Qualitative Evidence Aggregation using Argumentation

Anthony HUNTER [a] and Matthew WILLIAMS [b]

[a] *UCL Department of Computer Science,*
Gower Street, London, WC1E 6BT, UK
[b] *Royal Free Hospital,*
Pond Street, London, NW3 2QG, UK

Abstract Evidence-based decision making is becoming increasingly important in many diverse domains, including healthcare, environmental management, and government. This has raised the need for tools to aggregate evidence from multiple sources. For instance, in healthcare, much valuable evidence is in the form of the results from clinical trials that compare the relative merits of treatments. For this, in a previous paper [5], we have proposed a general language for encoding, capturing and synthesizing knowledge from clinical trials and a framework that allows the construction and evaluation of arguments from such knowledge. Now, in this paper, we consider a specific version of the general framework for aggregating qualitative information about trials, and undertake an evaluation of this qualitative framework by comparing the results we obtain with those that are published in the biomedical literature. Whilst the results from our qualitative system are inferior, we show that they do offer a quick and useful aggregation of the evidence, and furthermore, we suggest that it could be coupled with information extraction technology to provide a valuable automated solution.

1. Introduction

The systematic use of evidence is already established in healthcare, and is being increasingly advocated in other domains, such as education and environmental management. However, the rapidly increasing amount of evidential knowledge on a subject means that it is difficult for a decision maker to locate, or even be aware of, new research that is relevant to their needs. Even if the decision maker locates the necessary evidence, it is difficult for them to effectively and efficiently assimilate and fully exploit it. In addition to the difficulty presented by the sheer volumes of information, the evidence is often conceptually complex, heterogeneous, incomplete and inconsistent. Not least, is the imperative to abstract away from the details of individual items of evidential knowledge, and to aggregate the evidence in a way that reduces the volume, complexity, inconsistency and incompleteness.

One important kind of evidence comes from superiority-testing clinical trials which compare the efficacy of two or more treatments in a particular class of patients. In order to have a global view of the relative merits of treatments for a particular condition, a potentially large number of publications needs to be reviewed. To address this, syntheses of the evidence on particular treatments are routinely produced using systematic search and statistical aggregation techniques (e.g., systematic reviews and meta-analyses). Often such syntheses involve groups of clinicians and statisticians. Such syntheses require significant time and effort, and they can quickly become out of date as new results are frequently being published.

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. Thus, it would be helpful to have a method for automatically analyzing and presenting the clinical trial results and the possible ways to aggregate those in an intuitive form, highlighting agreement and conflict present within the literature. Our proposal in [6,5] aims to suggest such a method. The first part of our proposal is a language that can be used to encode the published results in a semantically appropriate way, and methods for constructing a knowledge base from the encoded results. The second part of our framework allows the construction of arguments on the basis of evidence as well as their syntheses, published or generated on-the-fly. 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.

In this paper, we go beyond what we have done in [6,5] by presenting a specific set of inference rules and preference rules for qualitative evidence aggregation. The motivation for doing this is two-fold. First, with this very simple version of our general framework, we can investigate the quality with respect to published meta-analyses. Since meta-analyses are undertaken by clinicians, and medical statisticians, using standard techniques from statistics, for aggregating evidence, we see them as providing a “gold standard" for the aggregated evidence. So in this paper, we present some results comparing our approach with 15 meta-analyses obtained from three National Institute of Clinical Excellence (NICE) Guidelines (www.nice.org.uk).

The second reason we want to present the qualitative version of our framework is that it only requires a minimal amount of information from the published clinical trials. Indeed, the information we require can often be obtained from the abstract of the published paper covering the clinical trial. This then raises the possibility of using information extraction technology with the abstracts obtained from PubMed (www.pubmed.org) which catalogues all published clinical trials. The coupling of information extraction technology with argument-based aggregation could then be used for providing an automated and immediate abstract view on the relevant literature (highlighting where the evidence is in agreement and where there are key conflicts), and for generating rough drafts of meta-analysis, guidelines, and systematic review.

2. Representing clinical trials

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 assume a set of trials TRIALS where each trial is just an atomic name for which we associate information about the trial. We give an example in Table 1, and explain the attributes as follows. The first attribute is the **patient class** involved. In this example, it is

Table 1. Two results obtain from the NICE Glaucoma Guideline (Appendix pages 70-72) where PGA is an abbreviation for prostaglandin analogue and BB is an abbreviation for beta-blocker. The first row corresponds to a trial performed by Pfeifer *et al* in 2002 (Pfe02) and the second row corresponds to a trial performed by Felman *at al* in 2002 (Fel02).

Trial name	Patient class	Leftarm	Rightarm	Outcome indicator	Risk ratio	Statistically significant
Pfe02	glaucoma	PGA	BB	safe IOP	1.43	no
Fel02	glaucoma	PGA	BB	safe IOP	1.29	yes

patients who have glaucoma (a problem resulting from increased pressure in the eye causing damage to the optic nerve and retina). 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 [2]). See [16,6] for proposals for using a medical ontology in argumentation about clinical trials. However, in this paper, for simplicity we assume that the set of results in TRIALS concerns a particular, sensible patient class, and so we do not consider this aspect further here.

