Validation of a questionnaire to measure people’s ability to assess claims about treatment effects

Giordano Pérez-Gaxiola and Astrid Austvoll-Dahlgren

Hospital Pediátrico de Sinaloa “Dr. Rigoberto Aguilar Pico”, Sinaloa, México; Norwegian Institute of Public Health, Oslo, Norway

Abstract

Introduction: Every day, people are faced with claims about treatment effects through mass media, social media, or by word of mouth. Objective: To validate a Spanish-language questionnaire to measure the ability of an individual to assess claims about treatments effects. Method: A set of 22 multiple choice questions taken from the claim evaluation tools database were translated and applied to 172 children and 268 adults. Using a Rasch model, overall and individual item-person fit was explored, as well as the potential item differential functioning (based on gender, age and mode of administration), multidimensionality and local independence. Results: Overall item-person fit was appropriate. Four multiple-choice questions showed a poor fit. Reliability of the questionnaire was satisfactory with a person separation index of 0.7. Multiple-choice questions were unidimensional, and there was no specific dependency. Conclusion: A set of 18 multiple-choice questions with satisfactory fit was obtained. This is the first available questionnaire validated in Spanish to measure individuals’ ability to assess claims about treatment effects.


Introduction

The general population is faced every day with claims about the effects of treatments, either through mass media, social networks or family and friends. These claims may include advice on how to prevent diseases or on the effects of individual, public health or health systems’ therapeutic interventions.1-4 There are several reasons why these claims are made, ranging from good intentions to commercial interests. However, many assertions are unreliable or frankly wrong, and people may suffer unnecessarily or waste resources.5-11 Consequently, helping people to make shared decisions in health by improving their skills to critically evaluate such claims is an important public health initiative.12

Health literacy efforts in Mexico are mainly aimed at providing reliable information through specific websites.13 These programs and sites are directed by the Ministry of Health. The Federal Commission for Protection against Sanitary Risk (COFEPRIS – Comisión Federal para la Protección contra Riesgos Sanitarios) occasionally uses social networks, especially Facebook® and Twitter®, to educate on dubious-quality claims about health. On the other hand, the teaching of evidence-based health care and skills for critically reading scientific literature or claims about treatments is carried out only in some faculties of medicine and does not target the general population. Therefore, there is the need to provide this type of training to patients as well. In addition, to the best of our knowledge, there is no validated instrument available in Spanish to measure the ability of people to critically analyze claims about treatments.

Development and application of the claim evaluation tools database

Although there is a growing number of educational resources to improve people's critical thinking...
regarding claims about treatments, only few have been properly validated. A recent systematic review concluded that these pedagogical interventions have not been consistently measured and that there are no comprehensive instruments to evaluate critical thinking skills. In view of the above, the Claim Evaluation Tools (CET) database, a set of multiple choice questions (MCQ), was developed by the Informed Health Choices (IHC) project (www.informedhealthchoices.org). The IHC project is an international collaboration of researchers from Uganda, Rwanda, Kenya, the United Kingdom, Australia and Norway that has developed educational interventions in order for people to be able to evaluate claims about treatments. The CET database was initially developed to measure the results of randomized trials that were part of the IHC project, but it can only be used in school settings to create questionnaires or to carry out cross-sectional studies to measure critical thinking skills in different populations.

MCQs were developed based on qualitative and quantitative feedback both from experts in methodology and from the general population of several countries. Psychometric analyses of the MCQs have been carried out in Norway and Uganda and they were found to have a valid and reliable construct in those settings, where children, adults, health professionals and people of low resources and low educational level were included.

As a starting point for the development of educational resources and the CET database, the IHC group developed a list of key concepts that people must understand to evaluate claims about treatments. This list serves as a curriculum for researchers and teachers to develop interventions. The list is reviewed and updated annually. The CET database includes between 4 and 6 MCQs for each key concept. These MCQs were written in English and subsequently have been translated into Norwegian, Luganda, German and Chinese. The CET database is hosted on the Testing Treatments interactive website (www.testingtreatments.org), an interactive platform translated into multiple languages and its Spanish version was made known in 2012.

The purpose of this study is to describe the psychometric validation of a set of MCQs that measure the skills of the general population to analyze claims about treatments, translated into Spanish and applied in Culiacán, Sinaloa, a city of approximately 850,000 inhabitants in northwestern Mexico.

