Aggregating evidence about the positive and negative effects of treatments

Anthony Hunter[a,*] and Matthew Williams[b]

(a) Department of Computer Science, University College London, London, WC1E 6BT, UK

(b) Department of Clinical Oncology, University College Hospital, London, NW1 2PG, UK

September 20, 2012

Abstract

Objectives

Evidence-based decision making is becoming increasingly important in healthcare. Much valuable evidence is in the form of the results from clinical trials that compare the relative merits of treatments. In this paper, we present a new framework for representing and synthesizing knowledge from clinical trials involving multiple outcome indicators.

Method

The framework generates and evaluates arguments for claiming that one treatment is superior, or equivalent, to another based on the available evidence. 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 attacks arguments that are based on weaker evidence.

Results

We evaluated the framework with respect to the aggregation of evidence undertaken in three published clinical guidelines that involve 56 items of evidence and 16 treatments. For each of the three guidelines, the treatment we identified as being superior using our method is a recommended treatment in the corresponding guideline.

Conclusions

The framework offers a formal approach to aggregating clinical evidence, taking into account subjective criteria such as preferences over outcome indicators. In the evaluation, the aggregations obtained showed a good correspondence with published clinical guidelines. Furthermore, preliminary computational studies indicate that the approach is viable for the size of evidence tables normally encountered in practice.

Keywords

Computational models of argument; Argument systems; Knowledge aggregation; Evidence aggregation; Evidence-based medicine; Clinical recommendations

[*] Corresponding author (a.hunter@cs.ucl.ac.uk)

1 Introduction

The systematic use of evidence is already established in healthcare in the form of evidence-based decision making. However, the rapidly increasing amount of evidential knowledge on a subject means that it is difficult for a clinician or biomedical researcher to effectively and efficiently acquire and assimilate that evidence. Therefore, getting a quick, up-to-date review of the state of the art on treatment efficacy for a particular condition is not always feasible. This problem is exacerbated by the fact that the evidence is often conceptually complex, heterogeneous, incomplete and inconsistent.

To cope with these problems (of volume, complexity, inconsistency and incompleteness of evidence), the 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.

- 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.

- 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.

- 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.

- 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.

- 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.

- 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.

In conclusion, 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.

2 Our proposal

As a first step to addressing the needs raised in the previous section, we have presented a general framework for representing and synthesizing knowledge from clinical trials involving the *same* outcome indicator (e.g. overall survival, or disease-free survival) [1]. The framework allows for arguments and counterarguments to be constructed and compared, and it allows for diverse criteria concerning the quality of the evidence to be taken into account when considering which arguments prevail [2, 3]. The framework also allows for reasoning with the evidence according to the patient class to which it applies. For instance, when comparing a pair of treatments, some trials may have involved a broad class of patients, others may have involved quite restricted subclasses, and so, it can be useful to identify the evidence that pertains to a specific patient class of interest. To address this need, we have shown how ontological reasoning can be harnessed in the argumentation process [4].

In this paper, to further address these needs of aggregating evidence, we present a framework for representing and synthesizing knowledge from clinical trials involving *multiple* outcome indicators. Our framework allows 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 a system based on our framework** is a table of evidence comparing pairs of treatments. Each row in the table gives the pair of treatments, the kind of comparison (e.g. randomized clinical trial, meta-analysis, or network analysis), the outcome indicator (e.g. disease-free survival, or overall survival), the outcome, the statistical significance, etc. For any treatments τ_1 and τ_2 occurring in the evidence table, a system based on 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 a system based on 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. We illustrate three examples of superiority graphs in Figure 1.

In order to use the evidence given as input to generate the superiority graph as output, our system uses an argumentation process. We summarize below (and then explain fully in the paper) the features of this process in terms of Steps 1 to 5 that takes a set of evidence as input (at Step 1) and produces a superiority graph as output at (Step 5) as follows.

1. **Generation of inductive arguments** From the input evidence, the inductive arguments are generated. Each inductive 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, then the claim ϵ is that τ_1 is superior to τ_2. Whereas if all the

evidence in X indicates that τ_1 is on balance neither better not worse than τ_2, then the claim ϵ is that τ_1 is equivalent to τ_2.

2. **Identification of preferences over inductive arguments** Not all inductive 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 inductive arguments that takes into account the nature and magnitude of the outcomes presented in the evidence.

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

4. **Generation of evidential argument graph** 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 inductive arguments concerning these treatments, together with the meta-arguments that raise concerns with regard to the quality of the evidence in those inductive arguments. We then evaluate the graph to determine which arguments are acceptable (i.e. which arguments "win" in the argumentation). These criteria, which we will explain, are based on dialectical criteria developed in the field of computational models of argument.

5. **Generation of superiority graph** For each pair of treatments τ_1 and τ_2 we have an argument graph. If the winning arguments have the claim that that τ_1 is superior to τ_2, then this is reflected in the superiority graph by an arc from τ_1 to τ_2. The superiority graph is a summary of the argumentation. 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.

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.

In an earlier attempt at using an argument-based approach to aggregating evidence involving multiple outcome indicators [5], we proposed the use of a two-dimensional preference ordering over arguments. This involved a numerical assignment based on an evaluation of the benefits offered by the treatments used, and a numerical assignment based on an evaluation of the quality of the evidence. The proposal was a form of utility-theoretic framework. Following discussions with clinicians, we decided that a simpler framework was required to consider the benefits of the treatments and the quality of the evidence. This led us to the proposal in this paper that has the following features: (1) A simplified and more intuitive definition for a preference relation over

	Left	Right	Outcome indicator	Value	Net	Sig	Type
e_1	ACE	CCB	mortality	1.04	<	no	MA
e_2	ACE	CCB	stroke	1.15	<	yes	MA
e_3	ACE	CCB	heart failure	0.84	>	yes	MA
e_4	ACE	CCB	diabetes	0.85	>	yes	MA

Table 1: Four results obtained from the NICE Hypertension Guideline (GC34, pages 36-37) concerning angiotensin-converting inhibitors (ACE) and calcium channel blockers (CCB)

inductive arguments; (2) meta-arguments (which are important to identify important weaknesses in the evidence used in inductive arguments); (3) More substantial case studies for evaluating the proposal; and (4) A methodology for using the framework that takes into account the subjective criteria required for aggregating evidence.

We proceed as follows: (Section 3) We discuss how we can represent evidence in a tabular format; (Section 4) We discuss how clinical evidence is currently aggregated by healthcare professionals; (Section 5) We review an abstract model of argumentation that we will incorporate in our general framework; (Section 6) We show how we can generate inductive arguments based on the available evidence; (Section 7) We show how we can identify a preference relation over arguments based on the relative benefits for the treatments being considered; (Section 8) We introduce the notion of meta-arguments that allows for a refinement of the argumentation process; (Section 9) We show how we can use the framework to aggregate evidence to generate a superiority graph; (Section 10) We evaluate our proposal with three case studies (Glaucoma, Hypertension, and Preeclampsia) and we compare the results with NICE Guidelines; (Section 11) We discuss how we can deal with subjective criteria in our framework; (Section 12) We discuss implementation issues; (Section 13) We conclude with a discussion of the proposal in this paper, and how it relates to the literature.

3 Representing evidence

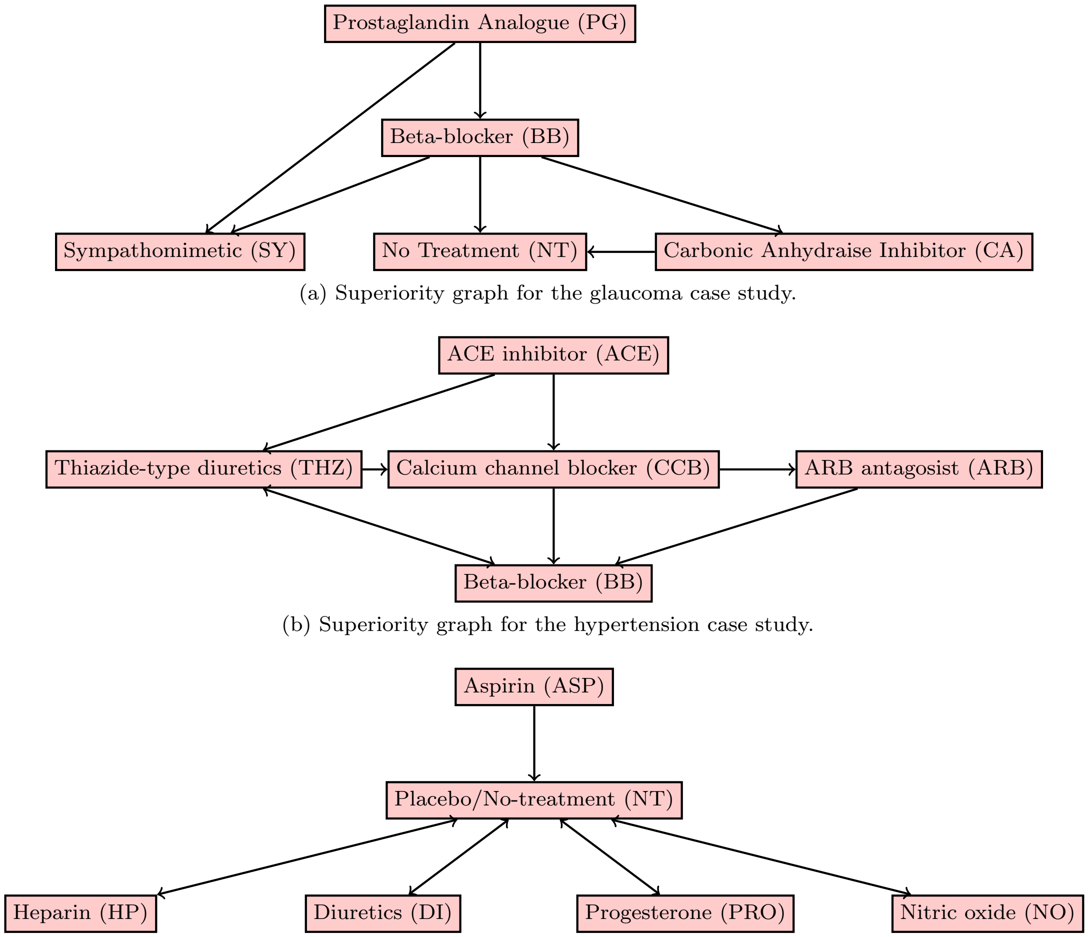
The types of evidence we consider in this paper include randomized clinical trial (RCT), meta-analyses (MA), network analyses (NA), and cohort study (CS). Our focus will be on 2-arm superiority trials, i.e., clinical trials whose purpose is to determine whether, given two treatments, one is superior to the other (strictly speaking, such a trial tries to disprove the hypothesis that the two treatments are identical). This is an extremely common trial design.

We represent evidence in a table. Each row is an item of evidence taken from an RCT, a CS, an MA or an NA. The choice of columns depends on the available information and the criteria that will be used for aggregating the evidence. We give an example in Table 1 concerning patients who require a treatment for hypertension (data from www.nice.org.uk). The table incorporates the columns normally required for our framework, and we explain them as follows.

- The **left** and **right** attributes signify the treatments compared in each item of evidence. In the table, these are angiotensin-converting inhibitors (ACE) for the left arm and calcium channel blockers (CCB) for the right arm.

- The **outcome indicator** is the specification of the particular outcome that is being considered when comparing the two treatments. In the table, in each row, it is relative risk (i.e. it is the proportion of patients who have the event or condition, i.e. “mortality”, “stroke”, “heart failure” or “diabetes”, within the period of the trial).

- The **value** of the outcome is the value obtained by the method applied to the outcome indicator. So for the first row, it is the proportion of patients who died during the trial taking ACE divided by the proportion of patients who died during the trial taking CCB.



(a) Superiority graph for the glaucoma case study.

(b) Superiority graph for the hypertension case study.

(c) Superiority graph for the pre-eclampsia case study.

Figure 1: Superiority graphs for the case studies. Each graph concerns a set of treatments considered for a particular class of patients. Each node is a treatment, and each edge denotes that there is evidence to suggest either the one treatment, τ_1, is superior to another τ_2, in which case there is a unidirectional edge from τ_1 to τ_2, or that τ_1 is equivalent to τ_2, in which case there is a bidirectional edge between τ_1 and τ_2. If there is no edge between a pair of treatments, then there is a lack of evidence to compare them.

- The **net outcome**, abbreviated by the column name Net, is a binary relation, denoted $>$ (superior), $\sim$ (equal), and $<$ (inferior), 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. For the first row, mortality is undesirable, and so a risk ratio value less than 1 means that the left arm is superior to the right arm, a risk ratio value equal to 1 means that the left arm is equal to the right arm, and risk ratio value greater than 1 means that the left arm is inferior to the right arm.
- The **statistically significant** attribute, abbreviated by the column name Sig, indicates whether the value is statistically significant. In this example, we just give a yes/no entry. In other examples, we given more detailed information such as the p value.
- The **evidence type**, abbreviated by the column name Type, specifies the type of study undertaken, e.g. randomized clinical trial (RCT), cohort study (CS), meta-analysis (MA), network analysis (NA), etc. It is an indicator of the quality of the evidence.