The next component of our representation concerns treatments. Again, medical ontologies cater for this task by providing categories and relationships on treatments, substances used, and other characteristics. We use the attributes **leftarm** and **rightarm** to signify the treatments compared in each trial in TRIALS.

A trial comparing two treatments will do so with respect to a particular outcome, which we call the **outcome indicator** e.g., in the case of the trials above, it is the proportion of patients for whom IOP (i.e. intra ocular pressure) is reduced to a safe level. As another example, for evaluating cancer treatments, it can be the proportion of patients who survive after 5 years.

A trial uses a statistical method to compare the two treatments. There is a range of methods, each appropriate to specific trial designs and outcomes. Here, **risk ratio** is used which in general means the measure of the outcome indicator obtained from the leftarm divided by measure of the outcome indicator obtained from the rightarm. For these trials, specifically it means the proportion of patients in the leftarm (i.e. those treated with prostaglandin) who during the trial period had the IOP reduced to a safe level divided by the proportion of patients in the rightarm (i.e. those treated with betablocker) who during the trial period had the IOP reduced to a safe level. So for both Pfe02 and Fel02, the risk ratio is greater than 1, which means that in both trials, prostaglandin is associated with more patients having a safe IOP than betablocker.

The final attribute is **statistical significance** for which if the entry is “yes" means that it is unlikely that the risk ratio result could have been obtained by chance (using a conventional cut-off such as 0.05), whereas if it is “no" then it means that it is quite likely to have been obtained by chance.

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), and it is straightforward to accommodate these extra attributes in our framework. For a general introduction to the nature of clinical trials, and a discussion of a wider range of attributes, see [7].

3. General framework

In this section, we review the general framework presented in [5] for constructing and comparing arguments based on the kind of information presented in the previous section.

For a superiority clinical trial comparing treatments τ_1 and τ_2 with respect to the outcome indicator μ, there are three possible interpretations of its results: (1) $\tau_1 >_\mu \tau_2$, meaning that we believe that the result supports the claim that treatment τ_1 is superior to τ_2 with respect to μ; (2) $\tau_1 <_\mu \tau_2$, meaning that we believe that the result supports the claim that treatment τ_1 is inferior to τ_2 with respect to μ; And (3) $\tau_1 \sim_\mu \tau_2$, meaning that we believe the result as supporting the claim that neither τ_1 nor τ_2 is superior to each other with respect to μ; Any formula of the form $\tau_1 >_\mu \tau_2$, $\tau_1 \sim_\mu \tau_2$ and $\tau_1 <_\mu \tau_2$ we will call a **claim**, denoted by ϵ, possibly subscripted. Note, we treat $\tau_1 > \tau_2$ as equivalent to $\tau_2 < \tau_1$ and $\tau_1 \sim \tau_2$ as equivalent to $\tau_2 \sim \tau_1$.

Given a set of results TRIALS one can informally think of an argument comprising of a set of evidence (i.e. a subset of TRIALS), an inferential rule and a conclusion or claim. For example, a plausible interpretation of Fel02 is that since the value for risk ratio is greater than 1, the first treatment is better than the second with respect to obtaining a safe IOP, i.e., that PGA $>_{\text{safeIOP}}$ BB. We define this process by an inference rule.

Definition 1. *An **inference rule**, λ, is a rule with conditions (employing set-theoretic expressions and equations utilizing attributes over the reals) on a set of results $\Phi \subseteq$ TRIALS and a claim ϵ.*

Example 1. *For* TRIALS, *let τ_1 be the leftarm, let τ_2 be the rightarm, let μ be the outcome indicator, and let $\gamma \in$* TRIALS.

$$(\lambda_s) \text{ For } \Phi = \{\gamma\}, \text{ if } \gamma \text{ is statistically significant and the risk ratio is greater than 1, then } \tau_1 >_\mu \tau_2.$$

Example 2. *For* TRIALS, *let τ_1 be the leftarm, let τ_2 be the rightarm, let μ let the outcome indicator, and let $\gamma \in$* TRIALS.

$$(\lambda_n) \text{ For } \Phi = \{\gamma\}, \text{ if } \gamma \text{ is not statistically significant, then } \tau_1 \sim_\mu \tau_2$$

Definition 2. *An **argument** is a triple $\langle \Phi, \lambda, \epsilon \rangle$ where $\Phi \subseteq$ TRIALS is a set of results, λ is an inference rule, Φ satisfies the conditions of λ and ϵ is the claim of λ applied to Φ.*

Example 3. *Using the data in the previous section concerning Fel02 and Pfe02, we obtain the following arguments.*

$$\langle \{\text{Fel02}\}, \lambda_s, \text{PGA } >_{\text{safeIOP}} \text{BB} \rangle \quad \langle \{\text{Pfe02}\}, \lambda_n, \text{PGA } \sim_{\text{safeIOP}} \text{BB} \rangle$$

In the above example, we see that the two arguments are in conflict. We capture this kind of conflict with the following definition. Note that this definition is symmetric, i.e., if A conflicts with B then B conflicts with A.