### Table 1. Key concepts selected from the Claim Evaluation Tools database and translated for validation in Spanish

<table>
<thead>
<tr>
<th>No.</th>
<th>Concept</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.2</td>
<td>Anecdotes are not reliable evidence.</td>
</tr>
<tr>
<td>1.3</td>
<td>Association is not necessarily causation.</td>
</tr>
<tr>
<td>1.7</td>
<td>Be aware of conflict of interest.</td>
</tr>
<tr>
<td>1.10</td>
<td>Avoid unrealistic expectations.</td>
</tr>
<tr>
<td>2.1</td>
<td>Treatment comparisons are necessary.</td>
</tr>
<tr>
<td>2.2</td>
<td>Comparisons should be made between similar groups.</td>
</tr>
<tr>
<td>2.5</td>
<td>Whenever possible, subjects should not know what treatment they are receiving.</td>
</tr>
<tr>
<td>3.1</td>
<td>Small studies may have misleading results.</td>
</tr>
<tr>
<td>4.1</td>
<td>Single-study can be misleading.</td>
</tr>
<tr>
<td>5.1</td>
<td>Treatments may have both beneficial and harmful effects.</td>
</tr>
<tr>
<td>6.1</td>
<td>Outcomes studies may not be relevant.</td>
</tr>
</tbody>
</table>

### Method

For this study, a pragmatic sample of MCQs encompassing 11 of the 32 original key concepts of the CET database was selected for validation in Spanish. These 32 key concepts were divided in their first version in six groups:

1. Recognizing the need for fair comparisons of treatments.
2. Judging whether a comparison of treatments is a fair comparison.
3. Understanding the role of chance.
4. Considering all the relevant fair comparisons.
5. Understanding the results of fair comparisons of treatments.
6. Judging whether fair comparisons of treatments are relevant.

Since its first publication, and simultaneously with this study, the list of key concepts has been updated. For this study, the numbering of the first publication was used. Updating the key concepts original list (Appendix 1) has no impact on the methods used or on the implications of the results of this study. The selected concepts include the six original groups (Table 1). The most recent version contains 34 key concepts divided in three groups and is available online (http://www.testingtreatments.org/wp-content/uploads/2016/10/Key-Concepts-2nd-edition-with-TTIshort-titles_14122016.pdf).

To allow the elimination of potentially problematic MCQs, two items were selected for each key concept of interest. All key concepts and the MCQs were translated by one of the authors. The translation was...
Table 2. Example of multiple choice questions format

Jorge has stomach ache. The last time Jorge had stomach pain was 2 months ago. That time, he drank a bit of warm milk and after one hour the pain was gone. Therefore, Jorge says that warm milk cures stomach pain.

Question: Is Jorge right?

Options:
A. You cannot know. It is possible that the pain would have disappeared without drinking the milk.
B. You cannot know, but it's probably true based on that Jorge had that experience.
C. Yes, Jorge's experience is enough to demonstrate that warm milk relieves stomach pain.

Answer:

reviewed and updated after feedback from one doctor and one patient. An example of a translated MCQ is shown in Table 2. In addition to the the MCQs, the questionnaire included questions about demographic characteristics such as age, gender, scientific experience (addressed to adults: having previously participated in a clinical trial or having taken an evidence-based medicine course) and socioeconomic/educational level (using the question: how many books do you have at home? as an indirect measurement in children).

Description of participants and administration of questionnaires

The final questionnaire was applied online to adults and to children of 10 to 15 years of age on paper. The online questionnaire online was carried out using QuestBack® and the participants were recruited using social networks (Facebook®, Twitter® and Whatsapp®). The paper questionnaire was applied in two junior high schools in Culiacán, Sinaloa: Escuela Activa Integral, a private school with 350 students, and Federal Junior High School # 2 “General Antonio Rosales Flores”, a public school with 600 students. The schools were selected by convenience from the central zone of the city. A classroom of each one of the three high school grades was randomly selected at each school for administration of the questionnaire. In the literature, there is no consensus on the required sample size according to the methodology of the Rasch model.21 The size was determined based on a pragmatic judgment that takes the number of evaluated items into account, and the statistical power required to evaluate the fit in the Rasch model.

From previous studies, a sample > 300 participants was found to be sufficient when a single questionnaire that includes approximately 22 MCQs is tested, as well as to test the fit and bias according to the participant’s gender or age.18,19

Rasch model

The Rasch model is a dynamic and unified way to evaluate various aspects of measurement in order to validate an instrument, including internal validity of construct (testing multidimensionality), invariance of the measure (item-person interaction) and item bias (differential by item).22,23 It can be used for dichotomous or polytomous data.23,25 Since items in the CET database have a multiple-choice format, they are scored dichotomously. Data were collected in Excel and the Rasch analysis23 was carried out with the RUMM2030® program (RUMM Laboratory).

