Qualitative Evidence Aggregation using Argumentation

Anthony HUNTER\textsuperscript{a} and Matthew WILLIAMS\textsuperscript{b}
\textsuperscript{a} UCL Department of Computer Science, Gower Street, London, WC1E 6BT, UK
\textsuperscript{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). Of-
ten 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 stan-
dard 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

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 publi-
shed paper covering the clinical trial. This then raises the possibility of using informa-
tion 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 auto-
mated and immediate abstract view on the relevant literature (highlighting where the evi-
dence 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 \( \text{TRIALS} \) 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 \textbf{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 et al in 2002 (Fel02).

<table>
<thead>
<tr>
<th>Trial name</th>
<th>Patient class</th>
<th>Leftarm</th>
<th>Rightarm</th>
<th>Outcome indicator</th>
<th>Risk ratio</th>
<th>Statistically significant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pfe02</td>
<td>glaucoma</td>
<td>PGA</td>
<td>BB</td>
<td>safe IOP</td>
<td>1.43</td>
<td>no</td>
</tr>
<tr>
<td>Fel02</td>
<td>glaucoma</td>
<td>PGA</td>
<td>BB</td>
<td>safe IOP</td>
<td>1.29</td>
<td>yes</td>
</tr>
</tbody>
</table>

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 $\tau_1$ and $\tau_2$ with respect to the outcome indicator $\mu$, 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 $\tau_1$ is superior to $\tau_2$ with respect to $\mu$; (2) $\tau_1 < \mu \tau_2$, meaning that we believe that the result supports the claim that treatment $\tau_1$ is inferior to $\tau_2$ with respect to $\mu$; And (3) $\tau_1 \sim \mu \tau_2$, meaning that we believe the result as supporting the claim that neither $\tau_1$ nor $\tau_2$ is superior to each other with respect to $\mu$; 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 $\epsilon$, 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, $\lambda$, 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 $\epsilon$.

Example 1. For TRIALS, let $\tau_1$ be the leftarm, let $\tau_2$ be the rightarm, let $\mu$ be the outcome indicator, and let $\gamma \in$ TRIALS.

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

Example 2. For TRIALS, let $\tau_1$ be the leftarm, let $\tau_2$ be the rightarm, let $\mu$ be 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 $(\Phi, \lambda, \epsilon)$ where $\Phi \subseteq$ TRIALS is a set of results, $\lambda$ is an inference rule, $\Phi$ satisfies the conditions of $\lambda$ and $\epsilon$ is the claim of $\lambda$ applied to $\Phi$.

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

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

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 $A = (\Phi_A, \lambda_A, \epsilon_A)$ and $B = (\Phi_B, \lambda_B, \epsilon_B)$ 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.

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

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 $\gamma_a$ is statistically significant and $\gamma_b$ 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 $\tau_1, \tau_2$ and an outcome indicator $\mu$, 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 $\epsilon$ is the claim of $A$, then $\epsilon$ 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 \( \text{TRIALS} = \{t_1, \ldots, t_n\} \) each of which uses the same outcome indicator and compares the same pair of treatments \( \tau_1 \) and \( \tau_2 \). We partition \( \text{TRIALS} \) into three sets \text{SUPERIOR}, \text{EQUITABLE}, and \text{INFERIOR}. Those in \text{SUPERIOR} are the trials for which \( \tau_2 \) was shown to be superior to \( \tau_1 \), those in \text{EQUITABLE} are the trials for which \( \tau_2 \) was shown to be equal to \( \tau_1 \), and those in \text{INFERIOR} are the trials for which \( \tau_2 \) was shown to be inferior to \( \tau_1 \). We also partition \( \text{TRIALS} \) into two sets \text{SIGNIFICANT} and \text{NONSIGNIFICANT}. Those in \text{SIGNIFICANT} are the trials for which the result is significant, and those in \text{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 \text{SUPERIOR}, \text{EQUITABLE}, \text{INFERIOR}, \text{SIGNIFICANT} and \text{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.