Definition 3. *If $A = \langle \Phi_A, \lambda_A, \epsilon_A \rangle$ and $B = \langle \Phi_B, \lambda_B, \epsilon_B \rangle$ are two arguments then we say that A **conflicts** with B whenever:*

1. *$\epsilon_A = \tau_1 >_\mu \tau_2$, and either $\epsilon_B = \tau_1 \sim_\mu \tau_2$ or $\epsilon_B = \tau_1 <_\mu \tau_2$.*
2. *$\epsilon_A = \tau_1 \sim_\mu \tau_2$, and either $\epsilon_B = \tau_1 >_\mu \tau_2$ or $\epsilon_B = \tau_1 <_\mu \tau_2$.*

3. $\epsilon_A = \tau_1 <_\mu \tau_2$, *and either* $\epsilon_B = \tau_1 >_\mu \tau_2$ *or* $\epsilon_B = \tau_1 \sim_\mu \tau_2$.

We organize the arguments into a graph. To do this, we first consider the conflict relation given above. It is easy to see that the graph induced is tripartite, and its independent sets are given by those arguments with claim $\tau_1 >_\mu \tau_2$, those arguments with claim $\tau_1 \sim_\mu \tau_2$, and those arguments with claim $\tau_1 <_\mu \tau_2$. In our example, this graph is as follows.

$$\langle\{\text{Fel02}\}, \lambda_s, \text{PGA } >_{\text{safeIOP}} \text{BB}\rangle \rightleftarrows \langle\{\text{Pfe02}\}, \lambda_n, \text{PGA } \sim_{\text{safeIOP}} \text{BB}\rangle$$

Since the argument graph is by definition symmetric (if we use the conflict relation), it would be beneficial to allow breaking the symmetry with user-defined preferences. We do this by defining preference rules.

Definition 4. *A* ***preference rule*** *is a set of conditions on an ordered pair of conflicting arguments A, B. When the conditions are satisfied, A is said to be preferred to B otherwise, we say that A is not preferred to B.*

Example 4. *For $A = \langle\{\gamma_a\}, \lambda_a, \epsilon_A\rangle$ and $B = \langle\{\gamma_b\}, \lambda_b, \epsilon_B\rangle$ such that A conflicts with B, A is preferred to B iff γ_a is statistically significant and γ_b is not statistically significant.*

Preference rules are not required to be infallible in any sense. Indeed the above example embodies one of the aspects of *publication bias*, where by preferring significant results to non-significant ones, one may miss evidence that supports the claim that the significant results are a chance occurrence.

We use the preference rules chosen by the user in breaking the symmetry present in the conflict relation, as developed by Amgoud and Cayrol [1], and capture the attack relation as follows.

Definition 5. *For any pair of arguments A and B, A* **attacks** *B iff A conflicts with B and A is preferred to B and it is not the case that B is preferred to A.*

The motivation here is that if A and B conflict with each other and A is preferred to B then B's conflict with A is cancelled. However, this wording leads to problems when A is preferred to B according to a preference rule and B is preferred to A according to a preference rule. In this case, cancelling both attacks will give the misleading impression that A and B are consistent together. For this reason we give the above, more complicated definition, which only cancels an attack if exactly one argument is preferred to the other.

Now we combine these components by defining an argument graph based on a set of trial results, a set of inference rules, and a set of preference rules as follows.

Definition 6. *Given a pair of treatments τ_1, τ_2 and an outcome indicator μ, and a set* TRIALS *concerning these treatments and outcome indicator, an* **argument graph** *is a graph where the set of nodes is the set of arguments formed using a set of inference rules as given by Definition 1 and the set of arcs is the attacks relation given by Definition 5.*

We can directly use the dialectical semantics given by Dung [4] to decide extensions of argument graphs. We regard a preferred set of arguments as an interpretation of a TRIALS (i.e. an aggregation of the evidence in TRIALS. So if X is an extension of the argument graph, and $A \in X$, and ϵ is the claim of A, then ϵ is a possible aggregation of the evidence.

4. Qualitative framework

Now we present a specific version of the framework including inference rules and preference rules. We start with a set of trials $\text{TRIALS} = \{t_1, .., t_n\}$ each of which uses the same outcome indicator and compares the same pair of treatments τ_1 and τ_2. We partition TRIALS 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, those in EQUITABLE are the trials for which τ_2 was shown to equitable with τ_1, and those in INFERIOR are the trials for which τ_2 was shown to be superior to τ_1. We also partition TRIALS into two sets SIGNIFICANT and NONSIGNIFICANT. Those in SIGNIFICANT are the trials for which the result is significant, and those in NONSIGNIFICANT are the trials for which the result is not significant. In this paper, we focus on qualitative aggregation based solely on the distribution of trials in SUPERIOR, EQUITABLE, INFERIOR, SIGNIFICANT and NONSIGNIFICANT.