The structure of the sample class-interval was explored (number and size of groups by skill) and initial analysis of the selected item was performed to explore the item-person interaction.

In a Rasch model, the ratio between two items must be constant between different groups of skills (skills to critically analyze claims about treatments). The response patterns towards an item are compared with that what would be expected according to the model, i.e., the probabilistic form of Guttman's scalogram.23 In other words, the easier the item is the more likely it is for it to be correctly answered, and the more skillful the person the more likely it is for him/her to answer adequately.26

In RUMM2030®, the item-person interaction is presented in measurement units called logits, which express the distance between the parameters of the model. Average item location is always 0. If the instrument has an adequate level of difficulty (neither too easy nor too difficult), the location of people will also be around zero.23 If the person’s location is above 0, the instrument is easy, if it is below 0 it indicates that it is difficult.

The residual item and person fit analyzes the degree of divergence between the observed data and those expected for each person-item when added for all items and persons, respectively. In RUMM2030® this is reported as an approximate z-score, which represents a normal standardized distribution.27 Ideally, item fit or person fit should have an average of 0 and a standard deviation (SD) of 1.23
Person fit was also investigated to identify those with possible poor fit. If all people answer as expected, they will fall within a residual fit of ± 2.5. Excluding people with poor fit from analysis was contemplated, since they might cause bias. Some reasons for poor fit could be that the person guesses, copies or provides “too perfect answers”, such as always selecting option A for all questions of the questionnaire.

Individual item fit in the Rasch model was explored with the chi-square test and residual fit. The MCQs with probabilities of significant chi-square did not fit to the significance level of 0.01 in the model. The MCQs with residual fit within a ± 2.5 range were considered potentially problematic and required further analysis.

When data fit to the Rasch model, capability is consistently measured by means of the trait scale, with a low measurement error. Consequently, fit to the Rasch model implies reliability. In a Rasch analysis this is shown with a separation index and indicates the capability of the construct to differentiate between people with high or low skills. This also indicates that fit statistics can be relied on. A value of 0.7 in the separation index was considered acceptable.

For the identification of MCQs with poor fit, the item characteristic curve (ICC) was used, which indicates the expected probability of correctly answering an item depending on the skill in the latent trait analysis (Fig. 1). The line in the ICC represents the expected scores and points, the scores observed at each class interval. There is good fit when the observed scores follow the expected ones; items with poor fit indicate measurement error. The ICC was reviewed individually for all MCQs searching for misfits.

Item characteristics analysis can also be used to identify the differential item functioning (DIF). Ideally, with exception of the skill to evaluate claims about treatments, CET database MCQs are expected to function equally for both genders and different age groups. Since the versions of the questionnaires of this study were administered to different populations (electronic to adults and on paper to children), the potential DIF for this variable was also explored. There are two types of DIF: the uniform type occurs when the difference between groups in relation to an item is uniform; for example, if adults had greater skills in comparison with children; in the non-uniform type, the difference between skill groups can be inconsistent. MCQs showing a non-uniform DIF were excluded.

In the Rasch model it is assumed that any subset of items administered to a person must provide the same estimate about the skill; therefore, multidimensionality evaluation is essential, which can be made by comparing the person’s location based on two subsets of the questionnaire items: the two subsets of more divergent MCQs within the questionnaire were identified and compared with the person’s location with an independent Student’s t-test.

In addition, local dependence was explored, i.e., to what extent an item influences the answer to other items. This was carried out with the residual correlations function in RUMM2030. A residual correlation between 0.2 and 0.3 above the average of residual correlations of all items was considered potentially problematic. Modifying or eliminating MCQs with poor fit was contemplated.

Results

Two hundred and sixty-eight adult participants were included in the online questionnaire and 172 children between 10 and 15 years in the paper version. All students of the selected classrooms did participate: 99 children of the public school and 73 of the private one; 67 % of adults were between 25 and 45 years of age, 26 % had participated in some clinical trial and
Table 3. Demographic characteristics of participants in the validation of an instrument that explores the skills for assessing claims about medical treatments (web-based and printed formats)