The set of attributes we have discussed here is only indicative. Often other attributes are useful for assessing and aggregating evidence (e.g. 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 for ensuring patients and clinician do not know which arm a patient is in, etc). For a general introduction to the nature of clinical trials, and a discussion of a wider range of attributes, see [6].

The patient class is an important attribute that can be captured about an item of evidence. For instance, in Table 1, the patient class is people with "persistent raised blood pressure of 160/100 mmHg or more". In our previous work, we showed how the patient class may involve a conjunction and/or disjunction of terms from a medical ontology and description logics can be used to provide inferencing (see [7]). Similarly, treatments presented in the left arm and right arm can be composed of a conjunction and/or disjunction of terms from an ontology. Again, medical ontologies cater for this by providing categories and relationships on treatments, substances used, and other characteristics. See [8, 4] for proposals for using a medical ontology in argumentation about clinical trials.

For simplicity, in the rest of this paper, we assume that the evidence concerns a particular, sensible patient class, and that each treatment in the left arm and right arm is atomic, and so we do not consider the ontological aspects of patient class or treatment further in the rest of this paper.

Later in the paper, we will evaluate our argument-based approach in three case studies: (Glaucoma case study) The evidence, which is presented in Table 3 (in the appendix), concerns treatments for raised intraocular pressure (raised IOP), which is raised pressure in the eye, where the evidence was obtained from the NICE Glaucoma Guideline [9]; (Hypertension case study) The evidence, which is presented in Table 4 (in the appendix), concerns treatments for hypertension, which is raised blood pressure, where the evidence was obtained from the NICE Hypertension Guideline [10]; and (Pre-eclampsia case study) The evidence, which is presented in Table 5 (in the appendix), concerns treatments for pre-eclampsia, which is hypertension arising during pregnancy, where the evidence was obtained from the NICE Hypertension in Pregnancy Guideline [11].

4 Background to aggregating evidence

In the previous section, we describe the kind of input that we use for our framework. This is also the kind of input that is currently used for aggregating evidence in guideline development (such as undertaken by NICE) and systematic reviewing (such as undertaken by the Cochrane Collaboration). We have broken the evidence down to relational data (each entry is a term from an ontology or a value), whereas in say guideline development, there may be some free text in the evidence tables (as seen in the appendices of guidelines produced by organizations such as NICE). Nonetheless, there is a good correspondence between our input and what is currently used.

When considering evidence, particularly in the context of studies such as randomized clinical trials, there is an emphasis on pairwise comparisons. Whilst trials may consider more than two treatments, these are normally built on pairwise comparisons. For example, if a trial considers three treatments τ_1, τ_2, and τ_3, then the statistical information is available to derive three pairwise comparisons, viz. τ_1 compared with τ_2, τ_2 compared with τ_3, and τ_3 compared with τ_1. This means that it is valid to decompose any such trial into pairwise comparisons, and importantly, it means that different trials can be compared by considering the results in terms of pairwise comparisons. In other words, pairwise comparisons provide a common format to compare trials. Furthermore, the utility of pairwise comparisons is seen in the meta-analysis techniques for aggregating evidence for a single outcome indicator (see for example [6]).

When evidence is aggregated according to multiple outcome indicators, in guideline development and systematic reviews, the aim is to determine whether one treatment is better than another. For a pair of treatments, there are two dimensions to this. The first dimension concerns the outcomes (good or bad) being considered. For instance, is one treatment more efficacious than another, or does one treatment have more side-effects than the other? The second dimension concerns the quality of the evidence. For example, if the evidence for claiming that the first treatment is better than the second is based on small non-statistically significant studies, then this might be a reason to not accept any argument claiming that the first treatment is better than the second.

We will investigate these dimensions in more detail in the subsequent sections. However, it is useful to point out here that these dimensions involve different kinds of uncertainty. The first dimension involves uncertainty about whether or not a particular outcome will be obtained. For instance, when comparing two treatments for a particular cancer. The proportion of patients who live for at least 5 years after treatment may be 80% with the first treatment and 70% with the second treatment. So any given patient who has the first treatment would appear to have a probability of 0.8 of surviving for at least 5 years. The second dimension involves uncertainty about whether or not the information about the first dimension is actually true. For instance, if the study concerning the proportion of patients who live for at least 5 years involves 10 patients, then the study would appear to be too small, and there would be low confidence that if the study were repeated, that the same results would be obtained. So the first dimension could be described to be a form "object-level" uncertainty, and the second dimension could be described to be a form of "meta-level" uncertainty.

In our framework, we consider these two dimensions by using preferences over inductive arguments to deal with the first dimension, and using meta-arguments to deal with the second dimension. As we are aiming for a framework that is usable by those involved in guideline development and systematic reviewing, as well as clinicians, we are keen to have a simple breakdown of the kinds of information, in a way that is consistent with how they currently assess the available evidence, and a process that they can use without undue complexity.

As we said in Section 2, the output from our framework is a superiority graph. This appears to be 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.

5 Abstract argumentation

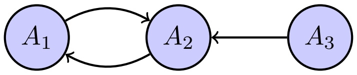
Our framework builds on more general developments in the area of computational models of argument. These models aim to reflect how human argumentation uses conflicting information to construct and analyze arguments. There is a number of frameworks for computational models of argumentation. They incorporate a formal representation of individual arguments and techniques for comparing conflicting arguments (for reviews see [12, 13, 14]). By basing our framework on these general models, we can harness theory and adapt implemented argumentation software as the basis of our solution.

In this section, we review the proposal for abstract argumentation by Dung [15]. The simplest way to formalize a collection of arguments consists of just naming arguments (so, in a sense, treating them as atomic) and merely representing the fact that an argument is challenged by another (and so not indicating what the nature of the challenge is). In other words, a collection of arguments can be formalized as a directed binary graph.

Definition 1. *An* **abstract argument graph** *is a pair* $(\mathcal{A}, \mathcal{R})$ *where* $\mathcal{A}$ *is a set and* $\mathcal{R}$ *is a binary relation over* $\mathcal{A}$ *(in symbols,* $\mathcal{R} \subseteq \mathcal{A} \times \mathcal{A}$*).*

Each element $A \in \mathcal{A}$ is called an **argument** and $(A_i, A_j) \in \mathcal{R}$ means that A_i **attacks** A_j (accordingly, A_i is said to be an **attacker** of A_j). So A_i is a **counterargument** for A_j when $(A_i, A_j) \in \mathcal{R}$ holds.

Example 1. *Consider arguments* A_1 = *"Patient has hypertension so prescribe diuretics",* A_2 = *"Patient has hypertension so prescribe beta-blockers", and* A_3 = *"Patient has emphysema which is a contraindication for beta-blockers". Here, we assume that* A_1 *and* A_2 *attack each other because we should only give one treatment and so giving one precludes the other, and we assume that* A_3 *attacks* A_2 *because it provides a counterargument to* A_2*. Hence, we get the following abstract argument graph.*



Arguments can work together as a coalition by attacking other arguments and by defending their members from attack as follows.

Definition 2. *Let* $S \subseteq \mathcal{A}$ *be a set of arguments.*

- S **attacks** $A_j \in \mathcal{A}$ *iff there is an argument* $A_i \in S$ *such that* A_i *attacks* A_j*.*
- S **defends** $A_i \in S$ *iff for each argument* $A_j \in \mathcal{A}$*, if* A_j *attacks* A_i *then* S *attacks* A_j*.*

The following gives a requirement that should hold for a coalition of arguments to make sense. If it holds, it means that the arguments in the set offer a consistent view on the topic of the argument graph.

Definition 3. *A set* $S \subseteq \mathcal{A}$ *of arguments is* **conflict-free** *iff there are no arguments* A_i *and* A_j *in* S *such that* A_i *attacks* A_j*.*

Now, we consider how we can find an acceptable set of arguments from an abstract argument graph. The simplest case of arguments that can be accepted is as follows.

Definition 4. *A set* $S \subseteq \mathcal{A}$ *of arguments is* **admissible** *iff* S *is conflict-free and defends all its arguments.*

The intuition here is that for a set of arguments to be accepted, we require that, if any one of them is challenged by a counterargument, then they offer grounds to challenge, in turn, the counterargument. There always exists at least one admissible set: The empty set is always admissible.

Clearly, the notion of admissible sets of arguments is the minimum requirement for a set of arguments to be accepted. We will focus on the following classes of acceptable arguments.

Definition 5. *Let* Γ *be a conflict-free set of arguments, and let* $\mathsf{Defended} : \wp(\mathcal{A}) \mapsto \wp(\mathcal{A})$ *be a function such that* $\mathsf{Defended}(\Gamma) = \{A \mid \Gamma \text{ defends } A\}$*.*

1. Γ *is a* **complete extension** *iff* $\Gamma = \mathsf{Defended}(\Gamma)$

2. Γ *is a* **grounded extension** *iff it is the minimal (w.r.t. set inclusion) complete extension.*

3. Γ *is a* **preferred extension** *iff it is a maximal (w.r.t. set inclusion) complete extension.*

The grounded extension is always unique, whereas there may be multiple preferred extensions. We illustrate these definitions with the following examples. As can be seen from the examples, the grounded extension provides a skeptical view on which arguments can be accepted, whereas each preferred extension take a credulous view on which arguments can be accepted.

Example 2. *Continuing Example 1, there is only one complete set, and so this is both grounded and preferred.*

	Conflict free	Admissible	Complete	Grounded	Preferred
$\{\}$	×	×			
$\{A_1\}$	×	×			
$\{A_2\}$	×				
$\{A_3\}$	×	×			
$\{A_1, A_2\}$					
$\{A_1, A_3\}$	×	×	×	×	×
$\{A_2, A_3\}$					
$\{A_1, A_2, A_3\}$					

Example 3. *Consider the situation where we have just two arguments A_4 and A_5 that attack each other. There are two preferred sets, neither of which is grounded.*

	Conflict free	Admissible	Complete	Grounded	Preferred
$\{\}$	×	×	×	×	
$\{A_4\}$	×	×	×		×
$\{A_5\}$	×	×	×		×
$\{A_4, A_5\}$					

The formalization we have reviewed in this section is abstract because both the nature of the arguments and the nature of the attack relation are ignored. In particular, the internal (logical) structure of each of the arguments is not made explicit. Nevertheless, Dung's proposal for abstract argumentation is ideal for clearly representing arguments and counterarguments, and for intuitively determining which arguments should be accepted (depending on whether we want to take a credulous or skeptical perspective).

We harness abstract argumentation in our general framework for aggregating evidence. We will introduce mechanisms for generating arguments from the evidence, and for generating the attacks relation based on the preferences over the arguments. In this way, we will instantiate abstract argumentation with logical arguments. This means that we can use Dung's definitions for determining which sets of arguments are acceptable, and thereby determine which aggregations of the evidence are acceptable.

6 Representing inductive arguments

We start with a set of evidence EVIDENCE $= \{e_1, .., e_n\}$ concerning a pair of treatments $\{\tau_1, \tau_2\}$. So each item has one of these treatments as the left arm and the other as the right arm. Each item in EVIDENCE is a result from an RCT, an MA, a CS, or an NA, represented as a row in a table of evidence (as described in the previous section).

We partition EVIDENCE into three sets SUPERIOR, EQUITABLE, and INFERIOR. Those in SUPERIOR are the trials for which τ_1 was shown to be superior to τ_2 with respect to some outcome indicator. By superior, we mean that if the outcome is desirable for the patient, then τ_1 is shown to be more efficacious for positive outcome than τ_2, and if the outcome is undesirable for the patient, then τ_1 is shown to be less susceptible to this negative outcome than τ_2. Similarly, those in EQUITABLE are the trials for which τ_2 was shown to be equitable with τ_1 with respect to an outcome indicator, and those in INFERIOR are the trials for which τ_2 was shown to be superior to τ_1 with respect to an outcome indicator.

Given treatments τ_1 and τ_2, there are three possible interpretations of a set of items of evidence (i.e. a set of rows from an evidence table such as Table 1):

1. $\tau_1 > \tau_2$, meaning the evidence supports the claim that treatment τ_1 is superior to τ_2.
2. $\tau_1 \sim \tau_2$, meaning the evidence supports the claim that treatment τ_1 is equivalent to τ_2
3. $\tau_1 < \tau_2$, meaning the evidence supports the claim that treatment τ_1 is inferior to τ_2.