The inference rules we use for the qualitative framework are given in Table 2. From these inference rules, we get four types of argument as follows.

- $\langle \text{SUPERIOR}, R_x, \tau_1 > \tau_2 \rangle$ where $R_x \in \{R_1, ..., R_{12}\}$
- $\langle \text{INFERIOR}, R_x, \tau_1 < \tau_2 \rangle$ where $R_x \in \{R_1, ..., R_{12}\}$
- $\langle \text{EQUITABLE}, R_x, \tau_1 \sim \tau_2 \rangle$ where $R_x \in \{R_{13}, .., R_{15}\}$
- $\langle \text{NONSIGNIFICANT}, R_{16}, \tau_1 \sim \tau_2 \rangle$

Note, the items of evidence in INFERIOR state that τ_1 is inferior to τ_2 which is equivalent to stating that τ_2 is superior to τ_1. Furthermore, as we specified earlier, $\tau_1 < \tau_2$ is equivalent to $\tau_2 > \tau_1$. Hence, by this correspondence, we may be able to apply the rules in $R_x \in \{R_1, ..., R_{12}\}$ to generate an argument $\langle \text{INFERIOR}, R_x, \tau_1 < \tau_2 \rangle$ where $R_x \in \{R_1, ..., R_{12}\}$.

Given a set TRIALS, we let $\mathsf{Args}(\text{TRIALS})$ denote the set of arguments that can be generated by this set of rules. Also, for an argument A, let $\mathsf{Rule}(A)$ be the inference rule used in the argument, and let $\mathsf{Claim}(A)$ be the claim of the argument (i.e. for $A = \langle \Phi, \lambda, \epsilon \rangle$, $\mathsf{Rule}(A) = \lambda$ and $\mathsf{Claim}(A) = \epsilon$).

Example 5. *For prostaglandin v beta-blocker (see Table 4), for obtaining a safe IOP, we have* $|\text{TRIALS}|$ *= 12,* $|\text{SUPERIOR}|$ *= 12, and* $|\text{SIGNIFICANT}|$ *= 7. Hence, we get the argument* $\langle \text{SUPERIOR}, R_2, \text{PGA} >_{\text{safeIOP}} \text{BB} \rangle$.

We motivate the inference rules as follows. First, $R_1, .., R_4$ are for when all the trials show superiority (of the leftarm over the rightarm), $R_5, .., R_8$ are for when the majority of the trials show superiority, and $R_9, .., R_{12}$ are for when a minority of the trials show superiority. Then each of these three groups is broken down according to the proportion of the trials that show superiority are also significant, i.e., $\frac{|\text{SUPERIOR} \cap \text{SIGNIFICANT}|}{|\text{SUPERIOR}|}$. So for instance, for R_1, all trials are significant, for R_2, it is the majority that are significant, for R_3, it is a minority that are significant, and for R_4, none are significant. Then, $R_{13}, .., R_{15}$ are for when some trials show equality (of the left and right arms). So R_{13} is when a minority show equality, R_{14} is when a majority show equality, and R_{15} is when all show equality. Note, $R_{13}, .., R_{15}$ are not broken down by significance since technically, when a trial shows equality it is a failure to show a difference, whereas significance is for showing whether a difference occurred by chance. Hence, for equality, significance is not meaningful. Finally, R_{16} is for when the proportion of trials that are nonsignificant is greater than or equal to 1/2.

Table 2. Inference rules for qualitative framework. Given TRIALS, let ρ_1 = |SUPERIOR/TRIALS|, ρ_2 = |SIGNIFICANT∩SUPERIOR/SUPERIOR|, ρ_3 = |SIGNIFICANT/TRIALS|, and ρ_4 = |EQUITABLE/TRIALS|.

