

# Argumentation about Treatment Efficacy

Nikos Gorogiannis<sup>1</sup>, Anthony Hunter<sup>1</sup>, Vivek Patkar<sup>2</sup>, and Matthew Williams<sup>3</sup>

<sup>1</sup> Department of Computer Science, University College London,  
Gower Street, WC1E 6BT, London, UK

{n.gorogiannis,a.hunter}@cs.ucl.ac.uk

<sup>2</sup> University College London Cancer Institute,  
72 Huntley Street, London WC1E 6BT, UK

v.patkar@medsch.ucl.ac.uk

<sup>3</sup> Mount Vernon Hospital, Northwood, Middlesex, HA6 2RN, UK

matt.williams@nhs.net

**Abstract** The volume and complexity of knowledge produced by medical research calls for the development of technology for automated management and analysis of such knowledge. In this paper, we identify scenarios where a researcher or a clinician may wish to use automated systems for analysing knowledge from clinical trials. For this, we propose a language for encoding, capturing and synthesising knowledge from clinical trials and a framework that allows the construction of arguments from such knowledge. We develop this framework and demonstrate its use on a case study regarding chemotherapy regimens for ovarian cancer.

## 1 Introduction

During the last few decades, medical research has grown rapidly, producing an enormous amount of results. This growth necessitates a change in how this knowledge is managed, searched and analysed in order to achieve the full potential offered. Information retrieval techniques are already indispensable for users. However, they do not address all the needs, as they do not enable one to *reason* with the available knowledge. This is where knowledge representation techniques can make a contribution. In this paper, we propose a framework for expressing and analysing this knowledge in an effective manner.

One important kind of such knowledge is associated with superiority-testing clinical trials, a type of clinical study which involves comparing the efficacy of two or more treatments when given to a particular class of patients. Each such trial reports the results of comparing two or more treatments. In order to have a global view of the relative comparisons between treatments for a particular condition, a potentially large number of publications needs to be reviewed and this is why syntheses of this kind of knowledge are undertaken. Normally, a group of scientists would search the literature and perform a study employing background medical knowledge and statistical techniques for aggregating clinical trial results (e.g., a systematic review or meta-analysis), requiring significant time and effort. In addition, such syntheses of the literature can quickly become out of date if new results are published in the interim.

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 analysing 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 aims to suggest such a method. The first part of the 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. Using such a knowledge base clinical scientists can access easily the desired set of evidence. The second part in our framework allows the construction of arguments on the basis of evidence as well as their syntheses, published or generated on-the-fly. These arguments encode judgments about the comparison of efficacy of two treatments, thus representing interpretations of the available evidence. These are then presented and organised 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.

To demonstrate our proposal, we present worked examples from a case study on chemotherapy treatments for ovarian cancer. We use an existing paper that reports the results of synthesising the available evidence in the area [1], and show how the user would navigate the evidence in the literature, create his/her own syntheses and produce arguments and counter-arguments for comparisons of treatment efficacy.

## 2 A Language for Representing Clinical Trial Results

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; other trial design types do exist, but we will ignore them here for simplicity.

We give below the details of two results used in [1] (which we will call the Ovarian Cancer Study from now on), encoded in the proposed language, which will be explained in this section.

Result	tr <sub>1</sub>	tr <sub>2</sub>	ind	hr	sig
$\gamma_{\text{Markman}}$	$PTC\bar{I}$	$PTCI$	$S$	1.23	$\perp$
$\gamma_{\text{Armstrong}}$	$PTC\bar{I}$	$PTCI$	$S$	1.33	$\top$

The first row corresponds to a superiority trial performed by Markman et al in 2001, which compared two chemotherapy regimens for ovarian cancer patients: the combination of agents paclitaxel and cisplatin, administered intravenously (IV) with the combination of carboplatin (IV), paclitaxel (IV) and carboplatin, administered intraperitoneally (IP). The study produced a hazard ratio of 1.23 for survival and did not find a statistically significant difference between the two regimens. We now explain what these data mean and how we encode them.