Any formula of the form $\tau_1 > \tau_2$, $\tau_1 \sim \tau_2$, and $\tau_1 < \tau_2$ is a **claim**, denoted by ϵ. We regard $\tau_1 > \tau_2$ as equivalent to $\tau_2 < \tau_1$, and $\tau_1 \sim \tau_2$ as equivalent to $\tau_2 \sim \tau_1$.

We use inference to derive a claim from a set of evidence. We use inference rules to define what are the allowed inferences. In this paper, we use three inference rules

Definition 6. *An* **inference rule** *is one of the following forms, where $X \subseteq$ EVIDENCE and $X \neq \emptyset$.*

1. *If $X \subseteq$ SUPERIOR, then $\tau_1 > \tau_2$.*
2. *If $X \subseteq$ EQUITABLE, then $\tau_1 \sim \tau_2$.*
3. *If $X \subseteq$ INFERIOR, then $\tau_1 < \tau_2$.*

For example, in the evidence given in Table 1, there is a subset $\{e_3, e_4\}$ of the evidence for which each item states that ACE is superior to CCB. From this subset, we may draw the conclusion that ACE is superior to CCB in general.

One can informally think of an argument comprising of a set of evidence (i.e. a subset of EVIDENCE), and a conclusion or claim that has been derived from the set of evidence using an inferential rule.

Definition 7. *An* **inductive argument** *is a tuple $\langle X, \epsilon\rangle$ such that ϵ follows from X using one of the three inferences rules given in Definition 6. We call X the support and ϵ the claim of the argument.*

Given a set EVIDENCE, let ARG(EVIDENCE) denote the set of inductive arguments that can be generated from the evidence according to the above definition.

Example 4. *Returning to the evidence in Table 1, concerning treatments* ACE *and* CCB*, we have* EVIDENCE $= \{e_1, e_2, e_3, e_4\}$, SUPERIOR $= \{e_3, e_4\}$, *and* INFERIOR $= \{e_1, e_2\}$. *From this, together with the inference rules, we get the following inductive arguments.*

$$\begin{array}{cc}
\langle\{e_3\}, \text{ACE} > \text{CCB}\rangle & \langle\{e_1\}, \text{ACE} < \text{CCB}\rangle \\
\langle\{e_4\}, \text{ACE} > \text{CCB}\rangle & \langle\{e_2\}, \text{ACE} < \text{CCB}\rangle \\
\langle\{e_3, e_4\}, \text{ACE} > \text{CCB}\rangle & \langle\{e_1, e_2\}, \text{ACE} < \text{CCB}\rangle
\end{array}$$

We refer to these arguments as inductive because from a set of evidence, we make the inductive inference concerning the claim. The better the evidence, the more likely the claim is correct. However, in general it is possible that it is incorrect, and therefore the claim is a defeasible inference. The way we manage this uncertainty is by considering counterarguments. For counterarguments, we will consider rebuttals and meta-arguments. We start by introducing rebuttals next.

Looking at Example 4, we see intuitively that the arguments with differing claims conflict. Obviously it cannot be the case that both of the arguments' claims are true. In this sense these arguments attack, or rebut, each other. We capture this relationship with the following definition. Note that this definition is symmetric, i.e., if A_i conflicts with A_j, then A_j conflicts with A_i.

Definition 8. *If the claim of argument A_i is ϵ_i and the claim of argument A_j is ϵ_j then we say that A_i* ***conflicts*** *with A_j each other whenever:*

1. *$\epsilon_i = \tau_1 > \tau_2$, and ($\epsilon_j = \tau_1 \sim \tau_2$ or $\epsilon_j = \tau_1 < \tau_2$).*

2. *$\epsilon_i = \tau_1 \sim \tau_2$, and ($\epsilon_j = \tau_1 > \tau_2$ or $\epsilon_j = \tau_1 < \tau_2$).*

3. *$\epsilon_i = \tau_1 < \tau_2$, and ($\epsilon_j = \tau_1 > \tau_2$ or $\epsilon_j = \tau_1 \sim \tau_2$).*

So when an argument A_i conflicts with an argument A_j, it denotes a rebuttal by A_i of A_j, and vice versa.

In order to take relative preferences over arguments into account, we can introduce a pre-order preference relation over arguments, and use this to define the attack relation between arguments, as proposed by Amgoud and Cayrol [16]. The next section is devoted to showing how we can define such a preference relation for taking into account the relative benefits of the treatments being considered.

Definition 9. *For any pair of arguments A_i and A_j, and a preference relation R, A_i* **attacks** *A_j with respect to R iff A_i conflicts with A_j and it is not the case that A_j is strictly preferred to A_i, according to R.*

Now we combine these components by defining an argument graph based on a set of trial results, a set of inference rules, and a preference relation over arguments as follows. For this, we assume a set containing two treatments $\{\tau_1, \tau_2\}$, which we call the topic, denoted Topic.

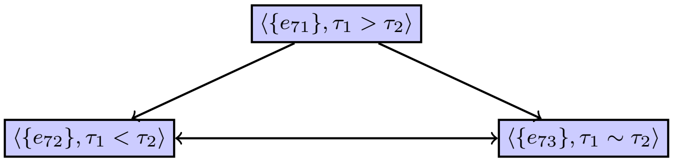
Definition 10. *Given a* Topic $= \{\tau_1, \tau_2\}$*, and a set* evidence*, an* **inductive argument graph** *is a graph where the set of nodes is the subset of* Arg(Evidence) *containing arguments with a claim in $\{\tau_1 > \tau_2, \tau_1 \sim \tau_2, \tau_1 < \tau_2\}$, and the set of arcs is the attacks relation given by Definition 9. We refer to the arguments in this graph as* Arg(Evidence,Topic)*.*

Clearly, an inductive argument graph is a particular kind of argument graph. Furthermore, it is an instance of a prioritized argument frameworks (PAFs) as proposed by Amgoud and Cayrol [16]. We leave the formalization of specific preference relations until the next section. In the meantime, we illustrate the use of an informally defined preference relation to get the following argument graph by applying the preference relation to a simple example of evidence.

Example 5. *Consider the following fictional evidence table, we get the argument graph below using the arguments with non-empty support.*

	Left	Right	Outcome indicator	Value	Net	Sig	Type
e_{71}	τ_1	τ_2	mortality	0.80	$>$	yes	RCT
e_{72}	τ_1	τ_2	palpitations	1.15	$<$	yes	NA
e_{73}	τ_1	τ_2	headaches	1.00	$\sim$	no	RCT

Suppose we prefer the argument with evidence showing superiority for the outcome indicator of "mortality" over other arguments (i.e. lower risk of mortality is preferred to lower risk of palpitations or equivalent risk of headaches). As a result, we get the following inductive argument graph.


$\langle\{e_{71}\}, \tau_1 > \tau_2\rangle$
$\langle\{e_{72}\}, \tau_1 < \tau_2\rangle$
$\langle\{e_{73}\}, \tau_1 \sim \tau_2\rangle$

When we consider an evidence table with just two treatments, as above, it is easy to see that the graph induced is tripartite, and its independent sets are given by those arguments with claim $\tau_1 > \tau_2$, those arguments with claim $\tau_1 \sim \tau_2$, and those arguments with claim $\tau_1 < \tau_2$.

We can directly use the dialectical semantics given by Dung [15] (i.e. Definition 5) to decide extensions of argument graphs. Here, there is one grounded and preferred extension and it contains just the argument $\langle\{\{e_{71}\}, \tau_1 > \tau_2\rangle$.

We regard a preferred extension as an interpretation of a set of EVIDENCE (i.e. an aggregation of the evidence in EVIDENCE). So if E is a preferred extension of the argument graph, and $A \in E$, and ϵ is the claim of A, then ϵ is a possible aggregation of the evidence. Furthermore, we regard a grounded extension as a higher quality interpretation than a preferred extension.

This section has provided a general framework for aggregating evidence concerning a pair of treatments according to multiple outcomes. To use the framework, a specific preference relation needs to be specified. In the next section, we consider specific criteria for preferring some arguments over others based on the relative benefits offered by the treatments.

7 Preferences over inductive arguments

Assuming a preference relation over arguments, as investigated by Amgoud and Cayrol [16], is a simple and intuitive idea. However, it does raise the question of where does the preference relation come from, what does it mean, what integrity constraints should we impose on it to ensure that it is sensible, and how can this be done in a way that is practical? We address these questions in this section.

7.1 Normalizing the benefits

We start by clarifying what the evidence is telling us about the treatments. Each item of evidence (i.e. each row in the evidence table) presents either a positive or negative effect of the left treatment *vis a vis* the right treatment. As we explained in Section 3, we assume that each evidence table that is given as input has a column called "Outcome indicator" and a column called "Value". The outcome indicator is what is being measured, and the value is the value of that measure.

Definition 11. *For an item of evidence e, the* **result of the evidence** *is the following pair,*

$$(OutcomeIndicator, Value)$$

where OutcomeIndicator is the entry for the column "Outcome Indicator" in the evidence table and $Value$ is the entry for the column "Value" in the evidence table.

Example 6. *Consider the following evidence table containing fictional evidence comparing use of the contraceptive pill (CP) with no contraception (NC). For this table, a user of contraceptive pills may be trading the benefit of a substantial reduction in risk of pregnancy against a small increased risk of breast cancer and thrombosis. Though there is also small positive effect for contraceptive users who get a reduced risk of ovarian cancer.*

	Left	Right	Outcome indicator	Value	Net	Sig	Type
e_{81}	CP	NC	breast cancer	1.04	<	yes	RCT
e_{82}	CP	NC	ovarian cancer	0.99	>	yes	MA
e_{83}	CP	NC	pregnancy	0.05	>	yes	RCT
e_{84}	CP	NC	thrombosis	1.02	<	yes	MA

For this evidence table, we have the following results

- e_{81} *has result* $(breast\ cancer, 1.04)$
- e_{82} *has result* $(ovarian\ cancer, 0.99)$

- e_{83} *has result* $(pregnancy, 0.05)$
- e_{84} *has result* $(thrombosis, 1.02)$

Because of our definition for the "Net" column, we know that the results for e_{81} *and* e_{84} *are negative results (from the point of view of the left treatment) and the results for* e_{82} *and* e_{83} *are positive results (from the point of view of the left treatment).*

If we have positive result for the left treatment (i.e. the entry for the "Net" column is $>$), then it means that the result says that the left treatment is better than the right treatment with respect to the outcome indicator (as given in the "Outcome Indicator" column), to the degree given by the value (as given in the "Value" column), and implicitly, it also means that the result is a negative result for the right treatment.

Similarly, if we have negative result for the left treatment (i.e. the entry for the "Net" column is $<$), then implicitly it means that the result says that the right treament is better than the left treatment with respect to the outcome indicator given in the "Outcome Indicator" column, to the degree given by the value given in the "Value" column, and implicitly, it also means that the result is a postive result for the right treatment.

From the result of each item of evidence in the support of an argument, we get the results of the support of the argument as follows.

Definition 12. *For an inductive argument* $A = \langle X, \epsilon \rangle$*, the* **results** *of* A*, denoted* $\mathsf{Results}(A)$*, is defined as follows*

$$\{(OutcomeIndicator, Value) \mid e \in X \text{ and } (OutcomeIndicator, Value) \text{ is the result of } e\}$$

Example 7. *Continuing with Example 6, there are six arguments that we can construct from this evidence table. Hence,* $\textsc{Arg}(\textsc{Evidence})$ *contains the following arguments.*

$$\begin{array}{ll} A_1 = \langle \{e_{82}, e_{83}\}, CP > NC \rangle & A_4 = \langle \{e_{81}, e_{84}\}, CP < NC \rangle \\ A_2 = \langle \{e_{82}\}, CP > NC \rangle & A_5 = \langle \{e_{81}\}, CP < NC \rangle \\ A_3 = \langle \{e_{83}\}, CP > NC \rangle & A_6 = \langle \{e_{84}\}, CP < NC \rangle \end{array}$$

So the sets of results are the following where oc is ovarian cancer, preg is pregnancy, bc is breast cancer, and th is thrombosis.

$$\begin{array}{l} \mathsf{Results}(A_1) = \{(ovarian\ cancer, 0.99), (pregnancy, 0.05)\} \\ \mathsf{Results}(A_2) = \{(ovarian\ cancer, 0.99)\} \\ \mathsf{Results}(A_3) = \{(pregnancy, 0.05)\}) \\ \mathsf{Results}(A_4) = \{(breast\ cancer, 1.04), (thrombosis, 1.02)\} \\ \mathsf{Results}(A_5) = \{(breast\ cancer, 1.04)\} \\ \mathsf{Results}(A_6) = \{(thrombosis, 1.02)\} \end{array}$$

So each argument has a set of evidence as support, and each item of evidence has a result (i.e. a $(OutcomeIndicator, Value)$ pair), and so each argument has an associated set of results (i.e. the results for the evidence in its support).

In order to compare sets of results, we need to normalize the results. To motivate this, consider the outcomes "breast cancer" and "ovarian cancer" in Example 6.

