Aggregation of Clinical Evidence using Argumentation: A Tutorial Introduction

Anthony Hunter[1] and Matthew Williams[2]

[1] Department of Computer Science, University College London, London, UK
[2] Department of Oncology, Charing Cross Hospital, London, UK

Abstract. In this tutorial, we describe a new framework for representing and synthesizing knowledge from clinical trials involving multiple outcome indicators. The framework offers a formal approach to aggregating clinical evidence. Based on the available evidence, arguments are generated for claiming that one treatment is superior, or equivalent, to another. Evidence comes from randomized clinical trials, systematic reviews, meta-analyses, network analyses, etc. Preference criteria over arguments are used that are based on the outcome indicators, and the magnitude of those outcome indicators, in the evidence. Meta-arguments attack (i.e they are counterarguments to) arguments that are based on weaker evidence. An evaluation criterion is used to determine which are the winning arguments, and thereby the recommendations for which treatments are superior. Our approach has an advantage over meta analyses and network analyses in that they aggregate evidence according to a single outcome indicator, whereas our approach combines evidence according to multiple outcome indicators.

1 Introduction

Evidence-based decision making is well established in medicine. However, the scale and pace of new evidence makes it difficult for clinicians and researchers to acquire and assimilate that evidence. As a consequence, understanding and reviewing the literature is difficult and time-consuming. This problem is exacerbated by the fact that the evidence is uncertain, incomplete and inconsistent. In this tutorial, we describe a new framework for aggregating evidence from clinical trials. This provides a systematic, transparent, and robust process that operates over multiple outcome indicators. The formal presentation of our framework has been presented in [1], but given the novelty of our approach can seem forbidding for a non-technical audience. So with this tutorial, we provide a more accessible introduction for clinical and scientific readers interested in reasoning with clinical evidence. We assume the the reader has some basic familiarity with clinical trials, in particular randomised clinical trials.

2 Motivation

To cope with the problems of volume, complexity, inconsistency and incompleteness of evidence, organizations supporting decision makers, such as the UK National Institute for Clinical Excellence, (NICE, www.nice.org.uk), compile and aggregate evidence into evidence-based guidelines for decision makers. Such guidelines systematically appraise available evidence so as to encode best-practice *recommendations*. These typically specify what tests should be done, and what treatments should be considered, for particular classes of patient. The advice is supported by reference to the primary literature (such as published randomized clinical trials, cohort studies, etc), together with available systematic reviews of evidence, such as by the Cochrane Collaboration (www.cochranecollaboration.org).

As valuable as guidelines are for drawing the best available evidence into decision making in healthcare, there are some important limitations.

1. Constructing guidelines can involve **assimilating massive amounts of evidence**. For instance, medical guidelines are based on a rapidly growing body of biomedical evidence, such as clinical trials and other scientific studies (for example, PubMed, the online repository of biomedical abstracts run by the US National Institute of Health has over 20 million articles). Production of evidence-based guidelines therefore requires **considerable human effort and expenditure** since the evidence needs to be systematically reviewed and aggregated.
2. Guidelines can become **out-of-date** quite quickly. For example, in medicine, even when major trials are published on topics, it may take years before the guidelines are rewritten to take account of the large amounts of newly available evidence (for example, PubMed is growing at the rate of 2 articles per minute). Decision makers are thus faced with the problem of assimilating and processing guidelines in combination with large amounts of newly available evidence which may warrant recommendations that conflict with, and so suggest revisions to, those recommendations provided by the guidelines.
3. Often there are **overlapping guidelines** to consider (from different agencies or bodies, and international, national, and local sources), and when there are multiple problems to be resolved (e.g. a patient with both cancer and liver problems). Thus, different guidelines may offer conflicting guidance.
4. Guideline recommendations are often written keeping in mind a **general population** so they need to be interpreted for individual cases with specific features. For example, given a patient with some particular symptoms and test results, the clinician needs to decide if the patient falls into any of the classes of patients for which the guideline offers guidance (e.g. if the patient is from a particular ethnic group, or if they are very young, or if their symptoms do not exactly correspond). If the clinician has doubts, then turning to the primary literature for fuller descriptions of the relevant clinical trials may be useful. However, the clinician may then need to assimilate and aggregate the results from a number of articles which can be challenging. So after what

may be an incomplete study of the evidence, the clinician decides whether or not to accept the recommendation from the guideline for the specific case.
5. Guidelines are **not sensitive to local needs** or circumstances. This may also result in non-compliance by the decision maker in using a guideline. For example, an international guideline may recommend a particular kind of scan for patients with a particular combination of symptoms, but a particular hospital using the guideline might not be able to provide such a scan, and would deviate from the recommendations by the guideline.
6. Use of guidelines can **decouple a decision maker from the evidence** which can be problematical since the decision maker may have valuable knowledge and experience for use in interpreting the evidence.

These shortcomings suggest that there is a need for knowledge aggregation technologies for making evidence-based recommendations based on large repositories of complex, rapidly expanding, incomplete and inconsistent evidence. These technologies should aim to overcome the limitations of guidelines listed above, and offer tools for users who need to make evidence-based decisions, as well as users who need to draft systematic reviews and guidelines, and users who need to undertake research in order to fill gaps or resolve conflicts in the available evidence.

3 Argument-based evidence aggregation

In this section, we provide some background to our approach. We consider the kind of input we assume, and we briefly discuss what we mean by argumentation.

3.1 Input to our aggregation process

We concentrate on clinical trials that compare two different treatments (i.e. "two-armed" trials"), but where different trials may measure and report different outcome indicators.

Consider two treatments τ_1 and τ_2 for some heart condition. These may be compared on their efficacy in treating the condition, and on their side-effects. For example, we may have evidence from a trial that compares treatment τ_1 with τ_2 on the relative risk of mortality within 5 years is 0.95 (i.e. the risk of mortality with τ_1 is 0.95 of that with τ_2), and we may have evidence from a trial that compares treatment τ_2 with τ_1 on the relative risk of causing drowsiness is 0.5 (i.e. the risk of drowsiness with τ_2 is 0.5 of that with τ_1). Our framework takes this evidence as input, and determines which treatment is superior. In order to do this, we need to also take into account preferences (of clinicians or patients) over the outcome indicators and their magnitude.