To represent a trial result we will use  $\gamma$ , possibly with subscripts, e.g.,  $\gamma_{\text{Markman}}$  for the trial result by Markman et al. We also will use  $\Gamma$  to denote the set of results under consideration. The characteristics of a trial result will be called **attributes**. We use a set  $Attr$  of partial functions from trial results to values, to represent the set of available characteristics.

The first attribute of interest of a trial is the patient class involved. *Medical ontologies* provide languages for capturing patient characteristics (among other things) as well as logical machinery for answering queries about them. Description logics are usually employed to provide the necessary inference tools (see [2]). Ontologies, however, fall outside the scope of this paper; for simplicity we assume that the set of results  $\Gamma$  concerns a particular, sensible patient class.

The next component of our language concerns treatments. Again, medical ontologies cater for this task by providing categories and relationships on treatments, substances used, and other characteristics. For simplicity, we will elide such complexity and assume that there is a set algebra of treatments  $\mathcal{T}$  whose members, denoted by  $\tau$  possibly with subscripts, represent specific treatments. In the example, the phrases “paclitaxel and cisplatin, administered intravenously (IV)” and “carboplatin (IV), paclitaxel (IV) and carboplatin, administered intraperitoneally (IP)” would be denoted by, e.g.  $\tau_1$  and  $\tau_2$ . We will use  $\text{tr}_1(\gamma)$  and  $\text{tr}_2(\gamma)$  to signify the treatments compared in the trial result  $\gamma$ .

Clinicians and statisticians, however, may want to group treatments together in order to allow for more inferential power. In such cases, it is customary to conflate the details of the treatments down to the level of abstraction required. For example, the Ovarian Cancer Study is only concerned with four characteristics of the treatment: whether or not it employs platinum-based agents ( $P$ ), taxanes ( $T$ ), whether it combines several agents ( $C$ ), and whether it employs at least one agent with intraperitoneal administration ( $I$ ). We will use four symbols to denote each class of treatments, with a bar over a symbol to denote its negation. Therefore,  $\tau_1$  is a member of the treatment group  $PTC\bar{I}$  and  $\tau_2$  is a member of  $PTCI$ . From now on we use this treatment clustering scheme for defining  $\mathcal{T}$  and assume that  $\text{tr}_1, \text{tr}_2$  will range over the treatment groups allowed. We also use the term *treatment* to refer to any treatment group.

A trial comparing two treatments will do so with respect to particular outcomes, e.g., in the case of the Ovarian Cancer Study, survival. Moreover, a specific statistical measure will be used to evaluate outcomes. In our example, the hazard ratio for survival is used, which we shall explain below. We call the particular way an outcome is represented in a trial result, the **outcome indicator** and use the attribute `ind` to retrieve it, given a particular trial. Hence, in our example,  $\text{ind}(\gamma_{\text{Markman}}) = S$ , denoting hazard ratios for survival. Here, for simplicity, we only consider sets of results that involve the same outcome indicator.

Finally, a trial would report the results of the statistical method comparing the two treatments. There are many different measures, each appropriate to specific trial designs and outcomes. The trials examined in the Ovarian Cancer Study use *hazard ratios*. For a given trial it is possible, using statistical methods which are outside the scope of this paper, to estimate the number of expected

events (death in the case of survival) in each group on the basis of the assumption that the two treatments have identical effects (null hypothesis). Then, the *hazard rate* for each group can be computed, which is the ratio of the number of observed events to the number of expected events. The hazard ratio is the ratio of the hazard rates for the two groups. In the case of survival, a hazard ratio over 1 indicates that the group treated with  $\tau_1$  fared worse than the group treated with  $\tau_2$ . We will use the function  $\text{hr} \in \text{Attr}$  applied to the trial result  $\gamma \in \Gamma$  to retrieve the value of the hazard ratio reported,  $\text{hr}(\gamma)$ . In addition, a trial will report an indicator of whether this comparison is statistically significant. This can be done in various ways, e.g., whether the  $p$ -value is greater or lower than the significance level (commonly 5%), or whether the confidence interval includes the hazard ratio of 1 (in which case the possibility that the two treatments have exactly the same effect cannot be ruled out). However this is reported, the aim is to ascertain whether the result is statistically significant, so we will use a boolean attribute  $\text{sig}(\gamma)$  to denote whether a result  $\gamma$  is statistically significant. Therefore, for our example we would have  $\text{hr}(\gamma_{\text{Markman}}) = 1.23$ , and  $\text{sig}(\gamma_{\text{Markman}}) = \perp$ .