- \((\text{SUPERIOR}, R_x, \tau_1 > \tau_2)\) where \( R_x \in \{R_1, \ldots, R_{12}\} \)
- \((\text{INFERIOR}, R_x, \tau_1 < \tau_2)\) where \( R_x \in \{R_1, \ldots, R_{12}\} \)
- \((\text{EQUITABLE}, R_x, \tau_1 \approx \tau_2)\) where \( R_x \in \{R_{13}, \ldots, R_{15}\} \)
- \((\text{NONSIGNIFICANT}, R_{16}, \tau_1 \approx \tau_2)\)

Note, the items of evidence in \text{INFERIOR} state that \( \tau_1 \) is inferior to \( \tau_2 \) which is equivalent to stating that \( \tau_2 \) is superior to \( \tau_1 \). 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 \( R_x \in \{R_1, \ldots, R_{12}\} \) to generate an argument \((\text{INFERIOR}, R_x, \tau_1 < \tau_2)\) where \( R_x \in \{R_1, \ldots, R_{12}\} \).

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

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

We motivate the inference rules as follows. First, \( R_1, \ldots, R_4 \) are for when all the trials show superiority (of the left arm over the right arm), \( R_5, \ldots, R_9 \) are for when the majority of the trials show superiority, and \( R_{10}, \ldots, R_{12} \) 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. \(|\text{SUPERIOR}=\text{SIGNIFICANT}|\)

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, \( R_{13}, \ldots, R_{15} \) 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, \( R_{13}, \ldots, R_{15} \) 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.
Given a set TRIALS comparing \( \tau_1 \) and \( \tau_2 \), the inference rules \( R_1 \) to \( R_{16} \) impose constraints on what combinations of arguments are possible together in Args(TRIALS).

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

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

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 Rule(\( A_i \)) and Rule(\( A_j \)).

<table>
<thead>
<tr>
<th>Preference rule</th>
<th>Rule(( A_i ))</th>
<th>Rule(( A_j ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P_2 )</td>
<td>( R_2 )</td>
<td>( { R_{16} } )</td>
</tr>
<tr>
<td>( P_3 )</td>
<td>( R_3 )</td>
<td>( { R_{16} } )</td>
</tr>
<tr>
<td>( P_4 )</td>
<td>( R_4 )</td>
<td>( { R_{16} } )</td>
</tr>
<tr>
<td>( P_5 )</td>
<td>( R_5 )</td>
<td>( { R_{11}, R_{12}, R_{13}, R_{16} } )</td>
</tr>
<tr>
<td>( P_6 )</td>
<td>( R_6 )</td>
<td>( { R_{11}, R_{12}, R_{13}, R_{16} } )</td>
</tr>
<tr>
<td>( P_7 )</td>
<td>( R_7 )</td>
<td>( { R_{12}, R_{13}, R_{16} } )</td>
</tr>
<tr>
<td>( P_8 )</td>
<td>( R_8 )</td>
<td>( { R_{12}, R_{13}, R_{16} } )</td>
</tr>
<tr>
<td>( P_9 )</td>
<td>( R_9 )</td>
<td>( { R_8, R_{11}, R_{12}, R_{13}, R_{16} } )</td>
</tr>
<tr>
<td>( P_{10} )</td>
<td>( R_{10} )</td>
<td>( { R_8, R_{11}, R_{12}, R_{13}, R_{16} } )</td>
</tr>
<tr>
<td>( P_{11} )</td>
<td>( R_{11} )</td>
<td>( { R_8, R_{13} } )</td>
</tr>
</tbody>
</table>

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

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 \( A_i \in \text{Args(\text{Trial}s)} \) where Rule(\( A_i \)) \( \in \{ R_1, ..., R_4 \} \) and \( \text{Claim}(A_i) = \tau_1 > \tau_2 \), then there is no argument \( A_j \in \text{Args(\text{Trial}s)} \) where Rule(\( A_j \)) \( \in \{ R_5, ..., R_{15} \} \) and either \( \text{Claim}(A_j) = \tau_1 < \tau_2 \) or \( \text{Claim}(A_j) = \tau_1 \sim \tau_2 \).

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 \( A_i \in \text{Args(\text{Trial}s)} \) s.t. \( \text{Claim}(A_i) = \tau_1 \sim \tau_2 \) and either Rule(\( A_i \)) = \( R_{13} \) or Rule(\( A_i \)) = \( R_{14} \), then there is an argument \( A_j \in \text{Args(\text{Trial}s)} \) where Rule(\( A_j \)) \( \in \{ R_5, ..., R_{12} \} \) and either \( \text{Claim}(A_j) = \tau_1 > \tau_2 \) or \( \text{Claim}(A_j) = \tau_1 < \tau_2 \).