- (Option 1) The relative risk of mortality within 5 years is 0.95 (if taking τ_1 instead of τ_2)
- (Option 2) The relative risk of causing drowsiness is 0.5 (if taking τ_2 instead of τ_1)

These preferences may vary from person to person. For some people, even a modest reduction in the risk of mortality is preferred to a reduced risk of drowsiness, and therefore they would prefer option 1, whereas for other people (e.g. HGV drivers), the risk of drowsiness would be problematical, and they would therefore prefer option 2. Whilst such preferences are subjective, once we have captured them we can use them systematically when aggregating evidence with multiple outcome indicators.

So to summarize, the input to our aggregation process is the evidence concerning pairwise comparisons of treatments, and the preferences over outcome indicators (and their magnitude) that appear in the evidence. Note, in Section 3, we consider how to consider different choices of preference when we do not have a specific preference.

3.2 Our aggregation process is based on argumentation

Argumentation is an important cognitive activity for handling incomplete and inconsistent information. It involves identifying individual arguments and counterarguments, and it may involve identifying winning arguments. For example, diagnosis involves argumentation. There may be competing diagnoses for a patient. For each diagnosis, there may be one or more arguments that support it. Furthermore, there may counterarguments to some of these arguments (perhaps based on conflicting results from tests, or other reasons to doubt individual diagnoses). Deciding on which is the diagnosis for the patient can be regarded as a process of deciding on which arguments win.

In recent years, there has been substantial interest in developing theoretical and computational models of argument that can be used in diverse applications (for a review, see [2]). In theoretical models of argument, each argument has a formally specified claim, and some specified premises from which the claim can be derived using some formal reasoning process. For example, consider the following premises

```
The shape is square
If the shape is square, then the shape has four sides
```

From these premises, we have the claim "`The shape has four sides`" by logical reasoning (syllogism). Hence, we can construct an argument with these premises and claim.

A counterargument is an argument that contradicts the premises or claim of an argument. So a counterargument is an argument that "attacks" another argument. For example, from the premise that "`The shape is triangular`", we could construct a counterargument to the above argument.

```
The shape is triangular
If the shape is triangular, then the shape does not have four sides
```

So that claim of the second argument contradicts the claim of the first argument, and so the second argument is a counterargument to the first argument.

Furthermore, the claim of the first argument contradicts the claim of the second argument, and so the first argument is a counterargument to the second argument. So each argument attacks the other in this example.

Argumentation is useful when there is uncertainty in the information available. Here for instance, it may be that there is uncertainty about the shape of the observed object. One source believes it is square and the other source believes it is triangular.

Different formalisms for argumentation provide different ways of formalizing arguments and counterarguments, and for deciding on which arguments win. We do not provide a review of the field in this tutorial. Rather, we just outline (in the next section) the notions we require for our framework. However, what is common amongst these formalisms is that they provide an explicit representation of the conflicts arising in the available information, and that they provide principled ways of deciding what are winning arguments.

4 Step-by-step tutorial on our approach

In this section, we provide an introduction to our process for aggregating evidence. We do this in seven steps starting with the representation of the set of evidence as input (at Step 1) and a decision on which treatment is superior as output at (Step 7).

4.1 Tabulating the evidence (Step 1)

We start with a set of 2-arm superiority trials, i.e., clinical trials whose purpose is to determine whether, given two treatments, one is superior to the other. Each trial will typically report more than one outcome (perhaps a measure of effectiveness, and a measure of a side-effect). We collect these as an evidence table. Each row represents data about the trial and a single outcome; thus each trial may generate more than one row. The columns of the table depend on the particular trial, but we assume the following columns as a minimum for an evidence table. We give an example of an evidence table in Example 1.

- The **left** and **right** attributes signify the treatments compared in each item of evidence (i.e. the left and right arms of the trial for each item of evidence).
- The **outcome indicator** attribute is the specification of the particular outcome that is being considered when comparing the two treatments. For example, it could be the relative risk of mortality.
- The **outcome value** attribute is the value obtained for the outcome indicator for the left arm compared to the right arm. For example, if the outcome indicator is relative risk of mortality, then it would be the value obtained for the left arm compared to the right arm.
- The **net outcome** attribute is a binary relation over the two treatments that is determined from the value of the outcome and an evaluation of whether the outcome indicator is desirable or undesirable for the patient class. In

this tutorial, we consider outcome indicators that are evaluated in terms of relative risk. In this case, there are four possibilities for this.

1. If the outcome indicator is something that we want to decrease, and the outcome value is less than 1, then the left arm is superior is to the right arm, and so the net outcome is "superior".
2. If the outcome indicator is something that we want to decrease, and the outcome value is greater than 1, then the left arm is inferior is to the right arm, and so the net outcome is "inferior".
3. If the outcome indicator is something that we want to increase, and the outcome value is less than 1, then the left arm is inferior is to the right arm, and so the net outcome is "inferior".
4. If the outcome indicator is something that we want to increase, and the outcome value is greater than 1, then the left arm is superior is to the right arm, and so the net outcome is "superior".

For example, if the outcome indicator is relative risk of mortality, and the value is below 1, then the net outcome is desirable, and so the the left arm is superior to the right arm. Whereas, if the outcome indicator is relative risk of mortality, and the value is above 1, then the net outcome is undesirable, and so the the left arm is inferior to the right arm.

The set of attributes we have discussed here is the minimum that we require. There are numerous other optional attributes that are useful for assessing and aggregating evidence, such as the following, and so each such attribute could be captured as a further column in the evidence table (depending on the kind of evidence available and how it might be regarded).

- the p-value for the study
- the number of patients involved in each trial
- the geographical location for each trial
- the drop-out rate for the trial
- the methods of randomization
- the evidence type (meta-analysis, cohort study, network analysis, etc)

For a general introduction to the nature of clinical trials, and a discussion of a wider range of attributes, see [3].

Example 1. For our running example, we will use the following evidence table. There are four items of evidence e_1 to e_4. For each item of evidence, the left arm is CP (standing for contraceptive pill) and the right arm is NT (standing for no treatment). For e_1, the outcome indicator is relative risk of pregnancy, for e_2, the outcome indicator is relative risk of ovarian cancer, for e_3, the outcome indicator is relative risk of breast cancer, and for e_4, the outcome indicator is relative risk of deep vein thrombosis (DVT). There is one optional column in this evidence table which is the p value for the RCT in each item of evidence.