### 3 Arguments drawn from Clinical Trial Results

For a superiority clinical trial comparing treatments  $\tau_1$  and  $\tau_2$  with respect to the outcome indicator  $I$ , there are four possible interpretations of its results:

1.  $\tau_1 >_I \tau_2$ , meaning that we believe that the result supports the inference that treatment  $\tau_1$  is superior to  $\tau_2$  with respect to  $I$ .
2.  $\tau_1 <_I \tau_2$ , as above.
3.  $\tau_1 \sim_I \tau_2$ , meaning that we interpret the result as supporting the inference that neither  $\tau_1$  nor  $\tau_2$  is superior to each other with respect to  $I$ .
4. Finally, it may be the case that we believe that the result does not support any of the above possibilities.

We will call the formulae employed in statements 1–3 **claims**. Formally, any formula of the form  $\tau_1 >_I \tau_2$ ,  $\tau_1 \sim_I \tau_2$  and  $\tau_1 <_I \tau_2$  will be called a claim, denoted by  $\epsilon$ , possibly subscripted. We will sometimes use  $\circ, \diamond$  as meta-variables for the symbols  $>, <$  and  $\sim$ . So, for example,  $\tau_1 \circ_I \tau_2$  will stand for either of:  $\tau_1 >_I \tau_2, \tau_1 <_I \tau_2$  or  $\tau_1 \sim_I \tau_2$ .

Note that the interpretation of the results of a trial is a very complex and, in some cases, open question and that conflicting answers exist. For example, a strict statistical interpretation would be that a superiority trial can never provide evidence supporting the equivalence of two treatments, *by design*. However, some clinicians would argue that a sufficiently large superiority trial that fails to show superiority is, in fact, in itself evidence of (rough) equivalence. We will refrain from debating these issues, as they are well outside the scope of this work. However, we will allow for different users to express their own semantics.

Given a set of results  $\Gamma$  one can informally think of an argument comprising of a set of evidence, an inferential rule and a conclusion or claim. For example, a plausible interpretation of  $\gamma_{\text{Armstrong}}$  is that since  $\text{hr}(\gamma_{\text{Armstrong}}) > 1$ , it indicates

that the first treatment is worse than the second with respect to  $S$ , i.e., that  $PTC\bar{I} <_S PTCI$ . We define this process of inference as an inference rule.

**Definition 1.** An *inference rule* is a set of conditions (employing set-theoretic expressions and equations utilising attributes over the reals) on a set of results  $X \subseteq \Gamma$  and a claim  $\epsilon$ .

From now on we will use  $\lambda$ , possibly subscripted, to denote an inference rule, and  $\Lambda$  for the set of all inference rules chosen by the user.

*Example 1.* One of the simplest inference rules interprets a statistically significant trial result as evidence of actual treatment superiority.

$\lambda_s$ : Let  $X = \{\gamma\}$ . If  $\text{sig}(\gamma) = \top$  then:

- if  $\text{hr}(\gamma) < 1$  then  $\text{tr}_1(\gamma) >_{\text{ind}(\gamma)} \text{tr}_2(\gamma)$ ,
- if  $\text{hr}(\gamma) > 1$  then  $\text{tr}_1(\gamma) <_{\text{ind}(\gamma)} \text{tr}_2(\gamma)$ .

*Example 2.* Conversely, a non-significant result can be viewed as saying that neither treatment is better than the other one in terms of the outcome indicator considered. This can be captured by the following inference rule:

$\lambda_n$ : Let  $X = \{\gamma\}$ . If  $\text{sig}(\gamma) = \perp$  then  $\text{tr}_1(\gamma) \sim_{\text{ind}(\gamma)} \text{tr}_2(\gamma)$ .

An inference rule can be thought of as an *argument generator* in that, on the basis of a set of results that stands as evidence, it supports a certain claim. Therefore, we will define an argument as an application of an inference rule.