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 \( A_i \in \text{Args(\text{Trial}s)} \) where Rule(\( A_i \)) = \( R_{15} \) then there is no argument \( A_j \in \text{Args(\text{Trial}s)} \) where 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 \( \nu \) 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 \( P_2 \), 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.

\[
\langle \text{SUPERIOR}, R_3, \text{PGA} \succ \text{respiratory BB} \rangle \rightarrow \langle \text{EQUITABLE}, R_{16}, \text{PGA} \sim \text{respiratory BB} \rangle
\]

**Example 7.** For prostaglandin v beta-blocker (\(m_4\) in Table 4), for lower risk of cardio-logical 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.

\[
\langle \text{SUPERIOR}, R_7, \text{PGA} > \text{cardio BB} \rangle \rightarrow \langle \text{INFERIOR}, R_{12}, \text{PGA} < \text{cardio BB} \rangle
\]

\[
\langle \text{NONSIGNIFICANT}, R_{16}, \text{PGA} \sim \text{cardio BB} \rangle
\]

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 = |\text{SUPERIOR} \cap \text{SIGNIFICANT}|, n_2 = |\text{SUPERIOR} \cap \text{NONSIGNIFICANT}|, n_3 = |\text{EQUITABLE}|, n_4 = |\text{INFERIOR} \cap \text{NONSIGNIFICANT}|, and \(n_5 = |\text{INFERIOR} \cap \text{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 \(n_1, \ldots, n_5, \) 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.

<table>
<thead>
<tr>
<th>( n_1 )</th>
<th>( n_2 )</th>
<th>( n_3 )</th>
<th>( n_4 )</th>
<th>( n_5 )</th>
<th>Their result</th>
<th>Rules used</th>
<th>Our result</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>sup sig</td>
<td>( R_2 )</td>
<td>( \tau_1 &gt; \tau_2 )</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>sup sig</td>
<td>( R_2 )</td>
<td>( \tau_1 &gt; \tau_2 )</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>sup non-sig</td>
<td>( R_2, R_{16} )</td>
<td>( \tau_1 &gt; \tau_2 )</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>sup non-sig</td>
<td>( R_7, R_{12}, R_{16} )</td>
<td>( \tau_1 &gt; \tau_2 )</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>inf non-sig</td>
<td>( R_8, R_{16} )</td>
<td>( \tau_1 &gt; \tau_2, \tau_1 \sim \tau_2, \tau_1 &lt; \tau_2 )</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>6</td>
<td>inf sig</td>
<td>( R_2 )</td>
<td>( \tau_1 &lt; \tau_2 )</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>( n_1 )</th>
<th>( n_2 )</th>
<th>( n_3 )</th>
<th>( n_4 )</th>
<th>( n_5 )</th>
<th>Their result</th>
<th>Rules used</th>
<th>Our result</th>
</tr>
</thead>
<tbody>
<tr>
<td>( m_7 )</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>sup non-sig</td>
<td>( R_8, R_{12}, R_{16} )</td>
</tr>
<tr>
<td>( m_8 )</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>inf non-sig</td>
<td>( R_8, R_{12}, R_{16} )</td>
</tr>
<tr>
<td>( m_9 )</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>inf non-sig</td>
<td>( R_8, R_{12}, R_{16} )</td>
</tr>
<tr>
<td>( m_{10} )</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>inf sig</td>
<td>( R_7, R_{12}, R_{16} )</td>
</tr>
<tr>
<td>( m_{11} )</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>sup sig</td>
<td>( R_2 )</td>
</tr>
</tbody>
</table>

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 left arm is biphasic insulin and the right arm is human insulin; for \( m_{14} \), the left arm is Glargin insulin and the right arm is human insulin; and for \( m_{15} \), the left arm is biphasic insulin and the right arm is glargin insulin.

<table>
<thead>
<tr>
<th>( n_1 )</th>
<th>( n_2 )</th>
<th>( n_3 )</th>
<th>( n_4 )</th>
<th>( n_5 )</th>
<th>Their result</th>
<th>Rules used</th>
<th>Our result</th>
</tr>
</thead>
<tbody>
<tr>
<td>( m_{12} )</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>sup non-sig</td>
<td>( R_8, R_{12}, R_{13}, R_{16} )</td>
</tr>
<tr>
<td>( m_{13} )</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>inf non-sig</td>
<td>( R_8, R_{16} )</td>
</tr>
<tr>
<td>( m_{14} )</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>sup sig</td>
<td>( R_1 )</td>
</tr>
</tbody>
</table>

...ther it means 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