- For ovarian cancer, the relative risk is 0.99. This means that for every 100 women not taking the contraceptive pill who develop ovarian cancer, only 99 of them would develop ovarian cancer if they were on the contraceptive pill. So the tuple $(oc, 0.99)$ captures this as an advantage of the left arm over the right arm in the study, because the undesirable outcome has a reduced relative risk. This means we can regard the result $(oc, 0.99)$ as a benefit that we may want, and according to the evidence, the benefit comes from taking the left arm as opposed to the right arm.

- For breast cancer, the relative risk is 1.04. This means that 104 women taking the contraceptive pill got breast cancer for every 100 women not taking the contraceptive pill. So the tuple $(\text{breast cancer}, 1.04)$ captures this as a disadvantage of the left arm over the right arm in the study, because the undesirable outcome has a raised relative risk. This means we need to take the inverse of 1.04 (i.e. 0.96) for the value if we want to turn the result into a benefit that we may want, i.e. $(\text{breast cancer}, 0.99)$, and according to the evidence, the benefit comes from taking the right arm as opposed to the left arm.

So in general, a result $(OutcomeIndicator, Value)$ is a benefit in either of the following two cases

- If $OutcomeIndicator$ is an outcome indicator for something good (e.g. survival rate, successful treatment of infection, etc), and the assignment to $Value$ means that the left arm is better than the right arm (e.g. for an outcome indicator measured in terms of relative risk, the assignment to $value$ is greater than 1), then $(OutcomeIndicator, Value)$ is a benefit.
- If $OutcomeIndicator$ is an outcome indicator for something bad (e.g. death rate, incidence of infection, etc), and the assignment to $Value$ means that the left arm is better than the right arm (e.g. for an outcome indicator measured in terms of relative risk, the assignment to $value$ is less than 1), then $(OutcomeIndicator, Value)$ is a benefit.

Therefore, for any result that is not a benefit, we need to turn it into a benefit, by normalizing the value. How we normalize a value depends on how the outcome indicator is measured. If it is a relative risk, then it is just the inverse. So we assume a function $N_{OutcomeIndicator}$ that takes a value and returns the normalized value. For the example considered above for the breast cancer outcome indicator, we have $N_{breast\ cancer}(1.04) = 0.96$. As another example, in the Glaucoma case study, "change in IOP" is measured as a mean difference in the reduction of intraocular pressure. So a negative value means the left arm is better than the right arm. So to normalize a mean difference value, we turn a positive number into a negative number, and turn a negative number into a positive number. For instance, $N_{changeInIOP}(2.03) = -2.03$.

Next we need to know when to normalize results. For an argument that shows superiority, or equivalence, for the left arm over the right arm, (i.e. the claim is of the form $\tau_1 > \tau_2$ or $\tau_1 \sim \tau_2$), then we do not need to normalize, whereas when an argument that shows superiority for the right arm over the left arm, (i.e. the claim is of the form $\tau_1 < \tau_2$), then we do need to normalize. Hence, we define the benefits emanating from evidence in the support of arguments as follows.

Definition 13. *Let A be an inductive argument where $\mathsf{Claim}(A)$ is $\tau_1 > \tau_2$, $\tau_1 \sim \tau_2$, or $\tau_1 < \tau_2$. The* **Benefits** *function, denoted* $\mathsf{Benefits}$, *is defined as follows.*

$$\mathsf{Benefits}(A) = \begin{cases} \mathsf{Results}(A) \textit{ when } \mathsf{Claim}(A) \neq \tau_1 < \tau_2 \\ \mathsf{Normalize}(A) \textit{ when } \mathsf{Claim}(A) = \tau_1 < \tau_2 \end{cases}$$

where $\mathsf{Normalize}(A)$ *is the following set.*

$$\{(OutcomeIndicator, N_{OutcomeIndicator}(Value)) \mid (OutcomeIndicator, Value) \in \mathsf{Results}(A)\}$$

Example 8. *Continuing with Example 7, we get the following sets of benefits, where $N_{OutcomeIndicator}$ is the inverse function for each outcome indicator.*

$$\begin{aligned}
\mathsf{Benefits}(A_1) &= \{(\textit{ovarian cancer}, 0.99), (\textit{pregnancy}, 0.05)\} \\
\mathsf{Benefits}(A_2) &= \{(\textit{ovarian cancer}, 0.99)\} \\
\mathsf{Benefits}(A_3) &= \{(\textit{pregnancy}, 0.05)\}) \\
\mathsf{Benefits}(A_4) &= \{(\textit{breast cancer}, 0.96), (\textit{thrombosis}, 0.98)\} \\
\mathsf{Benefits}(A_5) &= \{(\textit{breast cancer}, 0.96)\} \\
\mathsf{Benefits}(A_6) &= \{(\textit{thrombosis}, 0.98)\}
\end{aligned}$$

So for an argument with claim $\tau_1 < \tau_2$, the results give the reasons why τ_1 is inferior to τ_2. By normalizing the results from the argument, we get the reasons why τ_2 is superior to τ_1.

7.2 Defining the preference relation

Our approach to defining a preference relation over arguments is based on defining a preference relation over the benefits offered by the treatments. In other words, for a pair of arguments A_i and A_j, we will ascertain that A_i is preferred to A_j when the benefits of A_i are preferred to the benefits of A_j (i.e. $\mathsf{Benefits}(A_i)$ are preferred to the benefits of $\mathsf{Benefits}(A_j)$). To do this, we need to clarify the intuition of having a preference relation over sets of benefits. We proceed with an example.

Example 9. *We return to Example 5. Here, we see that e_{71} has that τ_1 is superior to τ_2 because the relative risk of mortality is 0.8, whereas e_{72} has that τ_1 is inferior to τ_2 because relative risk of palpitation is 1.15. In other words, we have the results, where the first is a positive result for the left treatment, and the second is a negative result for the left treatment.*

- *e_{71} has result $(mortality, 0.8)$*
- *e_{72} has result $(palpitations, 1.15)$*

The first result is a substantial positive result for the left arm whereas the second is a modest positive result for the right arm. So after normalization it is reasonable to express the follow preference over the following two sets of benefits.

$$\{(mortality, 0.8)\} \textit{ is preferred to } \{(palpitations, 0.86)\}$$

This preference relation over benefits means that reduction in mortality (relative risk 0.8) is preferred to a reduction in palpitations (relative risk 0.86). Hence, for $A_1 = \langle\{e_{71}\}, \tau_1 > \tau_2\rangle$, and $A_2 = \langle\{e_{72}\}, \tau_1 < \tau_2\rangle$, we have

$$\mathsf{Benefits}(A_1) \textit{ is preferred to } \mathsf{Benefits}(A_2)$$

We assume that the preference relation over sets of benefits comes from the user. We expect the user bases the choices of which benefits are preferred by using subjective beliefs and judgments. In general, this subjectivity is unavoidable. Different people prioritize different things. If we return to Example 6, then for some people, a substantial reduction in the risk of pregnancy, such as with the benefits set $\{(pregnancy, 0.05)\}$, is preferred to a small decrease in risk of breast cancer and thrombosis, such as with the benefits set $\{(breastcancer, 0.96), (thrombosis, 0.98)\}$. Of course, there are some people who would have the opposite preference.

The idea of a preference relation over sets of benefits is that it focuses the attention on the outcomes, and allows for preferences to be assigned without consideration of the evidence that gives rise to it. We would expect that it is used to take into account both the outcome indicator and the magnitude of that indicator.

Example 10. *Consider the arguments $A_1 = \langle\{e_{86}\}, \tau_{51} < \tau_{52}\rangle$ and $A_2 = \langle\{e_{87}\}, \tau_{51} > \tau_{52}\rangle$ which can be formed from the evidence table below.*

- $\mathsf{Benefits}(A_1) = \{(overallsurvival, 1.02)\}$
- $\mathsf{Benefits}(A_2) = \{(diseasefreesurvival, 12.05)\}$

The outcome indicators are both measured in terms of relative risk. Here, τ_{52} is slightly better at overall survival, but τ_{51} over 12 times better at disease free survival. Here, it may be much better to accept a small loss in overall survival (i.e. there is a slightly greater probability that the patient will not survive the period) to gain a massive increase in disease free survival (i.e. there is a substantially greater probability that the patient will not have a recurrence of the disease in the period). This may be particularly desirable if the disease itself has painful or highly undesirable symptoms. Hence, we may adopt the preference that $\mathsf{Benefits}(\langle\{e_{87}\}, \tau_{52} > \tau_{51}\rangle)$ is preferred to $\mathsf{Benefits}(\langle\{e_{86}\}, \tau_{51} > \tau_{52}\rangle)$. Of course, other users may choose the opposite preference.

	Left	Right	Outcome indicator	Value	Net	Sig	Type
e_{86}	τ_{51}	τ_{52}	overall survival	0.98	$<$	yes	RCT
e_{87}	τ_{51}	τ_{52}	disease free survival	12.05	$>$	yes	RCT

Even though we assume that the preference relation over sets of benefits comes from the user, we expect that most choices would be common amongst many people. This means that a default preference relation could be generated by the system, and adapted by the user, thereby avoiding the need to get a lot of information from the user. For instance, for any outcome indicator concerning survival, denoted S, and any minor side-effect, denoted M, we would expect that most people would have that (S, x) is preferred to (M, y) when x and y are of the same order of magnitude, both x and y are relative risk, and $x < 1$.

In general, we would want the preference over sets of benefits to conform to various domain-specific constraints. For example, if we have $\{(\mathit{pregnancy}, 0.05)\}$ preferred to $\{(\mathit{thrombosis}, 0.98)\}$, then we would have $\{(\mathit{pregnancy}, 0.01)\}$ preferred to $\{(\mathit{thrombosis}, 0.98)\}$, since a risk ratio of 0.01 for pregnancy is better than a risk ratio of 0.05 for pregnancy. Similarly, we can see these relationships holding for some outcomes as well values. Let $OS1yr$ denote the outcome of overall survival at 1 year, and let $OS5yr$ denote the outcome of overall survival at 5 years. So if we have $\{(\mathit{OS1yr}, 1.05)\}$ preferred to $\{(\mathit{thrombosis}, 0.98)\}$, then we would have $\{(\mathit{OS5yr}, 1.05)\}$ preferred to $\{(\mathit{thrombosis}, 0.98)\}$, since if it is better to have a relative risk of survival at 1 year of 1.05, over an alternative, then it is better to have a relative risk of survival at 5 years of 1.05, over the same alternative.

For capturing the preference relation over sets of benefits, we assume that Indicators is the domain for the outcome indicator attribute (e.g. breast cancer, pregnancy, thrombosis, etc), and Reals is the set of real numbers which we use as the domain for the value attribute. Then, we let Results be Indicators × Reals. So, for each argument A_i, we assume that $\mathsf{Benefits}(A_i) \subset$ Benefits.

Definition 14. *A* **benefits preference relation** *, denoted $\succeq$, is a relation $\succeq \subseteq \wp(\textsc{Results}) \times \wp(\textsc{Results})$ that satisfies the properties of reflexivity, transitivity, monotonicity, weakening, and strengthening as defined in Table 2.*

For arguments, A_i, A_j, $\mathsf{Benefits}(A_i) \succeq \mathsf{Benefits}(A_j)$ means that the results of A_i are preferred to the results of A_j. As we have said above, we expect the user to give the benefits preference relation, or at least agree to a default benefits preference relation, for a specific evidence table. So we impose the condition that any benefits preference relation satisfies the properties in Table 2. If it does, then it means that the preference relation will be a pre-order relation since if satisfies montonicity and transitivity. As an abbreviation, we also assume the following.

- $\mathsf{Benefits}(A_i) \succ \mathsf{Results}(A_j) = (\mathsf{Benefits}(A_i) \succeq \mathsf{Benefits}(A_j)$ & $\mathsf{Benefits}(A_j) \not\succeq \mathsf{Benefits}(A_i))$.
- $\mathsf{Benefits}(A_i) \sim \mathsf{Results}(A_j) = (\mathsf{Benefits}(A_i) \succeq \mathsf{Benefits}(A_j)$ & $\mathsf{Benefits}(A_j) \succeq \mathsf{Benefits}(A_i))$.

For presentational purposes, in the examples, we only explicitly give the preference relation for pairs of arguments that conflict. The remainder of the relation can be obtained by applying the properties given in Table 2.