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

*Example 3.* Given  $\gamma_{\text{Armstrong}}$ , the following tuple is an argument:

$$A_1 = \langle \{\gamma_{\text{Armstrong}}\}, \lambda_s, PTC\bar{I} <_S PTCI \rangle.$$

By using  $\lambda_n$  on  $\gamma_{\text{Markman}}$  we obtain the following argument.

$$A_2 = \langle \{\gamma_{\text{Markman}}\}, \lambda_n, PTC\bar{I} \sim_S PTCI \rangle$$

Clearly,  $A_1$  and  $A_2$  are conflicting. We focus on conflict in the next section.

**Definition 3.** Given a claim  $\epsilon$ , we define the *set of arguments relevant to  $\epsilon$*  as

$$\text{args}(\epsilon) = \{ A \mid A = \langle X, \lambda, \epsilon \rangle \text{ is an argument with } X \subseteq \Gamma, \lambda \in \Lambda \}.$$

Given a pair of treatments  $\tau_1, \tau_2$  and an outcome indicator  $I$  we define the set of arguments relevant to  $\tau_1, \tau_2$  with respect to  $I$  as

$$\text{args}(\tau_1, \tau_2, I) = \text{args}(\tau_1 >_I \tau_2) \cup \text{args}(\tau_1 \sim_I \tau_2) \cup \text{args}(\tau_1 <_I \tau_2)$$

*Example 4.* Let  $\Gamma = \{\gamma_{\text{Markman}}, \gamma_{\text{Armstrong}}\}$ . Then,

$$\begin{aligned} \text{args}(PTC\bar{I} <_S PTCI) &= \{\langle \{\gamma_{\text{Armstrong}}\}, \lambda_s, PTC\bar{I} <_S PTCI \rangle\} \\ \text{args}(PTC\bar{I} \sim_S PTCI) &= \{\langle \{\gamma_{\text{Markman}}\}, \lambda_n, PTC\bar{I} \sim_S PTCI \rangle\} \\ \text{args}(PTC\bar{I} >_S PTCI) &= \emptyset \\ \text{args}(PTC\bar{I}, PTCI, S) &= \left\{ \left\langle \begin{array}{l} \{\gamma_{\text{Armstrong}}\}, \lambda_s, PTC\bar{I} <_S PTCI \\ \{\gamma_{\text{Markman}}\}, \lambda_n, PTC\bar{I} \sim_S PTCI \end{array} \right\rangle, \right\} \end{aligned}$$

The intention behind these definitions is that users should be able to define or select their own inference rules for argument construction.

## 4 Preferences over Arguments

In Example 4 we saw two arguments that were clearly in conflict:

$$\langle \{\gamma_{\text{Armstrong}}\}, \lambda_s, PTC\bar{I} <_S PTCI \rangle, \langle \{\gamma_{\text{Markman}}\}, \lambda_n, PTC\bar{I} \sim_S PTCI \rangle$$

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 kind of conflict with the following definition.

**Definition 4.** *If  $A = \langle X_A, \lambda_A, \epsilon_A \rangle$  and  $B = \langle X_B, \lambda_B, \epsilon_B \rangle$  are two arguments where  $\epsilon_A = \tau_1 \circ_I \tau_2$ , then we say that  $A$  and  $B$  **attack**, **conflict with** or **rebut** each other whenever:*

1.  $\epsilon_A = \tau_1 >_I \tau_2$ , and  $\epsilon_B \in \{\tau_1 \sim_I \tau_2, \tau_2 \sim_I \tau_1, \tau_1 <_I \tau_2, \tau_2 >_I \tau_1\}$ .
2.  $\epsilon_A = \tau_1 \sim_I \tau_2$ , and  $\epsilon_B \in \{\tau_1 >_I \tau_2, \tau_2 <_I \tau_1, \tau_1 <_I \tau_2, \tau_2 >_I \tau_1\}$ .
3.  $\epsilon_A = \tau_1 <_I \tau_2$ , and  $\epsilon_B \in \{\tau_1 \sim_I \tau_2, \tau_2 \sim_I \tau_1, \tau_1 >_I \tau_2, \tau_2 <_I \tau_1\}$ .

