withholding a treatment, are compared after random allocation of participants to treatment comparison groups. Random allocation ensures that each participant has a known (usually an equal) chance of being assigned to any of the comparison groups. This results in treatment comparison groups that are similar in terms of prognostic variables, whether or not all of them are known.
Recall bias	Recall bias occurs when study participants are systematically more or less likely to recall and report information on exposure (to a treatment or some other factor) depending on their outcome condition, or to recall information regarding their outcome condition dependent on their exposure.
Regression to the mean	The tendency of unusually large or small measurements of something that fluctuates, such as pain, to return to a more usual or average level on repeated measurements.
Relative effects	Relative effects are ratios. For example, if the probability of an outcome in one treatment comparison group is 10% (10 per 100) and the probability of that outcome in another comparison group is 5% (5 per 100), the relative effect is 5/10 = 0.50.
Reliable	The reliability of a claim or evidence about a treatment effect is the extent to which it is dependable or can be trusted. In the context of research, reliability often has a different meaning, which is the degree to which results obtained by a measurement procedure can be repeated.

Reporting bias	Bias resulting from decisions by researchers, or others (e.g. drug companies or journal editors) not to report or publish the results of a study, or not to provide full information about a study. Publication bias sometimes refers specifically to not publishing a study, and reporting bias sometimes refers specifically to not providing full information, such as not reporting some of the outcomes that were measured in a study.
Risk	Risk is the probability of an outcome occurring. See Probability.
Scale	A scale is a means for measuring or rating an outcome with a potentially infinite number of possible values within a given range, such as weight, blood pressure, pain, or depression.
Statistical significance	Statistical significance is a difference that is unlikely (below a specified level of confidence – typically 5%) to be explained by the play of chance.
Strength of recommendation	The strength of a recommendation is the extent to which people who made the recommendation are confident that the desirable consequences of adhering to the recommendation outweigh the undesirable consequences.
Research study	A research study is an investigation that uses specified methods to evaluate something. Different types of studies can be used to evaluate the effects of treatments, as well as to address other types of questions. Some studies are more reliable than others.
Subgroup	A subgroup is a subdivision of a group of people, a distinct group within a group. For example, in studies or systematic reviews of treatment effects, questions are often asked about whether there are different effects for different subgroups of people in the studies, such as women and men, or people of different ages.
Surrogate outcomes	Surrogate outcomes are outcome measures that are not of direct practical importance but are believed to reflect outcomes that are important. For example, blood pressure is not directly important to patients, but it is often used as an outcome in studies because it is a risk factor for stroke and heart attacks.
Systematic review	A systematic review is a summary of research evidence (studies) that uses systematic and explicit methods to summarise the research. It addresses a clearly formulated question using a structured approach to identify, select, and critically appraise relevant studies, and to collect and analyse data from the studies that are included in the review.
Theory	A theory is a supposition, or a system of ideas intended to explain something.
Treatment	A treatment is any intervention (action) intended to improve health, including preventive, therapeutic and rehabilitative interventions, and public health or health system interventions.
Treatment comparison	Treatment comparisons are studies comparing the effects of treatments.
Treatment comparison group	A treatment comparison group is a group of participants in a study allocated to receive one or more different treatments, usual care, or placebo. Treatment comparison groups are sometimes categorised as treatment groups (intervention groups or experimental groups) and control groups. However, control groups always receive some type of treatment, for example, usual care, a placebo, or active monitoring.
Treatment effects	Treatment effects are changes (increases or decreases) or differences in health outcomes as a result of treatments.
Validity	In treatment comparisons, validity, sometimes specified as internal validity, refers to the extent to which the design and conduct of a study eliminates or reduces bias in the effect estimate. For outcome measures, validity refers to the degree to which an outcome measure measures the construct it purports to measure.