Example 11. *Continuing Example 8, using the arguments A_1 to A_6, the following is a benefits preference relation.*

$$\begin{array}{lll}
\mathsf{Benefits}(A_1) \succ \mathsf{Benefits}(A_4) & \mathsf{Benefits}(A_4) \succ \mathsf{Benefits}(A_2) & \mathsf{Benefits}(A_3) \succ \mathsf{Benefits}(A_4) \\
\mathsf{Benefits}(A_1) \succ \mathsf{Benefits}(A_5) & \mathsf{Benefits}(A_2) \sim \mathsf{Benefits}(A_5) & \mathsf{Benefits}(A_3) \succ \mathsf{Benefits}(A_5) \\
\mathsf{Benefits}(A_1) \succ \mathsf{Benefits}(A_6) & \mathsf{Benefits}(A_2) \succ \mathsf{Benefits}(A_6) & \mathsf{Benefits}(A_3) \succ \mathsf{Benefits}(A_6)
\end{array}$$

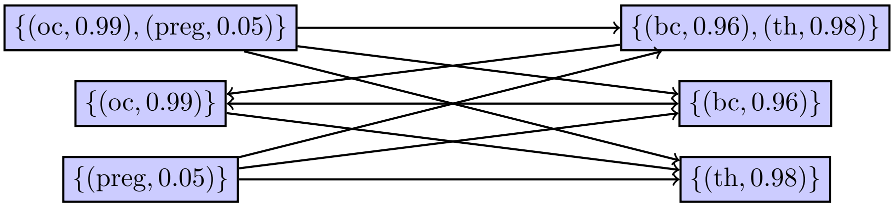
In order to illustrate the benefits from each argument, we form a benefits graph as follows. Each node is the benefits for an argument and each arc denotes that the benefits for the first node are preferred to the benefits of the second.

Property	Definition
Reflexivity	$\mathsf{Benefits}(A_i) \succeq \mathsf{Benefits}(A_i)$
Transitivity	$\mathsf{Benefits}(A_i) \succeq \mathsf{Benefits}(A_j)$ and $\mathsf{Benefits}(A_j) \succeq \mathsf{Benefits}(A_k)$ implies $\mathsf{Benefits}(A_i) \succeq \mathsf{Benefits}(A_k)$
Monotonicity	$\mathsf{Support}(A_i) \subseteq \mathsf{Support}(A_j)$ implies $\mathsf{Benefits}(A_j) \succeq \mathsf{Benefits}(A_i)$
Weakening	$\mathsf{Support}(A_i) \subseteq \mathsf{Support}(A_j)$ and $\mathsf{Benefits}(A_k) \succeq \mathsf{Benefits}(A_j)$ implies $\mathsf{Benefits}(A_k) \succeq \mathsf{Benefits}(A_i)$
Strengthening	$\mathsf{Support}(A_i) \subseteq \mathsf{Support}(A_j)$ and $\mathsf{Benefits}(A_i) \succeq \mathsf{Benefits}(A_k)$ implies $\mathsf{Benefits}(A_j) \succeq \mathsf{Benefits}(A_k)$

Table 2: Properties of the benefits preference relation (Definition 14) that hold for all arguments $A_i, A_j, A_k \in \textsc{Arg}(\textsc{Evidence})$. For an argument $A = \langle X, \epsilon \rangle$, $\mathsf{Support}(A) = X$. We explain these properties as follows: (Reflexivity) This ensures that the relation makes every benefits set be equally preferred to itself; (Transitivity) This ensures that if benefits set R_i is preferred to benefits set R_j, and benefits R_j is preferred to benefits R_k, then R_i is preferred to R_k (Monotonicity) This ensures that a benefits set subsumes another, then it will be more preferred; (Weakening) This ensures that when a benefits set R_i is dominated by another benefits set R_k, then R_k will be preferred to any subset of R_i; (Strengthening) This ensures that when a benefits set R_i is preferred to another R_k, then adding further benefits to it, to give a larger results set R_j, will not affect its preference over R_k. Note, weakening and strengthening follow from transitivity and monotonicity.

Definition 15. *For $\Phi \subseteq \textsc{Arg}(\textsc{Evidence},\textsc{Topic})$, the set of nodes of the benefits graph are $\{\mathsf{Benefits}(A_i) \mid A_i \in \Phi\}$. For each pair of arguments $A_i, A_j \in \Phi$ such that A_i and A_j conflict, there is an arc from $\mathsf{Benefits}(A_i)$ to $\mathsf{Benefits}(A_j)$ iff $\mathsf{Benefits}(A_i) \succeq \mathsf{Benefits}(A_j)$.*

Example 12. *Continuing Example 11, using the arguments A_1 to A_6, the following is a benefits graph corresponding to the benefits preference relation.*



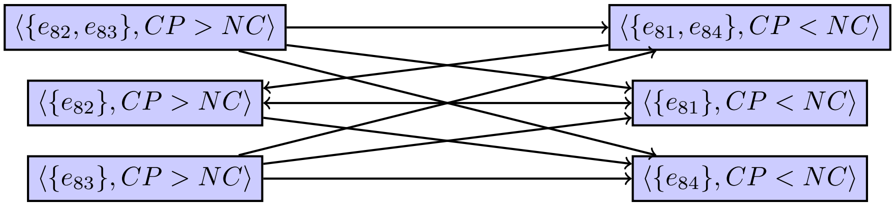
As we explained above, we use the preference relation over sets of benefits to define the preference relation over arguments. With that, we can then form an argument graphs, as suggested in the previous section, and this will constitutes a prioritized argument framework (as defined by Amgoud and Cayrol [16]).

Definition 16. *For arguments $A_i, A_j \in \textsc{Arg}(\textsc{Evidence},\textsc{Topic})$, such that A_i conflicts with A_j, and a benefits preference relation $\succeq$, the* **argument preference relation** *is defined as follows.*

$$A_i \textit{ is preferred to } A_j \textit{ iff } \mathsf{Benefits}(A_i) \succeq \mathsf{Benefits}(A_j)$$

For the rest of this paper, we assume that given a pair of treatments in Topic *and an evidence table* Table*, the inductive argument graph (specified in Definition 10) uses this definition for the preferences over arguments.*

Example 13. *Continuing Example 12, the following is the inductive argument graph constructed according to Definition 10.*


$\langle\{e_{82}, e_{83}\}, CP > NC\rangle$
$\langle\{e_{81}, e_{84}\}, CP < NC\rangle$
$\langle\{e_{82}\}, CP > NC\rangle$
$\langle\{e_{81}\}, CP < NC\rangle$
$\langle\{e_{83}\}, CP > NC\rangle$
$\langle\{e_{84}\}, CP < NC\rangle$

For small sets of arguments, it is possible to define directly the benefits preference relation. However, it is often easier to define one or more rules to define the benefits reference relation. For a modest set of outcome indicators, it is straightforward to identify a set of rules that would automatically generate the preference relation over arguments. The following is for a simple rule that just considers the cardinality of each benefits set.

Definition 17. *For arguments A_i and A_j, a benefits preference relation is a* **cardinality preference relation** *if it satisfies the following condition.*

$$\textit{If } |\mathsf{Benefits}(A_i)| \geq |\mathsf{Benefits}(A_j)|, \ \textit{then } \mathsf{Benefits}(A_i) \succeq \mathsf{Benefits}(A_j)$$

The above is useful when the outcome indicators are approximately equally important in terms of having a positive or negative effect, and the values are similar in magnitude. We will use the above to define the preference relation for the second and third case studies in Section 10.

In the following definition, we give a rule for specifying the preference relation that is specifically for the glaucoma case study in Section 10.1. Again it assumes that the values are similar in magnitude, but it divides the outcome indicators into two classes. The first class is for intended positive effects (i.e. an improvement in the medical problem being treated), and the second class is for modest negative effects (i.e. side-effects that are tolerable as long as there is a positive effect). We give an example of using the glaucoma preference relation in Figure 2.

Definition 18. *For arguments A_i and A_j, a benefits preference relation, $\succeq$, is a* **glaucoma preference relation** *if it satisfies the following condition.*

If there is a $(OutInd, Val) \in \mathsf{Benefits}(A_i)$ *such that* $Val > 1$
    and $OutInd \in \{$*visual field prog, change in IOP, acceptable IOP, IOP > 35mmHg*$\}$
and for all $(OutInd, Val) \in \mathsf{Benefits}(A_j)$
    $OutInd \in \{$*respiratory prob, cardio prob, allergy prob, hyperaemia, drowsiness*$\}$
then $\mathsf{Benefits}(A_i) \succeq \mathsf{Benefits}(A_j)$

We could refine this rule by specifying bounds on the values, and even reversing the preference if say all the intended positive effects were small (for example, for relative risk, the value is between 0.95 and 1, thereby suggesting there is little difference in relative risk), and one or more side-effects effects were large (for example, for relative risk of allergy problem, the value is 50, thereby suggesting that there is a 50 fold increase in risk of allergy problems).

To conclude this section, we have introduced a way of defining a preference relation over inductive arguments in terms of the evidence used in the support of each inductive argument. This approach is based on defining a preference relation over sets of benefits. In this way, we compare sets of outcome indicators and their magnitude. This allows for a simple and intuitive approach to capturing subjective criteria. We have assumed that the benefits preference relation is given by the user, or at least agreed to by the user, and so it an important part of the input to the argument-based aggregation process. We have also assumed that any benefits preference relation satisfies the properties of reflexivity, transitivity, monotonicity, strengthening, and weakening. In order to render the approach practical, we can use rules to define specific benefits preference relations such as the cardinality preference relation (Definition 17) and the glaucoma preference relation (Definition 18).

8 Generating meta-arguments

The general definition of an argument graph permits any kind of argument in the graph. In our framework, we have focused so far on one type of argument which is based on inductive inference. In this section, we consider further kinds of argument that we call meta-arguments. These are arguments against the quality of the evidence used in the inductive arguments. We present them as atomic (i.e. there is no internal structure to them), and they will be used as counterarguments to inductive arguments. Examples of meta-arguments that we can consider include the following.

- 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.

So each of these meta-arguments attacks an inductive argument on the basis of the evidence used in its support. We will only define a few of the meta-arguments that we may use in this framework.

Definition 19. *For $A \in \textsc{Arg}(\textsc{Evidence})$, if there is an $e \in \textsc{Support}(A)$ such that e is not statistically significant, and the outcome indicator of e is not a side-effect, then the following is a meta-argument that attacks A.*

$$\langle \texttt{Not statistically significant} \rangle$$

For the above definition, we assume that we have categorized some outcome indicators as "side effect". Normally, randomized clinical trials (which are regarded by many clinicians as the most reliable form of evidence) are set up to determine whether one treatment is better than another for addressing a specific clinical condition. It is unusual for randomized clinical trials to be set up to determine whether one treatment is better than another with respect to specific side effects. This is partly because the occurrence of side effects can be quite low, and therefore very large, and therefore very expensive, trials are needed to study them. So when a trial is run, side-effects can be studied, and reported on, as a secondary goal. As a consequence, reports of side-effects are frequently not statistically significant. Yet, for guideline recommendations, it can be important to take the non-significant reports of side-effects into account. With this in mind, the above meta-argument only attacks arguments where there is evidence outcome indicators that are not side-effects and not statistically significant. We illustrate this in the next example.

Example 14. *Consider Table 1. For this, $\langle \{e_1, e_2\}, ACE < CCB \rangle$ is an inductive argument. Since, e_1 is not statistically significant, and the outcome indicator of e_1 is not a side-effect, then $\langle$*`Not statistically significant`*$\rangle$ is a meta-argument that attacks it.*

For the next two forms of meta-argument, we assume that the evidence table has attributes concerning quality of the trials, as for example in Table 5.

Definition 20. *For $A \in \textsc{Arg}(\textsc{Evidence})$, if there is an $e \in \textsc{Support}(A)$ such that e is a non-randomized and non-blind trial, then the following is a meta-argument that attacks A.*

$$\langle \texttt{Non-randomized \& non-blind trials} \rangle$$

Definition 21. *For $A \in \textsc{Arg}(\textsc{Evidence})$, if $\textsc{Support}(A) = \{e\}$ such that e is a meta-analysis that concerns a narrow patient group then the following is a meta-argument that attacks A.*

$$\langle \texttt{Meta-analysis for a narrow patient group} \rangle$$

We now pull together the definitions for using inductive arguments and for using meta-arguments.