Note that this definition is symmetric, i.e., if  $A$  attacks  $B$  then  $B$  attacks  $A$ .

We will organise the arguments into a graph that we will use as an argumentation framework in the sense of [3]. To do this, we first define an attack relation on arguments in the obvious way:  $\mathcal{R}(A, B)$  is true iff  $A$  attacks  $B$ , for arguments  $A, B \in \text{args}(\tau_1, \tau_2, I)$ . It is easy to see that the graph induced is tripartite, and its independent sets are  $\text{args}(\tau_1 >_I \tau_2)$ ,  $\text{args}(\tau_1 \sim_I \tau_2)$ ,  $\text{args}(\tau_1 <_I \tau_2)$ . In our example, this graph would be as follows.

$$\begin{array}{c} \langle \{\gamma_{\text{Armstrong}}\}, \lambda_s, PTC\bar{I} <_S PTCI \rangle \\ \downarrow \uparrow \\ \langle \{\gamma_{\text{Markman}}\}, \lambda_n, PTC\bar{I} \sim_S PTCI \rangle \end{array}$$

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

**Definition 5.** A *preference rule*  $\pi$  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  $B$  according to  $\pi$ , and we write  $\pi(A, B) = \top$ . Otherwise, we say that  $A$  is not preferred to  $B$  and we write  $\pi(A, B) = \perp$ .

We will use  $\pi$ , possibly with subscripts, to denote a preference rule, and  $\Pi$  for the set of preference rules chosen by the user.

*Example 5.* A preference rule that considers statistically significant results as more authoritative than non significant ones can be seen below.

$$\pi_s(\langle \{\gamma_a\}, \lambda_a, \tau_1 \circ_I \tau_2 \rangle, \langle \{\gamma_b\}, \lambda_b, \tau_1 \diamond_I \tau_2 \rangle) = \top \text{ iff } \begin{cases} \text{sig}(\gamma_a) = \top \\ \text{and sig}(\gamma_b) = \perp \end{cases}$$

As mentioned previously,  $\circ$  and  $\diamond$  are meta-variables for the symbols  $<, >, \sim$ , in other words  $\circ$  and  $\diamond$  stand for  $<, >$  or  $\sim$ . Here, since by definition the arguments are conflicting, we do not need to specify the values of these meta-variables.

Preference rules are not required to be infallible in any sense. Indeed,  $\pi_s$  embodies one of the aspects of *publication bias*, where by preferring significant results to non-significant ones, one may miss evidence that support 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  $\mathcal{R}$ , and capture the new subrelation  $\mathcal{R}_\Pi$  of  $\mathcal{R}$  as follows.

**Definition 6.** For any pair of arguments  $A, B \in \mathcal{A}$ ,  $\mathcal{R}_\Pi(A, B)$  is true whenever

- $\mathcal{R}(A, B)$  is true and,
- if there is a preference rule  $\pi \in \Pi$  such that  $\pi(B, A) = \top$  then there is also a preference rule  $\pi' \in \Pi$  such that  $\pi'(A, B) = \top$ .

The motivation here is that if  $A$  and  $B$  attack each other and  $A$  is preferred to  $B$  then  $B$ 's attack on  $A$  is cancelled. However, this wording leads to problems when  $A$  is preferred to  $B$  according to a rule  $\pi_1$  and  $B$  is preferred to  $A$  according to  $\pi_2$ . 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 can put together these components by defining an abstract argumentation framework in the sense of Dung's work [3].

**Definition 7.** Given a pair of treatments  $\tau_1, \tau_2$  and an outcome indicator  $I$ , we define the argumentation framework  $AF_{\tau_1, \tau_2, I}$  as a pair  $\langle \mathcal{A}, \mathcal{R}_\Pi \rangle$  where  $\mathcal{A} = \text{args}(\tau_1, \tau_2, I)$ .

Dung [3] defines a notion of admissibility: an admissible set of arguments  $\mathcal{S}$  is one which contains no argument that attacks another argument in  $\mathcal{S}$  and for any argument  $A_1 \in \mathcal{S}$  that is attacked by argument  $A_2 \notin \mathcal{S}$ , there is another argument  $A_3 \in \mathcal{S}$  that attacks  $A_2$ . Admissible sets of arguments that are maximal with respect to set inclusion are called *preferred*.