<table>
<thead>
<tr>
<th>Age group (years)</th>
<th>Web-based questionnaire (n = 268)</th>
<th>Printed questionnaire (n = 172)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10-15</td>
<td>4</td>
<td>172</td>
</tr>
<tr>
<td>16-25</td>
<td>26</td>
<td>0</td>
</tr>
<tr>
<td>26-35</td>
<td>88</td>
<td>0</td>
</tr>
<tr>
<td>36-45</td>
<td>91</td>
<td>0</td>
</tr>
<tr>
<td>46-55</td>
<td>26</td>
<td>0</td>
</tr>
<tr>
<td>56-65</td>
<td>27</td>
<td>0</td>
</tr>
<tr>
<td>66-75</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>Female gender</td>
<td>107 (40 %)</td>
<td>86 (50 %)</td>
</tr>
<tr>
<td>Previous participation in a clinical trial</td>
<td>70 (26 %)</td>
<td>NA</td>
</tr>
<tr>
<td>Adults: had taken a course on critical reading or evidence-based medicine</td>
<td>105 (39 %)</td>
<td>NA</td>
</tr>
<tr>
<td>Children: how many books do you have at home?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Many</td>
<td>NA</td>
<td>84 (49 %)</td>
</tr>
<tr>
<td>Some</td>
<td>NA</td>
<td>71 (41 %)</td>
</tr>
<tr>
<td>Few</td>
<td>NA</td>
<td>17 (10 %)</td>
</tr>
</tbody>
</table>

39 % had taken critical reading courses. Demographic characteristics are shown in Table 3.

A class interval of seven groups of skills was identified. Residual item and person fit was satisfactory and close to 0: -0.247 ± 2.083 and -0.0643 ± 0.60, respectively; however, ideally, the standard error for residual item fit should be below 1.4.

Average location of person’s skill was 1.348 logits, which indicated that, in general, participants found the test somewhat easy (they had skills that were higher than MCQs level of difficulty). When the samples were separately assessed, location was better in children, and skill in adults (Figures 2 and 3), probably because a proportion of adults had previously received training on key concepts and critical reading.

Item and person residual fit averages were satisfactory and close to 0, although with higher standard errors than ideally (< 1.4); the person-item map is shown in figure 4, where the upper part represents the participating groups and their level of skill, and bottom part, items’ location and distribution.27 The questionnaire reliability was satisfactory, with a separation index of 0.7.

The person fit analysis showed that almost all items were within the ± 2.5 differential range. When item fit was explored, only one MCQ fell outside the ± 2.5 residual range and the chi-square value reached statistical significance.

When inspecting the MCQ ICCs to determine the discriminating capability of each question, we found that almost all MCQs had a satisfactory fit in the curve, except for two.

DIF was explored by age, gender and questionnaire application mechanism. No MCQ showed DIF for sex. Regarding age, non-uniform DIF was found in two MCQs, and uniform DIF in 4. Six MCQs had DIF for the application mechanism, where the participants of the online version had more skill, i.e., adults, who had had previous training on critical reading.

By means of Student’s t-test, MCQs were found to be satisfactory and unidimensional. In addition, there was no significant local dependence, which suggests that there was no item redundancy.

Therefore, eliminating MCQs with poor adjustment in the ICC curve, with non-uniform DIF, with significant chi-square value and residual fit ± 2.5 was decided (4 MCQs), by means of which a set of 18 multiple-choice questions with satisfactory fit to the Rasch model and covering all selected key concepts was obtained.

Discussion

This study explores the psychometric properties of a questionnaire developed using a select sample of multiple choice questions from the CET database that were translated into Spanish. In general, the questionnaire had a good fit to the model and a satisfactory reliability. Based on the findings of the Rasch analysis, individual MCQs with poor fit were eliminated.

All MCQs of said series can be obtained for non-commercial use, on express request, at the Testing Treatments interactive website (www.testingtreatments.org). Information for the psychometric evaluation in different settings is also available upon request.

There has been interest in validating MCQ sets from the database in different countries. At present, validations are being carried out in China, Germany, the United Kingdom and Norway and there are plans to validate the Spanish translation of the rest of the key concepts, as well as cross-sectional studies in Chile and Spain. This will allow exploring the validity and applicability in other Spanish-speaking regions. Our intention is also to reformulate the four questions with poor fit in future studies.

For this study, we assessed the MCQs we consider to be important the Mexican context, by means of which we covered 11 of the 32 original key
concepts. With the exception of key concepts 1.10 and 6.1, the remaining MCQs were also prioritized in the educational interventions of the IHC project in Uganda. Although these prioritization judgments were made separately, almost identical conclusions were reached.

Judgments about the relevance of each key concept must be made according to the context and MCQs intended use, for example, when evaluating educational tools in a particular population. Some items are considered more basic than others and might be a starting point when the curriculum for an educational intervention is created.