ID	Left	Right	Outcome indicator	Outcome value	Net outcome	p
e1	CP	NT	pregnancy	0.05	superior	0.01
e2	CP	NT	ovarian cancer	0.99	superior	0.07
e3	CP	NT	breast cancer	1.04	inferior	0.01
e4	CP	NT	DVT	1.02	inferior	0.05

4.2 Generation of structured arguments (Step 2)

From the input evidence, a particular kind of argument that we call an structured argument is generated. Each structured argument is a pair $\langle X, \epsilon \rangle$ where X is a subset of the evidence concerning two treatments τ_1 and τ_2. If all the evidence in X indicates that τ_1 is better in some respects than τ_2 (i.e. for the evidence in X, the net outcome is superior), then the claim ϵ is that τ_1 is superior to τ_2. Whereas if all the evidence in X indicates that τ_2 better in some respects to τ_1, then the claim ϵ is that τ_1 is inferior to τ_2 (i.e. for the evidence in X, the net outcome is inferior). And if all the evidence in X indicates that τ_2 equal in some respects to τ_1, then the claim ϵ is that τ_1 is equal to τ_2 (i.e. for the evidence in X, the net outcome is equal). Note, we assume the evidence in an argument is homogeneous in the sense that X only contains evidence that indicates τ_1 better in some respects to τ_2, or X only contains evidence that indicates τ_1 equal in some respects to τ_2, or X only contains evidence that indicates τ_2 better in some respects to τ_1

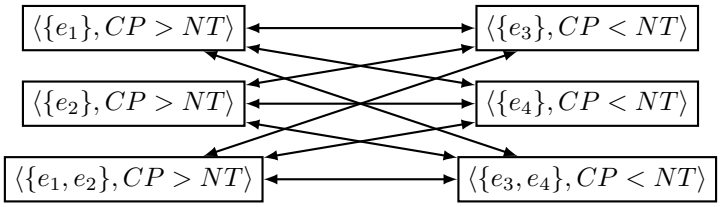
Example 2. Continuing Example 1, we have six structured arguments. Given two items of evidence that support the claim $CP > NT$, we get three arguments with the claim $CP > NT$. Similarly given two items of evidence that support the claim $CP < NT$, we get three arguments with the claim $CP < NT$

$$\begin{array}{cc} \langle \{e_1\}, CP > NT \rangle & \langle \{e_3\}, CP < NT \rangle \\ \langle \{e_2\}, CP > NT \rangle & \langle \{e_4\}, CP < NT \rangle \\ \langle \{e_1, e_2\}, CP > NT \rangle & \langle \{e_3, e_4\}, CP < NT \rangle \end{array}$$

Each of the arguments on the left provides the case for the claim that τ_1 is superior to τ_2, and each of the arguments on the right provides the case for the claim that τ_2 is superior to τ_1 (or equivalently $.\tau_1$ is inferior to τ_2). Informally, we want to have each of the possible subsets of the evidence that supports a claim as an argument because we want to consider all possible ways that the evidence could be used as a winning argument. We will explain this in the rest of this section.

Looking at Example 2, we see intuitively that the arguments with differing claims conflict. Obviously it cannot be the case that both of the claims are true. So in this sense these arguments attack, or rebut, each other. We can represent the arguments and the attacks between them by a network (technically, a directed graph): Each node is an argument, and each arc (i.e. arrow) denotes one argument attacking another.

Example 3. Continuing Example 2, we can see that each argument with claim $CP > NT$ attacks each argument with claim $CP < NT$ and vice versa. In other words, each argument with claim $CP > NT$ is a counterargument to each argument with claim $CP < NT$ and vice versa. This is represented by the following directed graph.



4.3 Identification of preferences over structured arguments (Step 3)

Not all structured arguments are of the same weight. They vary in terms of the benefits that they offer, so for instance one argument may have the claim that τ_1 is superior τ_2 because of a substantial improvement in life expectancy, and another argument may have the claim that τ_2 is superior to τ_1 because the former has no side-effects, and the latter has some minor side-effects. To capture this, we use a preference relation over structured arguments that takes into account the nature and magnitude of the outcomes presented in the evidence (as we suggested in the introduction). This allows for a simple and intuitive approach to capturing subjective criteria.

Example 4. Continuing Example 1, given the outcome indicators presented in the evidence table, a clinician or patient may express the following preferences over them as following.

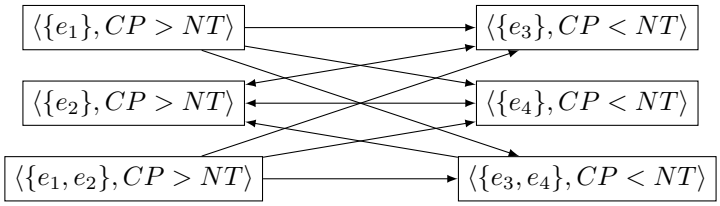
- (Preference 1) Substantial reduction in pregnancy is *more preferred* to modest reduction in risk of either breast cancer or DVT.
- (Preference 2) Modest reduction in risk of ovarian cancer is *equally preferred* to modest reduction in risk of either breast cancer or DVT.
- (Preference 3) Modest reduction in risk of ovarian cancer is *less preferred* to modest reduction inower risk in both DVT and breast cancer.

In our framework, preferences over outcomes are used to refine the symmetrical (bidirectional) attacks between structured arguments. For each pair of structured arguments A and B, if the outcome indicators and their magnitude in the evidence in A are preferred to the the outcome indicators and their magnitude in the evidence in B, then A attacks B and B does not attack A.

Example 5. The preferences in Example 4 can be used to refine the directed graph in Example 3 to give the following directed graph.

- Preference 1 is used to prefer arguments involving evidence e_1 over arguments involving evidence e_3 or e_4, and so the top and bottom arguments on the left attack each of the arguments on the right (but not vice versa).
- Preference 2 is used to identify that an argument involving just evidence e_2 is equally preferred to an argument involving just evidence e_3 and that an argument involving just evidence e_2 is equally preferred to an argument involving just evidence e_4, and so the middle argument on the left attacks the top and middle arguments on the right, and top and middle arguments on the right each attack the middle argument on the left.