## 5 Case Study

In this section, we expand our running example and look at how a user might navigate the information contained in the Ovarian Cancer Study. We will see how statistical methods for aggregating results (*meta-analysis*) can be seen as types of arguments that may conflict or agree with other arguments.

In general, a meta-analysis consists of taking a weighted average of a set of results, e.g., a weighted average of hazard ratios. The way the weights are computed depends on statistical assumptions relating to the clinical trials involved. The two main methods, corresponding to different kinds of assumptions, for meta-analyses are the *fixed effects model* and the *random effects model*. Here, we will not look at the details of how each method is defined, but assume that one has been chosen, based on the trial results available. We will use it as a function  $\text{MA}$  that takes a set of results as argument and returns a *result* that is to be thought of as a virtual trial result. For instance,  $\text{hr}(\text{MA}(X))$  represents the hazard ratio of the result produced by a meta-analysis on the set of results  $X$ . The inference rule  $\lambda_{MA}$  will then be defined as follows.

$\lambda_{MA}$ : Let  $X$  be a set of results such that for all  $\gamma \in X$ ,  $\text{tr}_1(\gamma) = \tau_1$ ,  $\text{tr}_2(\gamma) = \tau_2$  and  $\text{ind}(\gamma) = I$  for given treatments  $\tau_1, \tau_2$  and outcome indicator  $I$ .

If  $\text{sig}(\text{MA}(X)) = \top$  then:

- if  $\text{hr}(\text{MA}(X)) < 1$  then  $\tau_1 >_I \tau_2$ ,
- if  $\text{hr}(\text{MA}(X)) > 1$  then  $\tau_1 <_I \tau_2$ .

Else, if  $\text{sig}(\text{MA}(X)) = \perp$  then  $\tau_1 \sim_I \tau_2$ .

To avoid bias, a meta-analysis should be performed on *all* eligible results, and that is why the definition of  $\lambda_{MA}$  puts such conditions on  $X$ .

Returning to the example, below are listed all the trial results reported in the Ovarian Cancer Study that compare the treatments  $T_1 = \bar{P}\bar{T}\bar{C}\bar{I}$  with  $T_2 = \bar{P}\bar{T}\bar{C}\bar{I}$ . We will denote this set of results as  $\Gamma'$ .

Result	hr ? 1	sig
$\gamma_{\text{Trope}}$	<	$\top$
$\gamma_{\text{Barlow}}$	<	$\perp$
$\gamma_{\text{Bruckner}}$	<	$\perp$
$\gamma_{\text{Aabo}}$	<	$\perp$
$\gamma_{\text{Delgado}}$	<	$\perp$
$\gamma_{\text{Gronroos}}$	<	$\perp$

Result	hr ? 1	sig
$\gamma_{\text{Adams}}$	<	$\perp$
$\gamma_{\text{Carmo-Pereira}}$	>	$\perp$
$\gamma_{\text{Scott}}$	<	$\perp$
$\gamma_{\text{Park}}$	<	$\perp$
$\gamma_{\text{Omura}}$	<	$\perp$
$\gamma_{\text{Young}}$	<	$\perp$

Kyrgiou et al compute the result of the meta-analysis on these trials, and it turns out that  $\text{hr}(\text{MA}(\Gamma')) = 0.83 < 1$  and  $\text{sig}(\text{MA}(\Gamma')) = \top$ .

Using the inference rules  $\lambda_s, \lambda_n, \lambda_{MA}$  on these results we get:

$$\mathcal{A}_1 = \text{args}(T_1 >_S T_2) = \left\{ \begin{array}{l} \langle \{\gamma_{\text{Trope}}\}, \lambda_s, T_1 >_S T_2 \rangle, \\ \langle \Gamma', \lambda_{MA}, T_1 >_S T_2 \rangle \end{array} \right\}$$

$$\mathcal{A}_2 = \text{args}(T_1 \sim_S T_2) = \left\{ \begin{array}{l} \langle \{\gamma_{\text{Barlow}}\}, \lambda_n, T_1 \sim_S T_2 \rangle, \\ \vdots \\ \langle \{\gamma_{\text{Young}}\}, \lambda_n, T_1 \sim_S T_2 \rangle \end{array} \right\}$$

It should be clear that for every  $A_1 \in \mathcal{A}_1$  and  $A_2 \in \mathcal{A}_2$ ,  $A_1$  attacks  $A_2$  and vice versa. We now look at using preference rules with these sets.