Rule	Condition	Explanation	Claim
R_1	(1) $\rho_1 = 1$ (2) $\rho_2 = 1$	all trials show superiority of which all are significant	$\tau_1 > \tau_2$
R_2	(1) $\rho_1 = 1$ (2) $0.5 < \rho_2 < 1$	all trials show superiority of which a majority are significant	$\tau_1 > \tau_2$
R_3	(1) $\rho_1 = 1$ (2) $0 < \rho_2 \leq 0.5$	all trials show superiority of which a minority are significant	$\tau_1 > \tau_2$
R_4	(1) $\rho_1 = 1$ (2) $\rho_2 = 0$	all trials show superiority of which none are significant	$\tau_1 > \tau_2$
R_5	(1) $0.5 < \rho_1 < 1$ (2) $\rho_2 = 1$	a majority of trials show superiority of which all are significant	$\tau_1 > \tau_2$
R_6	(1) $0.5 < \rho_1 < 1$ (2) $0.5 < \rho_2 < 1$	a majority of trials show superiority of which a majority are significant	$\tau_1 > \tau_2$
R_7	(1) $0.5 < \rho_1 < 1$ (2) $0 < \rho_2 \leq 0.5$	a majority of trials show superiority of which a minority are significant	$\tau_1 > \tau_2$
R_8	(1) $0.5 < \rho_1 < 1$ (2) $\rho_2 = 0$	a majority of trials show superiority of which none are significant	$\tau_1 > \tau_2$
R_9	(1) $0 < \rho_1 \leq 0.5$ (2) $\rho_2 = 1$	a minority of trials show superiority of which all are significant	$\tau_1 > \tau_2$
R_{10}	(1) $0 < \rho_1 \leq 0.5$ (2) $0.5 < \rho_2 < 1$	a minority of trials show superiority of which the majority are significant	$\tau_1 > \tau_2$
R_{11}	(1) $0 < \rho_1 \leq 0.5$ (2) $0 < \rho_2 \leq 0.5$	a minority of trials show superiority of which a minority are significant	$\tau_1 > \tau_2$
R_{12}	(1) $0 < \rho_1 \leq 0.5$ (2) $\rho_2 = 0$	a minority of trials show superiority of which none are significant	$\tau_1 > \tau_2$
R_{13}	$0 < \rho_4 \leq 0.5$	a minority of trials show equality of τ_1 and τ_2	$\tau_1 \sim \tau_2$
R_{14}	$0.5 < \rho_4 < 1$	a majority of trials show equality of τ_1 and τ_2	$\tau_1 \sim \tau_2$
R_{15}	$\rho_4 = 1$	all trials show equality of τ_1 and τ_2	$\tau_1 \sim \tau_2$
R_{16}	$0.5 \leq \rho_3 \leq 1$	half or more trials are statistically nonsignificant	$\tau_1 \sim \tau_2$

Given a set TRIALS comparing τ_1 and τ_2, the inference rules R_1 to R_{16} impose constraints on what combinations of arguments are possible together in $\mathsf{Args}(\text{TRIALS})$.

Proposition 1. *If there is an argument $A_i \in \mathsf{Args}(\text{TRIALS})$ where $\mathsf{Claim}(A_i) = \tau_1 > \tau_2$, then there is at most one argument $A_j \in \mathsf{Args}(\text{TRIALS})$ where $\mathsf{Claim}(A_j) = \tau_2 > \tau_1$, and there is at most two arguments $A_k \in \mathsf{Args}(\text{TRIALS})$ where $\mathsf{Claim}(A_k) = \tau_1 \sim \tau_2$,*

So the above says that there is at most one argument showing superiority, at most two showing equivalence, and at most one showing inferiority, and the following says that there is always argument with at least one of these claims.

Proposition 2. *If TRIALS $\neq \emptyset$, then there is an argument $A_i \in \mathsf{Args}(\text{TRIALS})$ where $\mathsf{Claim}(A_i) = \tau_1 > \tau_2$ or $\mathsf{Claim}(A_i) = \tau_1 \sim \tau_2$ or $\mathsf{Claim}(A_i) = \tau_2 > \tau_1$.*

Being able to use rules R_5 to R_{12} means that conflicting arguments can be generated from R_5 to R_{15} as captured by the following proposition.

Table 3. For arguments A_i and A_j, A_i is preferred to A_j iff one of P_2 to P_{11} holds for $\mathsf{Rule}(A_i)$ and $\mathsf{Rule}(A_j)$.

Preference rule	$\mathsf{Rule}(A_i)$	$\mathsf{Rule}(A_j)$
P_2	R_2	$\{R_{16}\}$
P_3	R_3	$\{R_{16}\}$
P_4	R_4	$\{R_{16}\}$
P_5	R_5	$\{R_{11}, R_{12}, R_{13}, R_{16}\}$
P_6	R_6	$\{R_{11}, R_{12}, R_{13}, R_{16}\}$
P_7	R_7	$\{R_{12}, R_{13}, R_{16}\}$
P_8	R_8	$\{R_{12}, R_{13}, R_{16}\}$
P_9	R_9	$\{R_8, R_{11}, R_{12}, R_{13}, R_{16}\}$
P_{10}	R_{10}	$\{R_8, R_{11}, R_{12}, R_{13}, R_{16}\}$
P_{11}	R_{11}	$\{R_8, R_{13}\}$

Proposition 3. *If there is an argument* $A_i \in \mathsf{Args}(\textsc{trials})$ *s.t.* $\mathsf{Rule}(A_i) \in \{R_5, ..., R_{12}\}$ *and* $\mathsf{Claim}(A_i) = \tau_1 > \tau_2$*, then there is an argument* $A_j \in \mathsf{Args}(\textsc{trials})$ *where* $\mathsf{Rule}(A_j) \in \{R_5, ..., R_{15}\}$ *and either* $\mathsf{Claim}(A_j) = \tau_1 < \tau_2$ *or* $\mathsf{Claim}(A_j) = \tau_1 \sim \tau_2$.