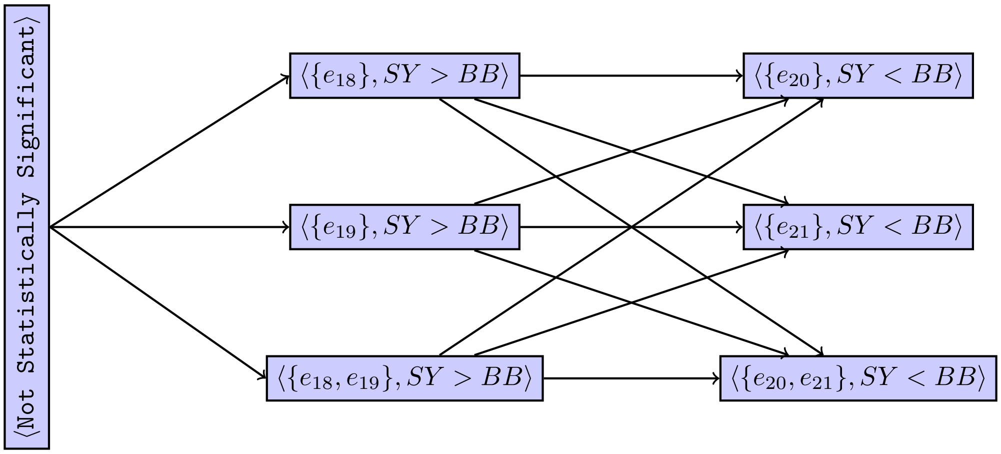

⟨Not Statistically Significant⟩
$\langle\{e_{18}\}, SY > BB\rangle$
$\langle\{e_{19}\}, SY > BB\rangle$
$\langle\{e_{18}, e_{19}\}, SY > BB\rangle$
$\langle\{e_{20}\}, SY < BB\rangle$
$\langle\{e_{21}\}, SY < BB\rangle$
$\langle\{e_{20}, e_{21}\}, SY < BB\rangle$

Figure 2: Evidential argument graph generated for evidence from Table 3 with the treatments SY and BB. There is one grounded extension of this argument graph which contains just the meta-argument and the inductive arguments $\langle\{e_{20}\}, SY < BB\rangle$, $\langle\{e_{21}\}, SY < BB\rangle$, and $\langle\{e_{20}, e_{21}\}, SY < BB\rangle$.

Definition 22. *Given a set* Evidence*, a set* Topic*, a benefits preference relation* $\succeq$ *(such as Definition 17 or Definition 17), and a set of definitions for meta-arguments (such as Definitions 19 – 21), let* iGraph *be the inductive argument graph constructed using Definition 10. The* **evidential argument graph***, denoted* eGraph*, is the smallest graph that satisfies the following two conditions:*

- *Each argument and each attack in* iGraph *is in* eGraph*.*
- *For each argument A in* iGraph*, if M is a meta-argument that attacks A, then M is an argument in* eGraph *and M attacks A is in* eGraph*.*

An evidential argument graph is an argument graph. It is a directed graph where each node is either an inductive argument or a meta-argument, and each arc is either an attack by a preferred inductive argument or an attack by a meta-argument. We give an example of an evidential argument graph in Figure 2.

So for each set of evidence, and each topic, Definition 22 gives us the argument graph we require for aggregating the evidence. In the next section, we show how we obtain the aggregation result from this graph.

9 Aggregating evidence

Now, we consider how we harness the ideas introduced so far for systematically aggregating evidence. We assume that we have a set Evidence, and we want to find the best treatments amongst those that are considered in the evidence. Let Treatments be the set of treatments that occur either as a left or right arm of an item of evidence in Evidence. Then for each pair of treatments $\tau_1, \tau_2 \in$ Treatments, we let Topic $= \{\tau_1, \tau_2\}$, and then we construct an evidential argument graph eGraph from Evidence and Topic. We use the following criteria for interpreting an argument graph that has been generated from an evidence table, and thereby show how we obtain an aggregation of that evidence.

- If there is a non-empty grounded extension (see Definition 5), and ϵ is the claim of the arguments in the extension (note, all arguments in a grounded or preferred extension will have the same claim), the result of the aggregation is ϵ.

- If there is an empty grounded extension (see Definition 5), then there are multiple preferred extensions (say $E_1, ..., E_n$), and so the result of the aggregation is ϵ_1 or ... or ϵ_n where ϵ_1 is the claim of the arguments in E_1 and ... and ϵ_n is the claim of the arguments in E_n.

We use superiority graphs as a way to summarize the analysis and to hang the evidential argument graph for each pairwise comparison. A superiority graph for an evidence table Evidence is a graph where each node is a treatment and there are two types of arc. The first type of arc is *strict superiority*, which when from τ_1 to τ_2 denotes that τ_1 is superior to τ_2, and the second type of arc is *equivalence* which when connecting τ_1 and τ_2 with a bidirectional arc denotes that τ_1 is equivalent to τ_2. To obtain this superiority graph, we require the following subsidiary function.

Definition 23. *For* Evidence *and* Topic $= \{\tau_1, \tau_2\}$, *let* eGraph *be the evidential argument graph given by Definition 22. The function* Winner(Topic,eGraph) *returns* $\{(\tau_1, \tau_2)\}$, $\{(\tau_2, \tau_1)\}$, *or* $\{(\tau_1, \tau_2), (\tau_2, \tau_1)\}$ *as follows.*

- *If there a grounded extension of* eGraph *where the claim of the inductive arguments is* $\tau_1 > \tau_2$ *then return an arc from* τ_1 *to* τ_2 *(i.e.* $\{(\tau_1, \tau_2)\}$*)*
- *Else if there a grounded extension of* eGraph *where the claim of the inductive arguments is* $\tau_1 < \tau_2$ *then return an arc from* τ_2 *to* τ_1 *(i.e.* $\{(\tau_2, \tau_1)\}$*)*
- *Else return a bidirectional arc between* τ_1 *and* τ_2 *(i.e.* $\{(\tau_1, \tau_2), (\tau_2, \tau_1)\}$*)*

A bidirectional arc is returned when the grounded extension of eGraph contains arguments with the claim $\tau_1 \sim \tau_2$, or when there is a preferred extension of eGraph containing no inductive arguments, or when there are multiple preferred extension of eGraph one of which contains inductive arguments with claim ϵ and another contains inductive arguments with claim ϵ' where ϵ and ϵ' conflict (e.g. ϵ is $\tau_1 > \tau_2$ and ϵ' is $\tau_1 > \tau_2$).

Example 15. *Consider the evidential argument graph* eGraph *given in Figure 2, where* Evidence *is listed in Table 3, and* Topic *is* $\{SY, BB\}$. *There is one grounded extension of this argument graph which contains just the meta-argument and the inductive arguments* $\langle\{e_{20}\}, SY < BB\rangle$, $\langle\{e_{21}\}, SY < BB\rangle$, *and* $\langle\{e_{20}, e_{21}\}, SY < BB\rangle$. *Therefore, the function* Winner(Topic,eGraph) *returns an arc from BB to SY.*

A superiority graph for an evidence table Evidence is a graph where each node is a treatment in Treatments, and for each pair τ_1, τ_2, if the comparison is non-empty (i.e. there is evidence in the set Evidence comparing τ_1 and τ_2), then there is an arc in the superiority graph between τ_1 and τ_2 given by Winner(Topic,eGraph), otherwise there is no arc between them.

So we have now completed the details of the process we outlined in Section 2. The input to the process is a set Evidence and the output is a superiority graph. The intermediate outputs produced by this process are summarized in Figure 3. Note, we regard the superiority graph as a summary of the argumentation. Therefore, the argument graph for each arc in the superiority graph should be available to the user so that they can see the reasons for the summary.

10 Case studies

To illustrate the aggregation of evidence in our framework, we report on case studies on treatments for glaucoma, for hypertension, and for pre-eclampsia. The outputs from our system are given in the superiority graphs in Figure 1.

10.1 Glaucoma case study

This case study concerns treatments for raised intraocular pressure (raised IOP), which is raised pressure in the eye. The evidence, which is presented in Table 3, was obtained from the NICE

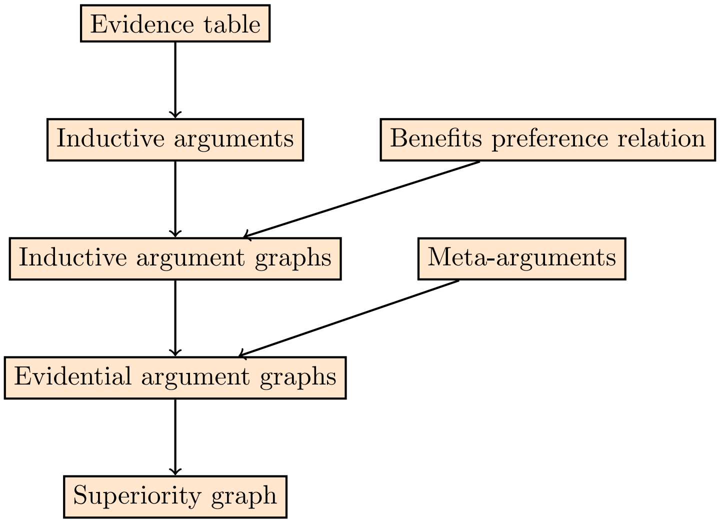

Evidence table
Inductive arguments
Benefits preference relation
Inductive argument graphs
Meta-arguments
Evidential argument graphs
Superiority graph

Figure 3: 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 inductive 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 inductive arguments.. One inductive argument attacks another if their claims conflict, and the benefits of the first argument are preferred to the second. Each meta-argument attacks an inductive argument when there is a weakness in the quality of the evidence used in the inductive 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.

Glaucoma Guideline [9]. We used Definition 18 for the benefits preference relation, and we used Definitions 19 – 21 for the meta-arguments.

We undertook a pairwise comparison of the five treatment options (i.e. beta-blockers versus no-treatment, prostaglandin analogues versus beta-blockers, prostaglandin analogues versus sympathomimetics, carbonic anhydrase inhibitors versus no treatment, carbonic anhydrase inhibitors versus beta-blockers, and sympathomimetics versus beta-blockers). We only considered these six comparisons because the guideline only has evidence that considers these pairs. For each of these comparisons, we generated an argument graph, and determined the arguments in the preferred or grounded extensions for each of these comparisons.

For this case study, we obtained the aggregations of the evidence concerning the treatment options that we summarize in Figure 1a. As with the NICE guideline, we obtained prostaglandin analogues as being superior. However, the NICE guideline also used an economic model to recommend beta-blocker for less serious cases (defined as being patients with a less thin central cornea).

10.2 Hypertension case study

This case study concerns treatments for hypertension, which is raised blood pressure. The evidence, which is presented in Table 4, where the evidence was obtained from the NICE Hypertension Guideline [10]. We used Definition 17 for the benefits preference relation, and we used Definitions 19 – 21 for the meta-arguments.

For this, we obtained the aggregations of the evidence concerning the treatment options which we summarize in Figure 1b. We compared this with the NICE guideline [10].

In the guideline, beta-blockers are regarded as inferior to the other treatments for hypertension, and this is reflected in our superiority graph, where beta-blockers are shown to not be superior to any of the alternatives. Beta-blockers are shown to be equivalent to thiazide-type diuretics, but by inspection of the argumentation, there is only one clinical trial comparing them and this was not statistically significant (i.e. the evidence concerning their direct comparison is insufficient to differentiate them). According to our superiority graph, ACE inhibitors were shown to be superior. In the guideline, ACE inhibitors are the recommended treatment for under 55s. In order to make this qualification to under 55s, the authors of the guideline took further factors into account including adverse events data that was not presented in the guideline.

10.3 Pre-eclampsia case study

This case study concerns treatments for pre-eclampsia, which is hypertension arising during pregnancy. The evidence, which is presented in Table 5, where the evidence was obtained from the NICE Hypertension in Pregnancy Guideline [11]. We used Definition 17 for the benefits preference relation, and we used Definitions 19 – 21 for the meta-arguments.

For this case study, we obtained the aggregations of the evidence concerning the treatment options which we summarize in Figure 1c. We compared this with the NICE guideline [11].

In the guideline, the recommendation is that women at high risk of pre-eclampsia, and women with more than one moderate risk factor, should be advised to take 75 mg of aspirin daily from 12 weeks until the birth of the baby. Women at high risk are those with any of the following: hypertensive disease during a previous pregnancy, chronic kidney disease, autoimmune disease such as systemic lupus erythematosis or antiphospholipid syndrome, type 1 or type 2 diabetes, or chronic hypertension. Factors indicating moderate risk are first pregnancy, age 40 years or older, pregnancy interval of more than 10 years, body mass index (BMI) of 35 kg/m or more at first visit, family history of pre-eclampsia, and multiple pregnancy.

As can be seen from the superiority graph, we also identified aspirin as best treatment. Though there is insufficient evidence in the evidence table for our approach to only make recommendations for women in a high risk category or for women with more than one factor indicating high risk. We believe that these restrictions on the recommendations in the NICE guideline come from the experience and broader medical knowledge of the guideline developers rather than from the evidence-base.

11 Managing subjective 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 benefits preference relation), and a set of meta-arguments. However, in practice it is normally not obvious that there is a single benefits preference relation or a single set of meta-arguments. This is because, in general, the selection of a benefits preference relation, and the selection of meta-arguments, are subjective criteria. Different clinicians, or their patients, may have different benefits preference relations. This is an intrinsic and unavoidable feature of dealing with preferences over benefits. 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 benefits 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 benefits preference relation, 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 benefits preference relation or a set of meta-arguments that is always the right choice, different combinations of benefits 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 benefits preference relations and/or sets of meta-arguments. Furthermore, if the superiority graph changes little over a wide range of sensible benefits 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 benefits preference relation or meta-arguments.