Observe that  $\text{args}(T_1, T_2, S) = \mathcal{A}_1 \cup \mathcal{A}_2$  and that in the argumentation framework  $\langle \mathcal{A}_1 \cup \mathcal{A}_2, \mathcal{R} \rangle$  there are exactly two preferred sets, namely  $\mathcal{A}_1$  and  $\mathcal{A}_2$ . This indicates that we do not have enough information and/or expressed preferences to distinguish between the two possibilities, i.e.,  $T_1 >_S T_2$  or  $T_1 \sim_S T_2$ .

Suppose the user applies preference rule  $\pi_s$ , meaning that statistically significant results are preferred to non-significant ones. Then, using  $\Pi = \{\pi_s\}$ , in the framework  $AF_{T_1, T_2, S}$  there remains only one maximal admissible set,  $\mathcal{A}_1$ , since it contains an argument based on a statistically significant result,  $\gamma_{\text{Trope}}$ .

Alternatively, suppose a new preference rule is used that prefers the result of a meta-analysis to a result that is included in the meta-analysis. In other words:

$$\pi_{MA}(\langle X, \lambda_{MA}, \tau_1 \circ_I \tau_2 \rangle, \langle \{\gamma\}, \lambda_x, \tau_1 \diamond_I \tau_2 \rangle) = \top \text{ iff } \lambda_x \in \{\lambda_s, \lambda_n\}$$

Here,  $\lambda_x$  is allowed to be any rule that interprets a single result used in the meta-analysis. Using  $\Pi = \{\pi_{MA}\}$ , we obtain exactly one preferred set,  $\mathcal{A}_1$ .

Another emerging method for aggregating trial results is *network meta-analysis*, which is a kind of meta-analysis that uses indirect treatment comparisons as well as direct ones. This means that when there is a trial result  $\gamma_{a,b}$  comparing  $\tau_a$  and  $\tau_b$ , and another result  $\gamma_{b,c}$  comparing  $\tau_b$  to  $\tau_c$ , an estimate of the comparison of  $\tau_a$  and  $\tau_c$  can be computed, and  $\gamma_{a,b}, \gamma_{b,c}$  can be thought of as a path from  $\tau_a$  to  $\tau_c$ . Network meta-analyses use such estimates together with direct comparisons (e.g., of  $\tau_a$  and  $\tau_c$ ) to provide a weighted average of higher inferential power. Without going into details, we use a function  $\text{NMA}(\tau_a, \tau_b, X)$ , where  $\tau_a$  and  $\tau_b$  are the treatments to compare, and  $X$  is the set of all results that belong in paths between  $\tau_a$  and  $\tau_b$ . Then, we will use an inference rule  $\lambda_{NMA}$  that resembles  $\lambda_{MA}$  in all aspects apart from using a network meta-analysis. We note that acceptance of network meta-analyses as an analytical method is not unanimous and this reinforces the utility of a framework where preferences are user-expressible.

The addition of  $\lambda_{NMA}$  to the set of rules  $\Lambda = \{\lambda_s, \lambda_n, \lambda_{MA}\}$  allows for another argument to be generated in our example. The result of the network meta-analysis for  $T_1$  and  $T_2$  was computed in the Ovarian Cancer Study, and it agrees with the standard meta-analysis:  $\text{hr}(\text{NMA}(I'')) = 0.87 < 1$  and  $\text{sig}(\text{NMA}(I'')) = \top$ . Note that here we use  $I''$  to denote all the results that lie on paths between

$T_1$  and  $T_2$ ; clearly  $I' \subseteq I''$ . Now, the set  $\mathcal{A}_1$  is as follows.