However, being able to use rules R_1 to R_4 means that no conflicting arguments can be generated by using rules R_5 to R_{15}.

Proposition 4. *If there is an argument* $A_i \in \mathsf{Args}(\textsc{trials})$ *where* $\mathsf{Rule}(A_i) \in \{R_1, ..., R_4\}$ *and* $\mathsf{Claim}(A_i) = \tau_1 > \tau_2$*, then there is no argument* $A_j \in \mathsf{Args}(\textsc{trials})$ *where* $\mathsf{Rule}(A_j) \in \{R_5, ..., R_{15}\}$ *and either* $\mathsf{Claim}(A_j) = \tau_1 < \tau_2$ *or* $\mathsf{Claim}(A_j) = \tau_1 \sim \tau_2$.

Being able to use rules R_{13} or R_{14} also means that conflicting arguments can be generated as captured by the following propositions.

Proposition 5. *If there is an argument* $A_i \in \mathsf{Args}(\textsc{trials})$ *s.t.* $\mathsf{Claim}(A_i) = \tau_1 \sim \tau_2$ *and either* $\mathsf{Rule}(A_i) = R_{13}$ *or* $\mathsf{Rule}(A_i) = R_{14}$*, then there is an argument* $A_j \in \mathsf{Args}(\textsc{trials})$ *where* $\mathsf{Rule}(A_j) \in \{R_5, ..., R_{12}\}$ *and either* $\mathsf{Claim}(A_j) = \tau_1 > \tau_2$ *or* $\mathsf{Claim}(A_j) = \tau_1 < \tau_2$.

However, being able to use rule R_{15} means that no conflicting arguments can be generated by using rules R_1 to R_{15}.

Proposition 6. *If there is an argument* $A_i \in \mathsf{Args}(\textsc{trials})$ *where* $\mathsf{Rule}(A_i) = R_{15}$ *then there is no argument* $A_j \in \mathsf{Args}(\textsc{trials})$ *where* $\mathsf{Rule}(A_j) \in \{R_1, ..., R_{14}\}$.

The preference rules are given in Table 3. Note, we do not consider R_1 because if it applies, no other rule could apply, and for $R_2, .., R_4$, the only other rule that can fire is R_{16}. Also, we do not consider $R_{12}, .., R_{16}$ since any argument based on them is not preferred to any other argument.

Example 6. *For prostaglandin v beta-blocker (m_3 in Table 4), for lower risk of respiratory problems as a side-effect, there are 2 trials, of which 1 shows superiority significantly and 1 shows superiority non-significantly (and so all the trials say that prostaglandin is superior to beta-blocker). By preference rule P2, the attack from the right* *argument to the left argument is suppressed. Therefore, we have the following argument graph, and we obtain the left argument in the resulting grounded extension.*

$$\langle \textsc{superior}, R_3, \text{PGA} >_{\text{respiratory}} \text{BB}\rangle \rightarrow \langle \textsc{equitable}, R_{16}, \text{PGA} \sim_{\text{respiratory}} \text{BB}\rangle$$

Example 7. *For prostaglandin v beta-blocker (m_4 in Table 4), for lower risk of cardiological problems as a side-effect, there are 5 trials, of which 1 shows superiority significantly, 2 show superiority non-significantly and 2 show inferiority non-significantly. So by preference rule P7, the attack from the right argument to the left argument is suppressed, and the attack from the lower argument to the left argument is also suppressed. Therefore, we have the following argument graph, and we obtain the left argument in the resulting grounded extension.*

$$\begin{tikzcd}
\langle \textsc{superior}, R_7, \text{PGA} >_{\text{cardio}} \text{BB}\rangle \arrow[rr] \arrow[dr] & & \langle \textsc{inferior}, R_{12}, \text{PGA} <_{\text{cardio}} \text{BB}\rangle \arrow[dl, shift right] \\
& \langle \textsc{nonsignificant}, R_{16}, \text{PGA} \sim_{\text{cardio}} \text{BB}\rangle \arrow[ur, shift right] &
\end{tikzcd}$$

With the qualitative framework, we have a simple set of inference rules and preference rules, that given a set of trial results TRIALS produces a small set of arguments and attack relationships. It allows for highlighting key conflicts in possible aggregations of the evidence, and as we show in the next section, it appears to perform well with real data.

5. Case study

In order to evaluate the qualitative framework, we have taken 14 meta-analyses from 3 NICE Guidelines (www.nice.org.uk), and we compare the results they obtained with the results that our qualitative evidence aggregation produced. We give a summary of this comparison in Tables 4, 5, and 6.