This research is the first initiative to translate and validate the MCQs of this database in Spanish; however, further studies are needed to evaluate the MCQs.
that address the rest of the key concepts, in order to provide researchers, clinicians and teachers with a complete set that serves to perform evaluations in multiple settings in Spanish speaking countries. Currently, there is a lack of evidence about the skills of the general population to evaluate claims about treatments, which can be assessed with a cross-sectional study using the MCOs we validated.

Conclusion

Based on the Rasch model, the questionnaire was found to be reliable; most multiple choice questions showed satisfactory fit and internal validity in the construct. The final product is a set of 18 MCOs that can be used in Mexico and Spanish-speaking countries for validity and applicability studies in other regions, as well as for cross-sectional studies that assess the population’s skills in general, children and adults, to analyze claims about treatments.

Acknowledgements

We are deeply grateful to the children and teachers of Escuela Activa Integral and Escuela Secundaria Federal # 2 “General Antonio Rosales Flores”, as well as to all people who contributed to this project. We also thank Sir Iain Chalmers and Andrew D. Oxman for their support for making this study possible.

References

Appendix 1

Key concepts that people need to understand to analyze claims about treatment effects

1. Recognizing the need for fair comparisons of treatments

Well-informed treatment decisions require reliable information. Not all claims about the effects of treatments are reliable.

<table>
<thead>
<tr>
<th>Concepts</th>
<th>Explanations</th>
<th>Implications</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 Treatments may be harmful</td>
<td>People often exaggerate the benefits of treatments and ignore or downplay potential harms. However, few effective treatments are 100% safe.</td>
<td>Always consider the possibility that a treatment may have harmful effects.</td>
</tr>
<tr>
<td>1.2 Personal experiences or anecdotes (stories) are an unreliable basis for assessing the effects of most treatments</td>
<td>People often believe that improvements in a health problem (e.g. recovery from a disease) was due to 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 got better after receiving a treatment does not mean that the treatment caused the improvement, or that others receiving the same treatment will also improve. The improvement (or undesirable health outcome) might have occurred even without treatment.</td>
<td>Claims about the effects of a treatment may be misleading if they are based on stories about how a treatment helped individual people, or if those stories attribute improvements to treatments that have not been assessed in systematic reviews of fair comparisons.</td>
</tr>
<tr>
<td>1.3 A treatment outcome may be associated with a treatment, but not caused by the treatment</td>
<td>The fact that a treatment outcome (beneficial or harmful) is associated with a treatment does not mean that the treatment caused the outcome. 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 of their being healthier and having better living conditions, rather than because of the treatment.</td>
<td>Unless other reasons for an association between an outcome and a treatment have been ruled out by a fair comparison, do not assume that the outcome was caused by the treatment.</td>
</tr>
<tr>
<td>1.4 Widely used treatments or treatments that have been used for a long time are not necessarily beneficial or safe</td>
<td>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.</td>
<td>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.</td>
</tr>
<tr>
<td>1.5 New, brand-named, or more expensive treatments may not be better than available alternatives</td>
<td>New treatments are often assumed to be better simply because they are new or because they are more expensive. However, 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 it may not be possible to know whether they will appear without long-term follow-up.</td>
<td>A treatment should not be assumed to be beneficial and safe simply because it is new, brand-named or expensive.</td>
</tr>
<tr>
<td>1.6 Opinions of experts or authorities do not alone provide a reliable basis for deciding on the benefits and harms of treatments</td>
<td>Doctors, researchers, patient organizations and other authorities often disagree about the effects of treatments. This may be because their opinions are not always based on systematic reviews of fair comparisons of treatments.</td>
<td>Do not rely on the opinions of experts or other authorities about the effects of treatments, unless they clearly base their opinions on the findings of systematic reviews of fair comparisons of treatments.</td>
</tr>
<tr>
<td>Concepts</td>
<td>Explanations</td>
<td>Implications</td>
</tr>
<tr>
<td>-------------------------------------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>1.7 Conflicting interests may result in misleading claims about the effects of treatments</td>
<td>People with an interest in promoting a treatment (in addition to wanting to help people), such as making money, may promote treatments by exaggerating benefits and ignoring potential harmful effects. Conversely, people may be opposed to a treatment for a range of reasons, such as cultural practices.</td>
<td>Ask if people making claims that a treatment is effective have conflicting interests. If they have conflicting interests, be careful not to be misled by their claims about the effects of treatments.</td>
</tr>
<tr>
<td>1.8 Increasing the amount of a treatment does not necessarily increase the benefits of a treatment and may cause harm</td>
<td>Increasing the dose or amount of a treatment (e.g. how many vitamin pills you take) often increases harms without increasing beneficial effects.</td>
<td>If a treatment is believed to be beneficial, do not assume that more of it is better.</td>
</tr>
<tr>
<td>1.9 Earlier detection of disease is not necessarily better</td>
<td>People often assume that early detection of disease leads to better outcomes. However, screening people to detect 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 tests can be inaccurate (e.g. misclassifying people who do not have disease as having disease). Screening can also cause harm by labelling people as being sick when they are not and because of side effects of the tests and treatments.</td>
<td>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.</td>
</tr>
<tr>
<td>1.10 Hope or fear can lead to unrealistic expectations about the effects of treatments</td>
<td>Hope can be a good thing, but sometimes people in need or desperation hope that treatments will work and assume they cannot do any harm. Similarly, fear can lead people to use treatments that may not work and can cause harm. As a result, they may waste time and money on treatments that have never been shown to be useful, or may actually cause harm.</td>
<td>Do not assume that a treatment is beneficial or safe, or that it is worth whatever it costs, simply because you hope that it might help.</td>
</tr>
<tr>
<td>1.11 Beliefs about how treatments work are not reliable predictors of the actual effects of treatments</td>
<td>Treatments that should work in theory often do not work in practice, or may turn out to be harmful. An explanation of how or why a treatment might work does not prove that it works or that it is safe.</td>
<td>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.</td>
</tr>
<tr>
<td>1.12 Large, dramatic effects of treatments are rare</td>
<td>Large effects (where everyone or nearly everyone treated experiences a benefit or a harm) are easy to detect without fair comparisons, but few treatments have effects that are so large that fair comparisons are not needed.</td>
<td>Claims of large effects are likely to be wrong. Expect treatments to have moderate, small or trivial effects, rather than dramatic effects. Do not rely on claims of small or moderate effects of a treatment, which are not based on systematic reviews of fair comparisons of treatments.</td>
</tr>
</tbody>
</table>
2. Judging whether a comparison of treatments is a fair comparison