$$\mathcal{A}_1 = \text{args}(T_1 >_S T_2) = \left\{ \begin{array}{l} \langle \{\gamma_{\text{Trope}}\}, \lambda_s, T_1 >_S T_2 \rangle, \\ \langle I', \lambda_{MA}, T_1 >_S T_2 \rangle, \\ \langle I'', \lambda_{NMA}, T_1 >_S T_2 \rangle \end{array} \right\}$$

Similarly to the case of the meta-analysis preference rule, one is needed for network meta-analyses too.

$$\pi_{NMA}(\langle X, \lambda_{NMA}, \tau_1 \circ_I \tau_2 \rangle, \langle \{\gamma\}, \lambda_x, \tau_1 \diamond_I \tau_2 \rangle) = \top \text{ iff } \lambda_x \in \{\lambda_s, \lambda_n\}$$

Using  $\Pi = \{\pi_{MA}, \pi_{NMA}\}$  yields as preferred the set  $\mathcal{A}_1$  once again.

## 6 Discussion and Conclusions

We have presented a framework for argumentation on treatment efficacy. Its major components are: (1) A language for encoding results from clinical trials, which we believe has many other potential uses such as intelligent querying. (2) A definition delineating the ways arguments on treatment efficacy comparisons can be produced from trial results, along with a definition of a notion of conflict. (3) A definition of a preference rule that allows the potential resolution of conflict depending on the characteristics of the conflicting arguments. Using these components along with standard argumentation tools, users can describe their preferences and analyse the available evidence in terms of agreement and conflict. Finally we have presented the potential use of this framework in a case study on ovarian cancer.

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 ([4,5] and others) exists that uses argumentation as a tool in medical decision support, but as such, assumes the existence of a hand-crafted set of facts around treatment efficacy. Work that is concerned with the capture of a wide spectrum of data about clinical trials exists, [6], and would potentially provide a useful basis for the continuation of our work.

The avenues for further work are several. While not discussed in this paper, the systematic selection and filtering of the trial data used is very important, and requires the formalisation and encoding of many more kinds of meta-data about trials, which we hope to investigate in the future. Such meta-data would include, e.g., the year a trial started (or ended), the location, the healthcare setting, the sample/power size and potentially many other attributes. Prior work exists in this area (see, e.g., [7]) and it would be interesting to see whether argumentation can help in this process as well.

A deeper case study involving clinicians as users would be beneficial in making more concrete the requirements and preferences of such users. Also, liaising with researchers on automated information extraction would provide means to constructing the repositories of trial results that would add significant value to this research. Finally, the construction of a user-friendly system will be beneficial to the evaluation and adoption of the framework presented here.

## References

1. Kyrgiou, M., Salanti, G., Pavlidis, N., Paraskevaidis, E., Ioannidis, J.: Survival Benefits With Diverse Chemotherapy Regimens for Ovarian Cancer: Meta-analysis of Multiple Treatments. *Journal of the National Cancer Institute* **98**(22) (2006) 1655–1663
2. Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P., eds.: The description logic handbook: theory, implementation, and applications. Cambridge University Press, New York, NY, USA (2003)
3. Dung, P.: On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence* **77**(2) (1995) 321–357
4. Patkar, V., Hurt, C., Steele, R., Love, S., Purushotham, A., Williams, M., Thomson, R., Fox, J.: Evidence-based guidelines and decision support services: a discussion and evaluation in triple assessment of suspected breast cancer. *British Journal of Cancer* **95**(11) (2006) 1490–1496
5. Tolchinsky, P., Cortés, U., Modgil, S., Caballero, F., López-Navidad, A.: Increasing human-organ transplant availability: Argumentation-based agent deliberation. *IEEE Intelligent Systems* **21**(6) (2006) 30–37
6. Sim, I., Owens, D., Lavori, P., Rennels, G.: Electronic trial banks: A complementary method for reporting randomized trials. *Medical Decision Making* **20**(4) (2000) 440–450
7. Wyatt, J., Altman, D., Heathfield, H., Pantin, C.: Development of design-a-trial, a knowledge-based critiquing system for authors of clinical trial protocols. *Computer Methods and Programs in Biomedicine* **43**(3-4) (1994) 283–291