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 the systematic review that has resulted from this 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.

With further case studies, we should be able to build a library of benefits preference relations and meta-arguments that can be used by others, thereby facilitating the use of the framework. In addition, we can measure how sensitive a superiority graph is to the choice of benefits preference relation and meta-arguments, for instance using measures of graph difference.

In general, we believe that a benefits 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 benefits 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 benefits 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 benefits 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.

Another kind of subjectivity in the aggregation process, whether in the existing approaches used by guideline developers and systematic reviewers, or in our framework, concerns the representation of evidence in the evidence table. There are two dimensions as listed below to this issue.

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.

Relaxation analysis offers a valuable tool for analyzing clinical evidence in order to make more insightful and robust recommendations, but there are ontological and wider knowledge representation issues to do with what relaxations should be considered (i.e. when two conditions are equivalent). To address this, we can couple the construction of arguments with an ontology based on description logic 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 have undertaken a theoretical analysis of how this may be done [4], and we would now like to harness this for developing our sensitivity analysis of superiority graphs. By grouping evidence, we may be able identify more robust superiority graphs, but it may mean that we loose discriminations between treatments (and thereby have too many pairwise comparisons resulting in the pair of treatments being "equal").

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 benefits 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).

12 Implementation issues

In order to investigate the viability of our framework, we have implemented a prototype system that takes an evidence table as input, and constructs a superiority graph as output. It is implemented in Python, and is available as a Google Code project[1]. Currently it has been coded for a specific case study in lung cancer chemo-radiotherapy. It incorporates a range of benefits preference relations and meta-arguments appropriate for the case study. These are hard-coded in the prototype, and any combination of them can be selected.

The input, given as a spreadsheet, is used to construct an argument graph for each pair of treatments that are compared in the evidence. The spreadsheet involves 37 pairwise clinical trials, and each trial considers a number of different outcome indicators. For example, survival after 1 year, survival after 2 years, survival after 3 years, death rate resulting from treatment,

[1] `http://code.google.com/p/cafe-computational-analysis-framework-for-evidence/`

toxicity, etc. This means that we actually have 283 items of evidence (where each item of evidence concerns a pairwise comparison according to a single outcome indicator, as discussed in Section 3). We also considered a number of relaxations of the treatments, which essentially meant grouping some treatments as being the same. Relaxation involved editing the spreadsheet directly, thereby creating a new version of the spreadsheet, and this was done by hand.

With the evidence for the lung cancer chemo-radiotherapy case study, we were able to investigate the viability of the algorithms for generating the inductive arguments, the meta-arguments, and the attack relation (using the selected preference relation). Because we considered relaxations, and different combinations of benefits preference relations and sets of meta-arguments, we constructed over 200 argument graphs. The majority of these argument graphs had between 10 and 25 arguments, though over 10% of the argument graphs had over 250 arguments, and we did generate a few argument graphs with over 8000 arguments.

Because each of the argument graphs conforms to a simple structure (given by Definition 22), the algorithm (to construct each argument graph and to determine its grounded extension) has been designed to deal with exactly this structure. Using a standard desktop PC, for argument graphs with approximately 20 arguments, the time was around 0.01 sec, with approximately 40 arguments, the time was around 0.05 sec, with approximately 400 arguments, the time was around 1.5 sec, with approximately 4000 arguments, the time was around 25 sec, and with approximately 8000 arguments, the time was around 415 sec.

Whilst the time taken by the algorithm appears to be exponential in the number of arguments, this preliminary implementation indicates that our framework is viable for the number of arguments that would arise in many cases in practice. Furthermore, the implementation is based on a naive algorithm, and it is reasonable to believe that with some further development, substantial improvements in the temporal performance could be achieved. In addition, the volume of evidence in the lung cancer chemo-radiotherapy case study is representative of a large number of situations where evidence is aggregated. Commonly for systematic reviews, and guideline developments, where the superior treatment is sought from a class of treatments, 10s or 100s of items of evidence are used, where each item is a pairwise comparison according to a single outcome indicator.

13 Discussion

In this section, we draw conclusions on our proposal, discuss related literature, and discuss some future directions for further research.

13.1 Conclusions on our proposal

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.

In this paper, we have drawn on argumentation techniques (in particular influenced by the approach of assumption-based argumentation [17, 18] and preference-based argumentation frameworks [16]) to provide a general framework for taking evidence involving multiple outcome indicators and aggregate it in terms of arguments. In our framework, we instantiate abstract argument graphs with arguments generated by inference rules applied to the evidence, and attacks relationships obtained via the preference relations. For any application of our framework, a specific set of benefits preference relations and meta-arguments needs to be given. Given an evidence table, the algorithms for generating arguments and the argument graph are simple. The output is a superiority graph with an argument graph associated with each arc in the superiority graph. We summarize how we use the framework for evidence aggregation in Figure 3.

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. 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.

We have discussed how we can use our approach in practice by using a number of choices for the preference relation and meta-arguments, rather than fixing on one choice. In this way, we can manage the subjective criteria that arise in aggregating evidence for making recommendations. We have also considered the viability of algorithms for generating and evaluating argument graphs. For this, we have used a larger case study concerning lung cancer chemo-radiotherapy.

Direct implementation of our proposal is therefore a viable option. Another option is to generate the argument graphs using our algorithms, and then use an existing engine for calculating the extensions for an abstract argument graph [19, 20]. A third option is to use an implementation of a logic-based argumentation system such as that for assumption-based argumentation [21].

An important part of our proposal is the use of inductive arguments balanced by the use of preferences based on the benefits they identify and by the use of meta-argument that act as counterarguments based on the quality of the evidence used in the inductive arguments. The resulting argument graphs and superiority graphs provide a simple and yet effective way to analyze the evidence. Furthermore, we believe that they are straightforward to understand by healthcare professionals. We have already shown how clinicians use preferences in evaluating evidence [3], and it is straightforward to use our framework to represent these preferences. The advantage of allowing the user to define their own benefits preference relations and their own meta-arguments is that they can systematically use the evidence in the context of their working environment.

13.2 Discussion of related literature

The problem of conflicting information is a general issue in handling knowledge and it arises in virtually all real-world domains. To address this, computational models of argument which aim to reflect how human argumentation uses conflicting information to construct and analyze arguments, are being developed (for reviews see [12, 13]). These are potentially valuable because they help the decision-maker see what the key perspectives are on the conflicting information, as well as help them see what would be reasonable conclusions to draw.

Our framework is based on a specific kind of instantiation of preference-based argumentation frameworks (PAFs) as proposed by Amgoud and Cayrol [16]. Using PAFs with our framework for generating inductive arguments and meta-arguments, and for defining the preference relation (in terms of the relative benefits offered by the evidence), we have a simple and effective solution to aggregating evidence.

There are number of other developments of abstraction argumentation that are possible alternatives to PAFs for aggregating evidence. Value-based argumentation frameworks (VAFs) [22] allow for each argument to be assigned a value, and preference relation can then be assigned to those values. Different users may have different preference relations over the values. We could consider this mechanism for our framework, by letting the value denote the outcome indicators and magnitude of the outcome indicators. Another possibility is extended argumentation frameworks (EAFs) [23], and argumentation frameworks with recursive attacks (AFRAs) [24], which allow for arguments to attack attacks. A more general approach that subsumes PAFs, VAFs, and EAFs, is meta-level argumentation frameworks (MAFs) which allow for meta-arguments in general in abstract argumentation [25]. Again it is possible that we could capture our approach using EAFs, AFRAs, or MAFs. However, this would then create the question of how the arguments that attack attacks would be defined. Also, it would involve using more complex definitions than we use with

PAFs, and it is not clear that there would be any advantages for our approach by basing it on EAFs, AFRAs, or MAFs.

In this paper, we have been concerned with aggregation of evidence, and we have been able to show that we can do this for a general class of input that can be given as an evidence table. For this, each inductive argument denotes an aggregation of evidence. However, it appears that there are other kinds of knowledge that can be brought into the aggregation process, and that it would be interesting to further develop our approach to aggregate arguments directly, building on approaches such as [26, 27, 28, 29]. Health economics arguments are important in guidelines, and it would be a valuable development to harness those in our framework via argument accrual.

Whilst we have drawn on preferences over sets of results (i.e. sets of tuples where the first item is an outcome indicator and the second item is the normalized magnitude of that outcome indicator) in our framework, we do not believe that utility theory can be used in a straightforward way to address the problems of aggregating clinical evidence. A central idea in utility theory is that of a lottery $[p_1, o_1; ..., p_n, o_n]$ that we get if we choose a particular action, where p_i is the probability of getting outcome o_i. For aggregating evidence, a lottery would be required for each treatment option, where each outcome would be a particular combination of possible benefits from that treatment. Unfortunately, the evidence is unlikely to be sufficiently detailed to allow for generating this probability distribution. Furthermore, even if this distribution were guessed, it would decouple the evidence used to justify the claims made. For this application, clinicians want to see clearly the link between the evidence used and the recommendations made, and we believe that our approach provides that link clearly and rationally.

Little work exists that aims to address the problem in focus in this paper. Medical informatics and bioinformatics research has not addressed the reasoning aspects inherent in the analysis of evidence of a primary nature, especially from clinical trials. Previous interesting work ([30, 31, 32, 33] and others) exists that uses argumentation as a tool in medical decision support, but as such, assumes the existence of a hand-crafted knowledgebase. So as far as we know, there are no other proposals for automated reasoning with evidence, nor for aggregating that evidence via a computational model of argument.

We see our approach as being consistent with the GRADE approach [34]. 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 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 [35]. 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 benefits preference relations, and the meta-arguments, in the argumentation. This gives a number of substantial advantages: (1) The way that the evidence is being aggregated is made explicit, with the benefits 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; and (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 on our approach [36].

13.3 Discussion of future work

In future work, we aim to build a library of meta-arguments based on discussions with clinicians and epidemiologists. We will also investigate whether it would be advantageous to generalise the use of meta-arguments so that we can have meta-arguments attacking meta-arguments. For instance, we could have a "general" meta-argument attacking an inductive argument, and then a more specific meta-argument attacking the general meta-argument. In addition, we aim to develop generic benefits preference relations based on an ontology of outcome indicators and the magnitude of the outcome, that can be used via logical reasoning (e.g. survival for 5 years is of greater benefit than survival for 1 year, or risk of recurrence of breast cancer is reduced by 95% is better than risk of endometrial cancer, as a side-effect, increased by 10%). We also aim to develop theoretical tools for effectively and efficiently acquiring and representing benefits preference relations based on lattice theory and logical constraints.

Also in future work, we want to consider how findings from health economics modelling can be incorporated in the aggregation of evidence. It is appears that often when the evidence is inadequate for making a clear choice between some treatment options, health economics is used to break the tie. For instance, if two treatments appear equally good for a particular patient class, then a simple model based on cost can mean that one treatment is the recommended treatment. For at least these simple cases, it would appear straightforward to take the finding of health economics and treat it as another kind of evidence in the evidence table. By then defining the preference relation accordingly, a cheaper treatment would be superior if all other outcome indicators (and their magnitude) are more or less equal.

The proposal in this paper would be further enhanced by automated, or semi-automated construction of the evidence table. This may be possible by developing information extraction technology, based on hybrid techniques from natural language processing, that are tailored for extracting specific kinds of information in restricted domains, For developing information extraction of clinical trials, it may be possible to build on a set of open source tools and resources for clinical text mining that are available as part of the well-established GATE framework [37]. These resources include the CLEF corpus of annotated clinical documents [38], terminological resources such as the Unified Medical Language System (UMLS) [39], and machine learning methods (such as SVMs) that have been tailored to statistical named entity recognition (NER) and relationship extraction. Using GATE, Roberts and co-workers have recently developed and evaluated hybrid methods (combining terminological resources with statistical methods) for recognizing a set of entity types (medical condition, drug, intervention, etc.) relevant to our research [40], and statistical methods for the extraction of clinical relationships between these entities [41]. Also, there is the Trial Bank Project which is concerned with extracting detailed information about the patient class from published clinical trials [42]. Finally, machine learning technology has been used, with some success, to learn patterns for information extraction from abstracts of clinical trials [43].

Acknowledgments

The authors are very grateful to all the anonymous referees for their valuable suggestions for improving the presentation of this work.

References

[1] N. Gorogiannis, A. Hunter, V. Patkar, and M. Williams. Argumentation about treatment efficacy. In *Knowledge Representation for Healthcare (KR4HC)*, volume 5943 of *LNCS*, pages 169–179. Springer, Heidelberg, Germany, 2010.

[2] A. Hunter and M Williams. Qualitative evidence aggregation using argumentation. In P. Baroni, F. Ceruti, M. Giacomin, and G. Simari, editors, *Computational Models of Argument*, pages 287–298. IOS Press, Amsterdam, Netherlands, 2010.