Well-informed treatment decisions require fair comparisons of treatments; i.e., comparisons designed to minimize the risk of errors. Not all comparisons of treatments are fair comparisons.

<table>
<thead>
<tr>
<th>Concepts</th>
<th>Explanations</th>
<th>Implications</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 Evaluating the effects of treatments requires appropriate comparisons</td>
<td>If a treatment is not compared to something else, it is not possible to know what would happen without the treatment, so it is difficult to attribute outcomes to the treatment.</td>
<td>Always ask what the comparisons are when considering claims about the effects of treatments. Claims that are not based on appropriate comparisons are not reliable.</td>
</tr>
<tr>
<td>2.2 Apart from the treatments being compared, the comparison groups need to be similar</td>
<td>If people in the 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 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.</td>
<td>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 were 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.</td>
</tr>
<tr>
<td>2.3 People's experiences should be counted in the group to which they were allocated</td>
<td>Randomized allocation helps to ensure that the treatment and comparison groups share similar characteristics. However, people sometimes do not receive or take the allocated treatments. The characteristics of such people often differ from those who do. Therefore, excluding from the analysis people who did not receive the allocated treatment may mean that like is no longer being compared with like.</td>
<td>Be cautious about relying on the results of treatment comparisons if patients' outcomes are not counted in the group to which they were allocated. 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.</td>
</tr>
<tr>
<td>2.4 People in the groups being compared need to be cared for similarly (apart from the treatments being compared)</td>
<td>Apart from the treatments being compared, people in the treatment comparison groups should otherwise receive similar care. If, for example, people in one group receive more attention and care than people in the comparison group, differences in outcomes could be due to differences in the amount of attention each group received rather than due to the treatments that are being compared. One way of preventing this is to keep providers unaware (&quot;blind&quot;) of which people have been allocated to which treatment.</td>
<td>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 could be misleading.</td>
</tr>
<tr>
<td>2.5 If possible, people should not know which of the treatments being compared they are receiving</td>
<td>People in a treatment group may experience improvements (for example, less pain) because they believe they are receiving a better treatment, even if the treatment is not actually better (this is called a placebo effect), or because they behave differently (due to knowing which treatment they received, compared to how they otherwise would have behaved). If individuals know that they are receiving (they are not “blinded” to) a treatment that they believe is better, some or all of the apparent effects of the treatment may be due either to a placebo effect or because the recipients behaved differently.</td>
<td>Be cautious about relying on the results of treatment comparisons if the participants knew which treatment they were receiving; this may have affected their expectations or behavior. The results of such comparisons could be misleading.</td>
</tr>
</tbody>
</table>
2.6 Outcomes should be measured in the same way (fairly) in the treatment groups being compared

If an outcome is measured differently in two comparison groups, differences in that outcome may be due to how the outcome was measured rather than because of the treatment 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 observe better outcomes in those who have received the treatment. One way of preventing this is to keep outcome assessors unaware (“blind”) of which people have been allocated to which treatment.