– Preference 3 is used to prefer an argument involving both evidence e_3 and e_4 over an argument involving just evidence e_2, and so the bottom argument on the right attacks the middle argument on the left (but not vice versa).



4.4 Generation of meta-arguments (Step 4)

Structured arguments may vary also in terms of the quality of the evidence. For instance, one argument may be based on one small randomized clinical trial, and another may be based on a number of large randomized clinical trials. To address this, we use meta-arguments.

Each meta-argument is a counterargument to an structured argument that is generated because there is a weakness in the evidence of the structured argument. For example, if an structured argument is based entirely on evidence that is not statistically significant, then a meta-argument could be a counterargument to it.

Example 6. Continuing Example 1, we may choose the meta-argument M = "Not statistically significant" to attack each structured argument that has evidence that has a p value above 0.05. So M attacks each of the following arguments.

$$\langle\{e_2\}, CP > NT\rangle$$
$$\langle\{e_1, e_2\}, CP > NT\rangle$$

There is a wide range of possible meta-arguments that can be used, and more than one meta-argument can be used at any one time. Each meta-argument attacks the evidence in a structured argument, and examples include

– The evidence contains flawed RCTs.
– The evidence contains results that are not statistically significant.
– The evidence is from trials that are for a very narrow patient class.
– The evidence has outcomes that are not consistent.

There are various ways we can formalize each of these as criteria as meta-argument (e.g. the meta-argument "Not statistically significant" could be defined as $p < 0.1$, or $p < 0.05$, or $p < 0.01$, or indeed any appropriate value for p).

Furthermore, various refinements of a meta-argument can be considered. For example, we could have a meta-argument "Not statistically significant for the intended outcome". So for instance, this would attack an structured argument that contained evidence that was not statistically significant for the outcome indicator that we want to treat, but it would not attack an structured argument

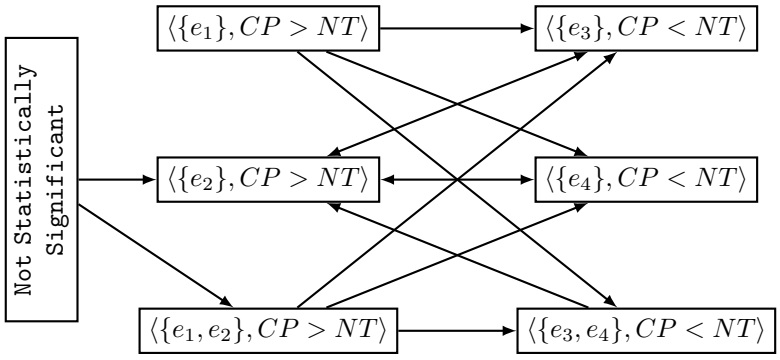
only because it contained evidence that was not statistically significant for a side-effect. The rationale behind such a refinement would be that the majority of trials are set up to determine the efficacy of treatments, rather than for side-effects, and so it is normal for outcomes concerning side-effects to not be statistically significant and yet they are important in aggregating evidence about a treatment.

Obviously, using meta-arguments can have various kinds of ramification in the aggregation process, but the aim is to reflect the choices that clinicians and researchers have for attacking evidence, and moreover make this an explicit and auditable process. So if an aggregation of the evidence involves specific meta-arguments, then these are documented precisely and clearly with the outcome of the aggregation so that we have a reproducible and transparent process.

4.5 Generation of evidential argument graph (Step 5)

An argument graph is a directed graph where each node denotes an argument, and each arc denotes an attack by one argument on another. So when one argument is a counterargument to another argument, this is represented by an arc. For each pair of treatments of interest, we construct an argument graph containing the structured arguments concerning these treatments, together with the meta-arguments that raise concerns with regard to the quality of the evidence in those structured arguments. In other words, this is the graph generated in Step 3 augmented with the meta-arguments generated in Step 4. We call this an evidential argument graph.

Example 7. Continuing Example 1, we have the following evidential argument graph. The structured arguments and the attacks between then come from Example 5, and the meta-argument and the attacks by the meta-argument come from Example 6.



An evidential argument graph provides a clear and useful summary of the evidence in terms of the claims that can be made, the preferences over the outcomes suggested by the evidence, and the weaknesses in the evidence.

4.6 Evaluating the argument graph (Step 6)

We then evaluate the evidential argument graph to determine which arguments are warranted (i.e. which arguments "win" in the argumentation) and which arguments are unwarranted (i.e. which arguments "loose" in the argumentation). Given the graph, any argument (structured or meta) that is unattacked is warranted. For each of the remaining arguments,

- if it is attacked by a warranted argument, then it is unwarranted
- if all the arguments that attack it are unwarranted, then it is warranted
- if it is attacked by an argument that is neither warranted nor unwarranted, then it is undecided

Using this argumentation process, an argument is undecided unless there are assignments to its attacking arguments to make it either warranted or unwarranted.

Example 8. Continuing Example 7, the meta-argument is unattacked, and the structured argument $\langle\{e_1\}, CP > NT\rangle$ is unattacked, and so both are warranted. Each of $\langle\{e_2\}, CP > NT\rangle$ and $\langle\{e_1, e_2\}, CP > NT\rangle$ are attacked by the meta-argument, and so both are unwarranted. Finally, all the arguments on the right are attacked by $\langle\{e_1\}, CP > NT\rangle$, and so they are unwarranted.

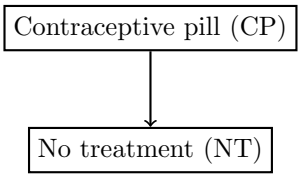
Example 9. Returning to Example 3, suppose we have no preferences over the arguments, and we have no meta-arguments, then the evidential argument graph would be the graph given in Example 3. So every argument is unattacked, and so we cannot identify any warranted arguments or any unwarranted arguments. Therefore, all the arguments are undecided.

Note, our framework is defined so that it is not possible to have an evidential argument graph with a warranted argument with claim $\tau_1 > \tau_2$ and a warranted argument with claim $\tau_1 < \tau_2$. It is a property of our framework that we have warranted arguments with one of the claims, or we have all the structured arguments being either unwarranted or undecided.