In these tables, each row is a based on a meta-analysis in the NICE guideline where $n_1 = |\textsc{superior} \cap \textsc{significant}|$, $n_2 = |\textsc{superior} \cap \textsc{nonsignificant}|$, $n_3 = |\textsc{equitable}|$, $n_4 = |\textsc{inferior} \cap \textsc{nonsignificant}|$, and $n_5 = |\textsc{inferior} \cap \textsc{significant}|$. The column "Their result" is the weighted average presented in the meta-analysis in the guideline where sup (respectively eq and inf) denotes superior (respectively equal and inferior) and sig (respectively non-sig) denotes significant (respectively non-significant). The column "Rule used" gives the rules that appear in the arguments we generate from the data in $n_1, .., n_5$, and "Our result" is the form of the claims of the arguments in the union of the preferred extensions. So for example, in Table 4, the first row labelled m_1, concerns a meta-analysis based on 12 clinical trials, of which 7 were statistically significant, and their weighted average result showed the leftarm was significantly superior to the rightarm, and our result showed leftarm was superior to the rightarm, and this was based on an argument involving inference rule R_2.

- From the NICE Glaucoma Guideline (CG85), we have investigated 6 meta-analyses, and give the data and results in Table 4. In each case where their result is superior significantly (respectively inferior significantly), we obtain $\tau_1 > \tau_2$ (respectively $\tau_1 < \tau_2$). For the cases where the their result is superior non-significantly, we obtain $\tau_1 > \tau_2$. For the case where their result is inferior non-

Table 4. Comparison of qualitative argument-based evidence aggregation with meta-analyses from NICE Glaucoma Guideline (CG85, Appendix, pp 218-221). Each row is a meta-analysis where the left arm is a prostaglandin analogue and the right arm is a beta-blocker. The treatment is intended to lower intraocular pressure (IOP). The outcome indicator for each meta-analysis is as follows: m_1 Decrease of IOP; m_2 Acceptable (safe) IOP; m_3 Respiratory problems; m_4 Cardiological problems; m_5 Allergic problems; m_6 Hyperaemia problems.

	n_1	n_2	n_3	n_4	n_5	Their result	Rules used	Our result
m_1	7	5	0	0	0	sup sig	R_2	$\tau_1 > \tau_2$
m_2	6	1	0	0	0	sup sig	R_2	$\tau_1 > \tau_2$
m_3	1	1	0	0	0	sup non-sig	R_2, R_{16}	$\tau_1 > \tau_2$
m_4	1	2	0	2	0	sup non-sig	R_7, R_{12}, R_{16}	$\tau_1 > \tau_2$
m_5	0	1	0	1	0	inf non-sig	R_8, R_{16}	$\tau_1 > \tau_2, \tau_1 \sim \tau_2, \tau_1 < \tau_2$
m_6	0	0	0	4	6	inf sig	R_2	$\tau_1 < \tau_2$

significantly, we obtain the vaguer result of a disjunction of $\{\tau_1 > \tau_2, \tau_1 \sim \tau_2, \tau_1 < \tau_2\}$ instead of $\tau_1 < \tau_2$. So overall, in 5 out of 6 cases, we get the same superiority/inferiority relation as their result, and in 1 out of 6 cases, we get a vaguer result (i.e. disjunction of $\{\tau_1 > \tau_2, \tau_1 \sim \tau_2, \tau_1 < \tau_2\}$).

- From the NICE Hypertension Guideline (CG34), we have investigated 5 meta-analyses, and give the data and results in Table 5. In each case where their result is superior significantly (respectively inferior significantly), we obtain $\tau_1 > \tau_2$ (respectively $\tau_1 < \tau_2$). There are 2 cases where their result is inferior non-significant, for which we obtain $\tau_1 > \tau_2$ in one case and $\tau_1 < \tau_2$ in the other case. Also there is 1 case where their result is superior non-significant, for which we obtain $\tau_1 < \tau_2$. So overall, in 3 out of 5 cases, we get the same as their result, and for 2 out of 5 cases, we get the opposite (i.e. either $\tau_1 > \tau_2$ instead of $\tau_1 < \tau_2$ or $\tau_1 < \tau_2$ instead of $\tau_1 > \tau_2$).
- From the NICE Type 2 Diabetes Guideline (CG66), we have investigated 3 meta-analyses, and give the data and results in Table 6. In the case where their result is superior significantly, we obtain $\tau_1 > \tau_2$. In the case where their result is superior non-significantly, we obtain $\tau_1 > \tau_2$. And in the case where their result is inferior non-significantly, we obtain the disjunction of $\{\tau_1 > \tau_2, \tau_1 \sim \tau_2, \tau_1 < \tau_2\}$.

From this consideration of 14 meta-analyses, it would appear that the qualitative evidence aggregation performs well. In 10 out of 14 meta-analyses, we get the same result (i.e. superiority or inferiority), in 2 out of 14 meta-analyses, we get a vaguer result (i.e. a disjunction), and in 2 out of 14 meta-analyses, we get the opposite result (i.e. the incorrect result) to the meta-analysis.