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

2.7 It is important to measure outcomes in everyone who was included in the treatment comparison groups

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

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 percentages of people lost to follow-up. The results of such comparisons could be misleading.

3. Understanding the role of chance

Well-informed treatment decisions require information about the risk of being misled by the role played by chance.

3.1 Small studies in which few outcome events occur are usually not informative and the results may be misleading

When there are only 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 between the treatments.

Be cautious about relying on the results of treatment comparisons with few outcome events. The results of such comparisons could be misleading.

3.2 The use of p-values to indicate the probability of something having occurred by chance may be misleading; confidence intervals are more informative

The observed difference in an outcome is the best estimate of how effective or safe a treatment is (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. The confidence interval is the range within which the true difference is likely to lie, after taking into account the play of chance. Although a confidence interval (margin of error) is more informative than a p-value, the latter is often reported. P-values are often misinterpreted to mean that treatments have or do not have important effects.

Understanding a confidence interval may be necessary to understand the reliability of an estimated treatment effect. Whenever possible, consider confidence intervals when assessing estimates of treatment effects. Do not be misled by p-values.

3.3 Saying that a difference is statistically significant or that it is not statistically significant can be misleading

Statistical significance is often 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 effects of the treatment are uncertain).

Claims that results were significant or non-significant usually mean that they were not statistically significant or non-significant. This is not the same as important or not important. Do not be misled by such claims.
4. Considering all of the relevant fair comparisons

Well-informed treatment decisions require systematic reviews of the evidence. Non-systematic summaries can be misleading.

<table>
<thead>
<tr>
<th>Concepts</th>
<th>Explanations</th>
<th>Implications</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1 The results of single comparisons of treatments can be misleading</td>
<td>A single comparison of treatments rarely provides conclusive evidence and results are often available from other comparisons of the same treatments. These other comparisons may have different results or may help to provide more reliable and precise estimates of the effects of treatments.</td>
<td>The results of single comparisons of treatments can be misleading.</td>
</tr>
<tr>
<td>4.2 Reviews of treatment comparisons that do not use systematic methods can be misleading</td>
<td>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 some studies may be biased, or the synthesis of the results of the selected studies may be inadequate or inappropriate.</td>
<td>Whenever possible, use systematic reviews of fair comparisons rather than non-systematic reviews of fair comparisons of treatments to inform your decisions.</td>
</tr>
<tr>
<td>4.3 Well done systematic reviews often reveal a lack of relevant evidence, but they provide the best basis for making judgements about the certainty of the evidence</td>
<td>The certainty of the evidence (the extent to which the research provides a good indication of the likely effects of treatments) can affect the treatment decisions people make. 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 should report an assessment of the certainty of the evidence based on these judgements.</td>
<td>When using the findings of systematic reviews to inform your decisions, always consider the degree of certainty of the evidence.</td>
</tr>
</tbody>
</table>

5. Understanding the results of fair comparisons of treatments

Well-informed treatment decisions require information about the size of effects. Research results may be presented in misleading ways.

<table>
<thead>
<tr>
<th>Concepts</th>
<th>Explanations</th>
<th>Implications</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1 Treatments usually have beneficial and harmful effects</td>
<td>Because treatments can have harmful effects as well as beneficial effects, decisions should be informed by the balance between the benefits and harms of treatments. Costs also need to be considered.</td>
<td>Always consider the trade-offs between the potential benefits of treatments and the potential harms and costs of treatments.</td>
</tr>
<tr>
<td>5.2 Relative effects of treatments alone can be misleading</td>
<td>Relative effects (e.g., the ratio of the probability of an outcome in one treatment group compared with that in a comparison group) are insufficient for judging the importance of the difference (between the probabilities of the outcome). A relative effect may give the impression that a difference is larger 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 a stroke by 50% but also has harms, and your risk of getting a stroke is 2 in 100, receiving the treatment is likely to be worthwhile. If, however, your risk of getting a stroke is 2 in 10,000, then receiving the treatment is unlikely to be worthwhile even though the relative effect is the same.</td>
<td>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.</td>
</tr>
</tbody>
</table>
6. Judging whether fair comparisons of treatments are relevant

Well-informed treatment decisions require relevant information. The results of fair comparisons may not be relevant to a particular individual.