[3] A. Hunter and M Williams. Using clinical preferences in argumentation about evidence from clinical trials. In T. Veinot, Ü. Çatalyürek, G. Luo, H. Andrade, and N. Smalheiser, editors, *Proceedings of the First ACM International Health Informatics Symposium*, pages 118–127. ACM Press, New York, NY, USA, 2010.

[4] N. Gorogiannis, A. Hunter, and M. Williams. An argument-based approach to reasoning with clinical knowledge. *International Journal of Approximate Reasoning*, 51(1):1 – 22, 2009.

[5] A. Hunter and M Williams. Argumentation for aggregating clinical evidence. In E. Grégoire, editor, *Proceedings of the International Conference on Tools with AI*, pages 361–368. IEEE Press, Los Alamitos, CA, USA, 2010.

[6] A. Hackshaw. *A Concise Guide to Clinical Trials.* WileyBlackwell, London, UK, 2009.

[7] F. Baader, D. Calvanese, D.L. McGuinness, D. Nardi, and P.F. Patel-Schneider, editors. *The description logic handbook: theory, implementation, and applications.* Cambridge University Press, New York, NY, USA, 2003.

[8] M. Williams and A. Hunter. Harnessing ontologies for argument-based decision-making in breast cancer. In *Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*, volume 2, pages 254–261. IEEE Computer Society Press, Los Alamitos, CA, USA, 2007.

[9] NICE. *Glaucoma: Clinical Guidelines CG85.* National Institute for Health and Clinical Excellence (www.nice.org.uk), London, UK, 2009. (accessed 1 April 2012).

[10] NICE. *Hyptertension: Clinical Guideline CG34.* National Institute for Health and Clinical Excellence (www.nice.org.uk), London, UK, 2006. (accessed 1 April 2012).

[11] NICE. *Hypertension in Pregnancy: Clinical Guidelines CG107.* National Institute for Health and Clinical Excellence (www.nice.org.uk), London, UK, 2010. (accessed 1 April 2012).

[12] T.J.M. Bench-Capon and P.E. Dunne. Argumentation in artificial intelligence. *Artificial Intelligence*, 171(10-15):619–641, 2007.

[13] Ph. Besnard and A. Hunter. *Elements of Argumentation.* MIT Press, Cambridge, MA, USA, 2008.

[14] I. Rahwan and G. Simari, editors. *Argumentation in Artifical Intelligence.* Springer, Hiedelberg, Germany, 2009.

[15] P. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming, and n-person games. *Artificial Intelligence*, 77:321–357, 1995.

[16] L. Amgoud and C. Cayrol. A reasoning model based on the production of acceptable arguments. *Annals of Mathematics and Artificial Intelligence*, 34:197–215, 2002.

[17] A. Bondarenko, P. Dung, R. Kowalski, and F. Toni. An abstract, argumentation-theoretic approach to default reasoning. *Artificial Intelligence*, 93:63–101, 1997.

[18] P. Dung, R. Kowalski, and F. Toni. Dialectical proof procedures for assumption-based admissible argumentation. *Artificial Intelligence*, 170:114–159, 2006.

[19] M South, G Vreeswijk, and J Fox. Dungine: A java dung reasoner. In Ph. Besnard, S. Doutre, and A. Hunter, editors, *Computational Models of Argument (COMMA'08)*, pages 360–368. IOS Press, Amsterdam, Netherlands, 2008.

[20] U. Egly, S. Gaggl, and S. Woltran. Aspartix: Implementing argumentation frameworks using answer-set programming. In M. Garcia de la Banda and E. Pontelli, editors, *Proceedings of the Twenty-Fourth International Conference on Logic Programming (ICLP'08),*, volume 5366 of *LNCS*, pages 734–738. Springer, Heidelberg, Germany, 2008.

[21] R. Craven, F. Toni, A. Hadad, and M. Williams. Efficient support for medical argumentation. In *Proceedings of the 13th International Conference on Principles of Knowledge Representation and Reasoning.* AAAI Press, Palo Alto, CA, USA, 2012. (in press).

[22] T. Bench-Capon. Persuasion in practical argument using value based argumentationframeworks. *Journal of Logic and Computation*, 13(3):429–448, 2003.

[23] S. Modgil. Reasoning about preferences in argumentation frameworks. *Artificial Intelligence*, 173(9-10):901–934, 2009.

[24] P. Baroni, F. Cerutti, M. Giacomin, and G. Guida. AFRA: Argumentation framework with recursive attacks. *International Journal of Approximate Reasoning*, 52(1):19–37, 2011.

[25] S. Modgil and T. Bench-Capon. Metalevel argumentation. *Journal of Logic and Computation*, 21(6):959–1003, 2011.

[26] H. Prakken. A study of accrual of arguments with applications to evidential reasoning. In *Proceedings of the Tenth International Conference on Artificial Intelligence and Law*, pages 85–94, 2005.

[27] B Verheij. Accrual of arguments in defeasible argumentation. In C. Witteveen and W van der Hoek, editors, *Proceedings of the Second Dutch German Workshop on Nonmonotonic Reasoning*, pages 217–224, 1995.

[28] T. Bench-Capon and H. Prakken. Justifying actions by accruing arguments. In *Proceeding of the 2006 conference on Computational Models of Argument*, pages 247–260. IOS Press, Amsterdam, Netherlands, 2006.

[29] M. Lucero, C. Chesñevar, and G. Simari. On the accrual of arguments in defeasible logic programming. In *Proceedings of the 21st international Joint conference on Artifical intelligence*, pages 804–809. AAAI Press, Palo Alto, CA, USA, 2009.

[30] J. Fox and S. Das. *Safe and Sound: Artificial Intelligence in Hazardous Applications.* MIT Press, Cambridge, MA, USA., 2000.

[31] V. Patkar, C. Hurt, R. Steele, S. Love, A. Purushotham, M. Williams, R. Thomson, and J. Fox. Evidence-based guidelines and decision support services: adiscussion and evaluation in triple assessment of suspectedbreast cancer. *British Journal of Cancer*, 95(11):1490–1496, 2006.

[32] P. Tolchinsky, U. Cortés, S. Modgil, and F. Caballeroand A. López-Navidad. Increasing human-organ transplant availability:argumentation-based agent deliberation. *IEEE Intelligent Systems*, 21(6):30–37, 2006.

[33] R Walton, C Gierl, P Yudkin, H Mistry, M Vessey, and J Fox. Evaluation of computer support for prescribing (CAPSULE). *British Medical Journal*, 315:791–795, 1997.

[34] G. Guyatt, A. Oxman, G. Vist, R. Kunz, Y. Falck-Ytter, P. Alonso-Coelle, and H. Schnemann. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *British Medical Journal*, 336:924–926, 2008.

[35] G. Guyatt, A. Oxman, G. Vist, R. Kunz, Y. Falck-Ytter, and H. Schnemann. GRADE: what is quality of evidence and why is it important to clinicians. *British Medical Journal*, 336:995–998, 2008.

[36] H. Schnemann, A. Oxman, J Brozek, P Glasziou, R Jaeschke, G Vist, J Williams, R Kunz, and J Craig. GRADE: grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *British Medical Journal*, 336:1106–1110, 2008.

[37] H Cunningham, D Maynard, K Bontcheva, and V Tablan. Gate: A framework and graphical development environment for robust NLP tools and applications. In E. Charniak and D. Lin, editors, *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*, 2002.

[38] A Roberts, R Gaizauskas, M Hepple, N Davis, G Demetriou, Y Guo, J Kola, I Roberts, A Setzer, A Tapuria, and B Wheeldin. The clef corpus: semantic annotation of clinical text. In *Proceedings of the Annual Symposium of American Medical Informatics Association (AMIA'07)*, pages 625–629, 2007.

[39] D Lindberg, B Humphreys, and A McCray. Unified medical language system. *Methods of Information in Medicine*, 32(4), 1993.

[40] A Roberts, R Gaizasukas, M Hepple, and Y Guo. Combining terminology resources and statistical methods for entity recognition: an evaluation. In *Proc. of the Sixth International Language Resources and Evaluation (LREC'08)*, 2008.

[41] A Roberts, R Gaizasukas, M Hepple, and Y Guo. Mining clinical relationships from patient narratives. *BMC Bioinformatics*, 9, 2008. Suppl 11,S3.

[42] I. Sim, D.K. Owens, P.W. Lavori, and G.D. Rennels. Electronic trial banks: A complementary method for reporting randomized trials. *Medical Decision Making*, 20(4):440–450, 2000.

[43] K. Hara and Y. Matsumoto. Extracting clinical trial design information from MEDLINE abstracts. *New Generation Computing*, 25(3):263–275, 2007.

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

Table 3: The evidence table for the case study. Each row is a meta-analysis from the NICE Glaucoma Guideline [9] (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 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).

	Left	Right	Outcome indicator	Value	Net	Sig	Type
e_{01}	BB	THZ	mortality	1.04	$<$	no	MA
e_{02}	ACE	CCB	mortality	1.04	$<$	no	MA
e_{03}	ACE	CCB	stroke	1.15	$<$	yes	MA
e_{04}	ACE	CCB	heart failure	0.85	$>$	yes	MA
e_{05}	ACE	CCB	diabetes	0.85	$>$	yes	MA
e_{06}	ARB	BB	mortality	0.89	$>$	no	MA
e_{07}	ARB	BB	myocardial infarction	1.05	$<$	no	MA
e_{08}	ARB	BB	stroke	0.95	$>$	no	MA
e_{09}	ARB	BB	heart failure	1.25	$<$	no	MA
e_{10}	ARB	BB	diabetes	0.75	$>$	yes	MA
e_{11}	ARB	CCB	mortality	1.02	$<$	no	MA
e_{12}	ARB	CCB	myocardial infarction	1.17	$<$	yes	MA
e_{13}	ARB	CCB	stroke	1.14	$<$	no	MA
e_{14}	ARB	CCB	heart failure	0.88	$>$	no	MA
e_{15}	ACE	THZ	mortality	1.00	$\sim$	no	MA
e_{16}	ACE	THZ	stroke	1.13	$>$	yes	MA
e_{17}	CCB	BB	mortality	0.94	$>$	no	MA
e_{18}	CCB	BB	myocardial infarction	0.93	$>$	no	MA
e_{19}	CCB	BB	stroke	0.77	$>$	yes	MA
e_{20}	CCB	BB	diabetes	0.71	$>$	yes	MA
e_{21}	CCB	THZ	mortality	0.97	$<$	no	MA
e_{22}	CCB	THZ	myocardial	1.02	$>$	no	MA
e_{23}	CCB	THZ	stroke	0.95	$<$	yes	MA
e_{24}	CCB	THZ	heart failure	1.38	$>$	yes	MA
e_{25}	CCB	THZ	diabetes	0.82	$<$	yes	MA

Table 4: The evidence table for the hypertension case study. Each row is a meta-analysis from the NICE Hypertension Guideline [10] (Appendix A pages 35-44) for the class of patients who have persistent high blood pressure. Each item is a meta-analysis (MA) generated by the guideline authors as presented in the guideline. The medications considered are beta-receptor blockers (BB), ACE inhibitors (ACE), thiazides (THZ), calcium channel blockers (CCB), and Angiotensin-II receptor antagonists (ARB).

	Left	Right	Outcome indicator	Value	Net	Sig	Type	Note
e_{01}	HEP	NT	pre-eclampsia	0.26	>	yes	MA	1,2
e_{02}	HEP	NT	fetal growth restriction	0.22	>	yes	MA	1,2
e_{03}	HEP	NT	gestational diabetes	0.48	>	no	MA	1,2
e_{04}	DI	NT	pre-eclampsia	0.68	>	no	MA	
e_{05}	PRO	NT	pre-eclampsia	0.21	>	no	MA	
e_{06}	NO	NT	pre-eclampsia	0.83	>	no	MA	
e_{07}	ASP	NT	pre-eclampsia	0.83	>	yes	MA	
e_{08}	ASP	NT	preterm	0.92	>	yes	MA	
e_{09}	ASP	NT	fetal & neonatal death	0.86	>	yes	MA	
e_{10}	ASP	NT	small gestational age	0.90	>	yes	MA	

Table 5: The evidence table for the pre-elampsia case study. Each row is a meta-analysis from the NICE Hypertension in Pregnancy Guideline [11] (Appendix G pages 1–23) for the class of patients who have pre-existing risk factors for the development of hypertensive disorders during pregnancy (such as diabetes, chronic hypertension, chronic kidney disease, etc). Each item is a meta-analysis / systematic review (MA) in the literature that the guideline authors considered. The medications considered are heparin (HEP), diuretics (DI), progesterone (PRO), nitric oxide (NO), aspirin (ASP), and placebo/no-treatment (NT). The final column is “Notes”. This is a commonly used category for extra information that might raise doubts in the item of evidence. Here, we consider two types: (1) The evidence from non-randomized and non-blind trials: and (2) The trials are for very narrow patient classes.