4.7 Generation of superiority graph (Step 7)

So far, we have only considered pairs of treatments, and for each pair of treatments τ_1 and τ_2 we have an argument graph. We summarise the result of the argument graph as a superiority graph. If the winning arguments have the claim that that τ_1 is superior to τ_2, then this is represented in the superiority graph by an arc from τ_1 to τ_2. For each arc in the superiority there is an associated argument graph which has been used to determine the direction of the arc. This argument graph is available to the user as an explanation for the direction of the arc.

Example 10. Continuing Example 8, there is an argument with the claim $CP > NT$ that is warranted, and all the arguments with the claim $CP < NT$ are unwarranted. So from the evidence table given in Example 1, we obtain the following superiority graph.



If an evidence table considers more than two treatments, as for example in Table 1, then an evidential argument graph needs to be generated for each pairs of treatments. So for the glaucoma evidence table, six evidential argument graphs were constructed, and the outcome from each of these gives one of the arcs in the superiority graph in Figure 1.

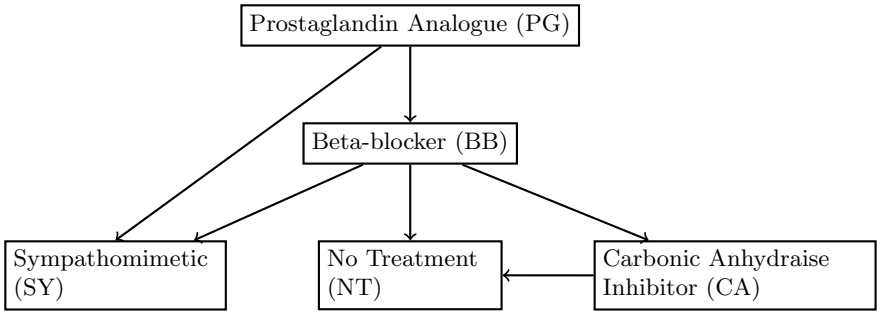


Fig. 1. Example of a superiority graph. This concerns treatments for glaucoma and it has been generated by our approach using the evidence table given in Table 1. There is an arc for each pair of treatments that we compared in one or more trials. If a pair of treatments were not compared in any trial, then there is no arc between them. When there is an arrow from treatment τ_1 to τ_2, then it means that our study found τ_1 to be superior to τ_2.

4.8 Summary of our approach

Our framework allows for the construction of arguments on the basis of evidence as well as their syntheses. The evidence available is then presented and organized according to the agreement and conflict inherent. In addition, users can encode preferences for automatically ruling in favour of the preferred arguments in a conflict.

The **input to our framework** is a table of evidence comparing pairs of treatments. Each row in the table concerns a specific item of evidence such as a randomized clinical trial, and it gives the pair of treatments, the outcome indicator (e.g. disease-free survival, or overall survival), the outcome value, and optionally further details such as the kind of comparison (e.g. randomized clinical trial, meta-analysis, or network analysis), the statistical significance, etc. For

any treatments τ_1 and τ_2 occurring in the evidence table, our framework would attempt to determine whether τ_1 is superior to τ_2, or τ_1 is equivalent to τ_2, or τ_1 is inferior to τ_2. This assessment would be justified by the arguments and counterarguments used to reach this conclusion.

The **output from our framework** is a **superiority graph** which is a directed graph where each node denotes a treatment (appearing in the input evidence table), each unidirectional arc from τ_1 to τ_2 denotes that τ_1 is superior to τ_2, and each bidirectional arc between τ_1 and τ_2 denotes that τ_1 is equivalent to τ_2.

So by determining in general whether one treatment is superior to another based on comparisons involving specific outcome indicators, we are using the items of evidence (concerning comparisons involving specific outcome indicators) as proxies for the general statement that in clinical and statistical terms one treatment is superior (or equivalent) to another. Furthermore, the items of evidence are normally incomplete and also disagree with each other as to which treatment is superior (for instance a treatment τ_1 may be superior to another τ_2 in suppressing the risk of mortality due to a particular disease, but τ_1 may be inferior to τ_2 because τ_1 has a substantial risk of a fatal side-effect and τ_2 has no risk of this side-effect). So to deal with the incomplete and inconsistent nature of the evidence, we have developed an approach that is based on a computational model of argumentation that takes into account the logical structure of individual arguments, and the dialectical structure of sets of arguments. We summarize our approach in Figure 2.

5 Managing subjectivity in aggregation criteria

So far in this paper we have explained how the evidence table is the input to the system, each pair of treatments is evaluated using an argument graph, and then a summary is produced in the form of a superiority graph. For this, we have assumed a single preference relation over the arguments (obtained from the preference relation), and a specific set of meta-arguments.

However, in practice it is normally not obvious that there is a single preference relation or a single set of meta-arguments. This is because, in general, the selection of a preference relation, and the selection of meta-arguments, are subjective criteria. Different clinicians, or their patients, may have different preference relations. This is an intrinsic and unavoidable feature of dealing with preferences over outcome indicators and their magnitude. Specification of the meta-arguments is also subjective because different experts judge evidence differently.

So irrespective of whether our proposal is used, aggregating clinical evidence involves subjective information. But the following are two key advantages of our approach for dealing with this subjective information:

Reproducibility The preference relation and the set of meta-arguments are presented explicitly with the superiority graph. This means that any aggregation of the evidence is reproducible. The evidence, the preference relation,