<table>
<thead>
<tr>
<th>Concepts</th>
<th>Explanations</th>
<th>Implications</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1 Fair comparisons of treatments should measure outcomes that are important</td>
<td>Patients, professionals and researchers may have different views about which outcomes are important. Studies often measure outcomes, such as heart rhythm irregularities, as surrogates for important outcomes, like death after heart attack. However, the effects of treatments on surrogate outcomes often do not provide a reliable indication of the effects on outcomes that are important.</td>
<td>Do not be misled by surrogate outcomes.</td>
</tr>
<tr>
<td>6.2 Fair comparisons of treatments in animals or highly selected groups of people may not be relevant</td>
<td>Systematic reviews of studies that only include animals or a selected minority of people are unlikely to provide results that are relevant to most people.</td>
<td>Results of systematic reviews of studies in animals or highly-selected groups of people may be misleading.</td>
</tr>
<tr>
<td>6.3 The treatments evaluated in fair comparisons may not be relevant or applicable</td>
<td>A fair comparison of the effects of a surgical procedure done in a specialized hospital may not provide a reliable estimate of the effects and safety of the same procedure performed in other settings. 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 good estimate of how the new drug compares to what is commonly done.</td>
<td>Be aware that your circumstances may be sufficiently different that the results of systematic reviews of fair comparisons of treatments may not apply to you.</td>
</tr>
<tr>
<td>6.4 Results for a selected group of people within fair comparisons can be misleading</td>
<td>Comparisons of treatments often report results for a selected group of participants in an effort 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 results’ are likely to be due to the play of chance and are unlikely to reflect true differences.</td>
<td>Findings based on results for subgroups of people within a treatment comparison may be misleading.</td>
</tr>
</tbody>
</table>
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.

**Allocation**
Allocation is the assignment of participants in comparisons of treatments to the different treatments (groups) being compared.

**Association**
Association is a relationship between two attributes, such as using a treatment and experiencing an outcome.

**Average difference**
The average difference is used to express treatment differences for continuous outcomes, such as weight, blood pressure or pain measured on 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.

**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; i.e., the likelihood that the effect will be substantially different from what the studies found (different enough that it might affect a decision). Judgements about the certainty of the evidence are based on 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 the occurrence of differences between comparison groups that are not due to treatment effects or bias. The role played by chance can lead to incorrect conclusions about treatment effects if too few outcomes occur in studies.

**Confidence interval**
A confidence interval is a statistical measure of a range within which there is a high probability (usually 95%) that the actual value lies. Wide intervals indicate lower confidence; narrow intervals greater confidence.

**Fair comparison**
Fair comparisons of treatments are comparisons designed to minimize the risk of systematic errors (biases) and random errors (resulting from the role played by chance).

**Outcome**
An outcome is a potential benefit or harm of a treatment measured in a treatment comparison. An outcome measure is how the outcome is measured in a study.

**p-value**
The p-value is the probability (ranging from zero to one) that the results observed in a study (or results more extreme) could have occurred by chance if in reality there were no treatment differences.

**Placebo**
A placebo is a treatment that does not contain active ingredients, which has been designed to be indistinguishable from the active treatment being assessed.

**Placebo effect**
A measurable, observable, or felt improvement in health or behavior not attributable to the treatment administered.

**Probability**
Probability is the chance or risk of something, such as an outcome, occurring. See Risk.

**Relative effects**
Relative effects are ratios. For example, if the probability of an outcome in the treatment group is 10% (10 per 100) and the probability of that outcome in a comparison group is 5% (5 per 100), the relative effect would be 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. It should be noted that reliability often has a different meaning in the context of research, which is the degree to which results obtained by a measurement procedure can be replicated.

**Risk**
Risk is the probability of an outcome to occur. See Probability.

**Scale**
A scale is an instrument 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 role played by chance.

**Study**
A study is an investigation that uses specified methods to evaluate something. Different types of studies can be used to evaluate the effects of treatments. Some 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 or indirect outcomes**  
A surrogate or indirect outcome is an outcome measure that is not of direct practical importance but is 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 heart attacks or stroke.

**Systematic review**  
A systematic review is a summary of research evidence (studies) that uses systematic and explicit methods to summarize the research. These reviews address a clearly formulated question using a structured approach to identify, select, and critically appraise relevant studies, and to collect and analyze 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 of the effects of treatments.