6. Discussion

We have presented a qualitative framework for argumentation on treatment efficacy. Using these components along with standard argumentation tools, users can describe their preferences and analyze the available evidence in terms of agreement and conflict.

The advantage of qualitative evidence aggregation is that it allows for abstraction from details of a meta-analysis, and it allows for modularity of analysis (thereby facilitating the aggregation according to multiple outcome indicators). Obviously such qualitative evidence aggregation is not able to replace statistical evidence aggregation. Ra-

Table 5. Comparison of qualitative argument-based evidence aggregation with meta-analysis data and results taken from NICE Hypertension Guideline (CG34, pp 36-43). Each row is a meta-analysis where the left arm is a calcium channel blocker and the right arm is a thiazide. The treatment is intended to lower blood pressure. The outcome indicator for each meta-analysis is as follows: m_7 Mortality; m_8 Myocardial infarction; m_9 Stroke; m_{10} Heart failure; and m_{11} Diabetes.

	n_1	n_2	n_3	n_4	n_5	Their result	Rules used	Our result
m_7	0	2	0	3	0	sup non-sig	R_8, R_{12}, R_{16}	$\tau_1 < \tau_2$
m_8	0	1	0	4	0	inf non-sig	R_8, R_{12}, R_{16}	$\tau_1 < \tau_2$
m_9	0	3	0	2	0	inf non-sig	R_8, R_{12}, R_{16}	$\tau_1 > \tau_2$
m_{10}	0	1	0	2	2	inf sig	R_7, R_{12}, R_{16}	$\tau_1 < \tau_2$
m_{11}	2	1	0	0	0	sup sig	R_2	$\tau_1 > \tau_2$

Table 6. Comparison of qualitative argument-based evidence aggregation with meta-analysis data and results taken from NICE Type 2 Diabetes Guideline (CG66, Appendix, p 18). Each row is a meta-analysis where the outcome indicator is the lowering of HbA1c (a protein involved in diabetes). For m_{13}, the leftarm is biphasic insulin and the rightarm is human insulin; for m_{14}, the leftarm is glargin insulin and the rightarm is human insulin; and for m_{15}, the leftarm is biphasic insulin and the rightarm is glargin insulin.

	n_1	n_2	n_3	n_4	n_5	Their result	Rules used	Our result
m_{12}	0	4	1	1	0	sup non-sig	$R_8, R_{12}, R_{13}, R_{16}$	$\tau_1 > \tau_2$
m_{13}	0	1	0	1	0	inf non-sig	R_8, R_{16}	$\tau_1 > \tau_2$,$\tau_1 \sim \tau_2$,$\tau_1 < \tau_2$
m_{14}	3	0	0	0	0	sup sig	R_1	$\tau_1 > \tau_2$

ther it is meant to complement it by addressing some of the shortcomings of statistical evidence aggregation including statistics suppresses conflict by using averages (whereas argumentation highlights conflict), statistics hides issues such as problems with particular sources of evidence (whereas in argumentation this can be made explicit by use of appropriate preference rules and/or further types of inference rule), and statistics is based on assumptions that either might be hidden or debatable.

Little work exists that aims to address the problem in focus here. Medical informatics and bioinformatics research does not address the reasoning aspects inherent in the analysis of evidence of primary nature, especially from clinical trials. Previous interesting work ([9,14,15] and others) exists that uses argumentation as a tool in medical decision support, but as such, assumes the existence of a hand-crafted knowledgebase.

We believe that the work presented here is a step towards an automated system for aggregating qualitative evidence. It is straightforward to implement the inference rules for generating arguments and the preference rules for generating the attack relation. Furthermore, given that there is only a small number of arguments generated per set TRIALS, a naive algorithm that considers each subset of arguments for calculating preferred extensions is viable.

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 [3]. These resources include the CLEF corpus of annotated clinical documents [12], terminological resources such as the Unified Medical Language System (UMLS) [8], 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 [10], and statistical methods for the extraction of clinical relationships between these entities [11]. Also, there is the Trial Bank Project which is concerned with extracting detailed information about the patient class from published clinical trials [13].

References

[1] L Amgoud and C Cayrol. Inferring from inconsistency in preference-based argumentation frameworks. *Journal of Automated Reasoning*, 29:125–169, 2002.

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

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

[4] 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.

[5] N Gorogiannis, A Hunter, V Patkar, and M Williams. Argumentation about treatment efficacy. In *Knowledge Representation for Healthcare (KR4HC)*, LNCS. Springer, 2010.

[6] 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.

[7] A Hackshaw. *A Concise Guide to Clinical Trials*. WileyBlackwell, 2009.

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

[9] 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.

[10] 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.

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

[12] 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.

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

[14] 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.

[15] 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.

[16] 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, 2007.