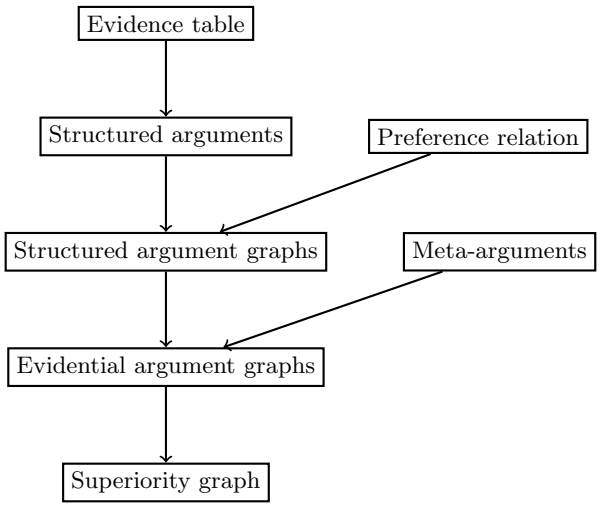


Fig. 2. Summary of our framework for evidence aggregation. The input is the evidence table and the output is the superiority graph. For each pair of treatments in the evidence table where there is a least one item of evidence comparing them, an evidential argument graph is produced. The evidential argument graph contains the structured arguments each of which takes a subset of the evidence to claim that one treatment is better (or equivalent) and meta-arguments that are counterarguments to structured arguments.. One structured argument attacks another if their claims conflict, and the benefits of the first argument are preferred to the second. Each meta-argument attacks an structured argument when there is a weakness in the quality of the evidence used in the structured argument. If "winners" of the evidential argument graph, are all arguments for one treatment being superior to another, then this is reflected in the superiority graph.

and the meta-arguments, can all be made available so that anyone can check exactly how the argument graphs and the superiority graph has been produced. This means the process is transparent and auditable.

Sensitivity analysis Since there is not a preference relation or a set of meta-arguments that is always the right choice, different combinations of preference relation and/or meta-arguments can be used. In this way, a form of sensitivity analysis can be undertaken and so a treatment can be identified as superior for a range of preference relations and/or sets of meta-arguments. Furthermore, if the superiority graph changes little over a wide range of sensible preference relation and meta-arguments, then the superiority graph could be regarded as robust. Such sensitivity analyses may allow researchers and clinicians to categorize their findings according to robustness, and it may allow them to focus their discussions on evidence that is sensitive to the choice of preference relation or meta-arguments.

In general, we believe that a preference relation and a set of meta-arguments should be justifiable in some sense. Therefore there should be some clinical or ethical reason for adopting a particular preference relation, and there should be some methodological or clinical reason for adopting a particular set of meta-arguments.

But it may also be worthwhile to go backwards from a particular superiority graph to identify a preference relation and a set of meta-arguments that would give that superiority graph. For instance, suppose we have some evidence concerning treatments τ_1 and τ_2, and we consider τ_1 superior to τ_2. Suppose we cannot find any combination of preference relation and set of meta-arguments that is justifiable, then we have a stronger case for saying that τ_1 is not superior to τ_2.

In conclusion, using our framework, we can investigate the sensitivity of aggregations of evidence according to different subjective choices concerning the evidence table (i.e. when deciding whether two trials concern the same treatment or the same patient class is a subjective decision), and in the aggregation process (i.e. when deciding which preference relation and which meta-arguments to use). This leads to investigation of the sensitivity of a superiority graph to these subjective choices, and the identification of treatments are superior for a wide range of subjective choices (for the evidence table and the aggregation process).

6 Managing subjectivity in representing evidence

Another kind of subjectivity in the aggregation process, concerns the way in which we group evidence. In many domains, the precise specification of the patient groups and treatments may vary across different trials. However, in order to make sense of the evidence, we accept that some treatments or patients can be grouped. This approach is common in existing systematic reviews, and also applies to our framework.

Patient class When aggregating a set of trial results, we need to assume that the patient group is the same, and that the same treatments are being used. Normally, this is not the case. There may be small differences in the inclusion and exclusion criteria, and therefore the specification of the patient class needs to be relaxed to allow the trials to be regarded as concerning the same patient class. For example, if trial A considers male patients over 21 and trial B considers male patients over 23, then it would be reasonable to relax the patient class to being male adults and so both trials concern the same patient class.

Treatments Similarly, the exact drug, the dosage, and the frequency of treatment might be slightly different, but for aggregation, they can be regarded as the same (e.g. for a particular drug 10% and 15% concentration may be regarded as the same treatment). Again this involves relaxation. As another example, many drugs for cancer are given in a cocktail (i.e. a mixture of therapies), and it is often difficult to find exactly the same cocktail used in more than a small number of trials. So again, the specification of the cocktail needs to be relaxed in order to aggregate the results.

Grouping of patients and treatments (relaxation) offers a valuable tool for analyzing clinical evidence in order to make more insightful and robust recommendations. To address this, we can couple the construction of arguments with an computer-readable model of the world, which contains accepted groupings of patients and treatments (an ontology), in order to automate the grouping of evidence according to patient class and/or treatment. By using the ontology to determine that two or more trials concern the same patient class and treatment, means that we have more evidence to consider for our arguments to any particular argument graph. We illustrate this idea in the next example.

Example 11. Suppose we have the following evidence table that is the same as the evidence table given in Example 1 except we have specific brands CP1 or CP2 instead of CP, where CP1 and CP2 are similar second generation low dose contraceptive pills.

ID	Left	Right	Outcome indicator	Outcome value	Net outcome	p
e1	CP1	NT	pregnancy	0.05	superior	0.01
e2	CP2	NT	ovarian cancer	0.99	superior	0.07
e3	CP1	NT	breast cancer	1.04	inferior	0.01
e4	CP2	NT	DVT	1.02	inferior	0.05

By using the ontological knowledge that CP1 and CP2 are similar, the above evidence table can be relaxed to the evidence table given in Example 1. In other words, by using this ontological knowledge, we can automatically replace CP1 and CP2 by CP in each entry in the Left column.

We have undertaken a theoretical analysis of how this may be done [4], and we can harness this for developing our sensitivity analysis of superiority graphs (whether by hand or by automated computer-readable ontologies).

7 Relationship of our approach with GRADE

One of the key questions when aggregating evidence is to what extent we can trust the evidence we have. There have been several approaches to considering the quality of evidence, including SIGN [5], and MERGE [6]. See [7] for a discussion. However, more recent work has aimed to achieve consensus via the GRADE guidelines [8].

We see our approach as being consistent with the GRADE approach. GRADE is a paper-based approach for making clinical recommendations based on evidence. It is an important tool for guideline development organizations such as NICE. In the approach, assignment of strength is made to each recommendation. Strong recommendations are made when the desirable effects of an intervention outweigh the undesirable effects, and weak recommendations are made when the trade-offs are less certain. Outcomes are graded according to their importance using a scale from 1 to 9. For instance, in considering phosphate lowering drugs in patients with renal failure, flatulence has grade 2, pain due to soft tissue calcification has grade 6, fractures has grade 7, myocardial infarction has grade 8, and mortality has grade 9 [9]. Allowing desirable and undesirable outcomes to be weighed. Furthermore, recommendations can be downgraded when the evidence is not of a sufficiently high quality. Items of evidence that are based on randomized clinical trials are *a priori* regarded as high quality evidence. But this assignment may be decreased for various reasons such as study limitations, inconsistency of results, indirectness of evidence, imprecisions, reporting bias, etc.

We can capture the GRADE approach in our framework using the preference relations, and the meta-arguments, in the argumentation. This means GRADE can benefit from a number of substantial advantages that come with our approach:

1. The way that the evidence is being aggregated is made explicit, with the preference relation and meta-arguments being made explicit, meaning that it is easier for third parties to inspect how the aggregation has been derived;
2. The same criteria (i.e. the same preference relations and meta-arguments) can be used systematically with new evidence tables, and so the aggregation process is consistent;
3. Different criteria (i.e. different combination of preference relation and meta-arguments) can be used in order to determine the sensitivity of ranking of treatments in a superiority graph;
4. Different strength of recommendation can be made by different choices of preference relation and meta-argument;
5. The process of generating superiority graphs can be automated.

Whilst, we have not considered diagnostic tests and strategies in our framework yet, we believe we can also capture the GRADE approach for diagnostic tests and strategies in our approach [10].

8 Discussion

For evidence-based decision making in healthcare, there is a need to abstract away from the details of individual items of evidence, and to aggregate the evidence in a way that reduces the volume, complexity, inconsistency and incompleteness of the information. Moreover, it would be helpful to have a method for automatically analyzing and presenting the clinical trial results and the possible ways to aggregate them in an intuitive form, highlighting agreement and conflict present within the literature.

We believe that our framework for aggregation of clinical evidence using argumentation addresses these needs. The output from our framework is a superiority graph. This is a useful summary of the aggregation of evidence for researchers and clinicians who need to aggregate evidence. Each arc connecting a pair of treatments in the graph is generated by an argumentation process that involves constructing an argument graph using the evidence concerning those two treatments, and this argument graph is available to the users of the superiority graph. They can look at the argument graph to inspect what arguments were considered and what preference criteria and meta-arguments were used. This means that it is explicit how the superiority graph was obtained, and thereby provides an audit trail of the aggregation process. Furthermore, different combinations of preference criteria and meta-arguments can be used to investigate the robustness of any superiority graphs produced.

We have already shown how clinicians use preferences in evaluating evidence [11], and it is straightforward to use our framework to represent these preferences. The advantage of allowing the user to define their own preference relations and their own meta-arguments is that they can systematically use the evidence in the context of their working environment.

We have evaluated our framework with three case studies involving 56 items of evidence, and 16 treatment options. The items of evidence come from three NICE Guidelines, and we have compared the results of our aggregation process with the recommendations made by NICE. In Table 1, we give one of the evidence tables used and in Figure 1, we give the resulting superiority graph. The results using our framework are consistent with the NICE recommendations, though in some cases, it is apparent that they bring extra knowledge (beyond the evidence) into the process such as health economics modelling, or experiential knowledge, and so in some cases their recommendations are more refined than ours. We made simple choices for the preference relations over sets of benefits, and we believe that they are robust in the sense that they could be changed quite considerably and still we would get the same results from our aggregation process. For more details on this evaluation of our approach, please see [1].

In another case study, on lung cancer chemo-radiotherapy, we have investigated a number of different benefits preference relation and kinds of meta-argument. For this, we constructed an evidence table with 283 items of evidence (where each item of evidence concerns a pairwise comparison according to a single outcome indicator). The primary evidence on which the evidence table was based was a superset for that used in a Cochrane Review on this topic [12]. For

the systematic review that has resulted from our case study, the different ways of aggregating the evidence gave various insights into the evidence, such as the identification of weaknesses in the evidence base, and suggestions being made for future clinical trials to better determine which of the available treatments is superior. By exploring various relaxations of the evidence, we were able to make more refined recommendations than obtained with the original Cochrane review.

As we explained in Section 7, our approach is consistent with GRADE, and the GRADE approach for interventions can be formalized and automated in our approach giving a number of benefits. By using GRADE in our approach, any assumptions are made explicit, and the aggregation process is reproducible.

Our approach is also consistent with standard techniques such as meta-analyses. If there are multiple trials with the same outcome indicator, then standard techniques such as taking the weighted average offer substantial advantages. However, standard meta-analysis techniques do not handle multiple outcome indicators [13, 3]. So if there are multiple trials with the same outcome indicator, then standard techniques can be applied, and the result of the standard techniques used as the input to our approach. In other words, for the evidence table, a row can be based on a meta-analysis. So our approach can harness the output of standard meta-analysis techniques, but our approach can address problems that cannot be addressed by standard meta-analysis techniques

Network analysis is an increasingly popular method for systematic reviews with over 30 published in 2011, and an estimate of over 50 in 2012 [14]. In network analysis, the pairwise superiority of interventions is considered transitively. For example, if τ_1 is superior to τ_2 and τ_2 is superior to τ_3, then by transitivity τ_1 is superior to τ_3. In general, such an inference can be error-prone (for a discussion of this, see [15]). But with further information about the trials (such as details about the populations, results, etc), then there are network analysis techniques that can qualify the transitive inference [16]. Also, see [17] for a discussion of network analysis. However, as with meta-analysis techniques, network analysis techniques assume a common outcome indicator. So again, we believe that our approach is consistent with network analysis techniques. Our approach can harness the output of network analysis techniques, but our approach can address problems that cannot be addressed by network analysis techniques.

Acknowledgements

The authors would like to thank Jiri Chard and Cristina Visintin for valuable feedback on this tutorial.

References

1. Hunter, A., Williams, M.: Aggregating evidence about the positive and negative effects of treatments. Artificial Intelligence in Medicine **56** (2012) 173–190
2. Besnard, Ph.., Hunter, A.: Elements of Argumentation. MIT Press, Cambridge, MA, USA (2008)

3. Hackshaw, A.: A Concise Guide to Clinical Trials. WileyBlackwell, London, UK (2009)
4. Gorogiannis, N., Hunter, A., Williams, M.: An argument-based approach to reasoning with clinical knowledge. International Journal of Approximate Reasoning **51**(1) (2009) 1 – 22
5. SIGN: SIGN 50: A Guideline Developers Handbook. Scottish Intercollegiate Guidelines Network (2011)
6. Liddle, J., Williamson, M., Irwig, L.: Method for Evaluating Research and Guideline Evidence. New South Wales Health Department (1996)
7. NICE: The Guidelines Manual. National Institute for Health and Clinical Excellence (2009)
8. Guyatt, G., Oxman, A., Vist, G., Kunz, R., Falck-Ytter, Y., Alonso-Coelle, P., Schnemann, H.: GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. British Medical Journal **336** (2008) 924–926
9. Guyatt, G., Oxman, A., Vist, G., Kunz, R., Falck-Ytter, Y., Schnemann, H.: GRADE: what is quality of evidence and why is it important to clinicians. British Medical Journal **336** (2008) 995–998
10. Schnemann, H., Oxman, A., Brozek, J., Glasziou, P., Jaeschke, R., Vist, G., Williams, J., Kunz, R., Craig, J.: GRADE: grading quality of evidence and strength of recommendations for diagnostic tests and strategies. British Medical Journal **336** (2008) 1106–1110
11. Hunter, A., Williams, M.: Using clinical preferences in argumentation about evidence from clinical trials. In Veinot, T., Çatalyürek, Ü., Luo, G., Andrade, H., Smalheiser, N., eds.: Proceedings of the First ACM International Health Informatics Symposium, ACM Press (2010) 118–129
12. O'Rourke, N., Roqué Figuls, I., Farré Bernadó, N., Macbeth, F.: Concurrent chemoradiotherapy in non-small cell lung cancer. Cochrane Database of Systematic Reviews **6** (2010) DOI: 10.1002/14651858.
13. Kirkwood, B., Sterne, J.: Essential Medical Statistics. Blackwell (2003)
14. Bafeta, A., Trinquart, L., Seror, R., Ravaud, P.: Analysis of the systematic reviews process in reports of network meta-analyses: methodological systematic review. British Medical Journal **347** (2013) f3675
15. Baker, S., Kramer, B.: The transitive fallacy for randomized trials: If a beats b and b beats c in separate trials, is a better than c? BMC Medical Research Methodology **2**(13) (2002)
16. Lumley, T.: Network meta-analysis for indirect treatment comparison. Statistics in Medicine **21** (2002) 2313–2324
17. Li, T., Puhan, M., Vedula, S., Singh, S., Dickersin, K.: Network meta-analysis-highly attractive but more methodological research is needed. BMC Medicine **(** (2011) 79
18. NICE: Glaucoma: Clinical Guidelines CG85. National Institute for Health and Clinical Excellence (www.nice.org.uk), London, UK (2009) (accessed 1 April 2012).

ID	Left	Right	Outcome indicator	Outcome value	Net outcome	Sig	Type
e_{01}	BB	NT	visual field prog	0.77	superior	no	MA
e_{02}	BB	NT	change in IOP	-2.88	superior	yes	MA
e_{03}	BB	NT	respiratory prob	3.06	inferior	no	MA
e_{04}	BB	NT	cardio prob	9.17	inferior	no	MA
e_{05}	PG	BB	change in IOP	-1.32	superior	yes	MA
e_{06}	PG	BB	acceptable IOP	1.54	superior	yes	MA
e_{07}	PG	BB	respiratory prob	0.59	superior	yes	MA
e_{08}	PG	BB	cardio prob	0.87	superior	no	MA
e_{09}	PG	BB	allergy prob	1.25	inferior	no	MA
e_{10}	PG	BB	hyperaemia	3.59	inferior	yes	MA
e_{11}	PG	SY	change in IOP	-2.21	superior	yes	MA
e_{12}	PG	SY	allergic prob	0.03	superior	yes	MA
e_{13}	PG	SY	hyperaemia	1.01	inferior	no	MA
e_{14}	CA	NT	convert to COAG	0.77	superior	no	MA
e_{15}	CA	NT	visual field prog	0.69	superior	no	MA
e_{16}	CA	NT	IOP > 35mmHg	0.08	superior	yes	MA
e_{17}	CA	BB	hyperaemia	6.42	inferior	no	MA
e_{18}	SY	BB	visual field prog	0.92	superior	no	MA
e_{19}	SY	BB	change in IOP	-0.25	superior	no	MA
e_{20}	SY	BB	allergic prob	41.00	inferior	yes	MA
e_{21}	SY	BB	drowsiness	1.21	inferior	no	MA

Table 1. An evidence table concerning treatments for glaucoma. Each row is a meta-analysis from the NICE Glaucoma Guideline [18] (Appendix pages 213-223) for the class of patients who have raised intraocular pressure (i.e. raised pressure in the eye) and are therefore at risk of glaucoma with resulting irreversible damage to the optic nerve and retina. Each item is a meta-analysis (MA) generated by the guideline authors as presented in the appendix of the guideline. The medications considered are no treatment (NT), beta-blocker (BB), prostaglandin analogue (PG), sympathomimetic (SY), and carbonic anhydrase inhibitor (CA). The Net outcome column gives an interpretation of the value with respect to the type of outcome indicator: For the outcome indicator "change in IOP", if the value is negative, the left arm is superior, otherwise it is inferior. For the outcome indicator "acceptable IOP", which is a desirable outcome for the patient, if the value is greater than 1, the left arm is superior, otherwise it is inferior. For each of the remaining outcome indicators (i.e. for "respiratory problems", "cardiovascular problems", "allergy problems", "hyperaemia", "convert to COAG", "visual field progression", "IOP > 35mmHg", and "drowsiness"), which are undesirable for the patient, if the value is less than 1, then the left arm is superior, otherwise it is inferior. Note, "hyperaemia" means redness of eyes, "convert to COAG" means the patient develops chronic open angle glaucoma, "visual field progression" means that there is damage to the retina and/or optic nerve resulting in loss of the visual field and "IOP > 35mmHg" means that the intraocular pressure is above 35mmHg (which is very high).