Argumentation for Aggregating Clinical Evidence

Anthony Hunter
Department of Computer Science
University College London
London, WC1E 6BT, UK
Email: a.hunter@cs.ucl.ac.uk

Matthew Williams
Department of Academic Oncology
Royal Free Hospital
London, NW3 2QG, UK
Email: mhw@doctors.net.uk

***Abstract*—Evidence-based decision making is becoming increasingly important in healthcare. Much valuable evidence is in the form of the results from clinical trials that compare the relative merits of treatments. For this, in previous papers [1], [2], we have proposed a general framework for representing and synthesizing knowledge from clinical trials involving the same outcome indicator. Now, in this paper, we present a new framework for representing and synthesizing knowledge from clinical trials involving *multiple* outcome indicators. In this framework, evidence from randomized clinical trials, systematic reviews, meta-analyses, network analyses, etc., comparing a pair of treatments τ_1 and τ_2 according to desired and/or undesired outcomes is aggregated to give an overall evaluation of the treatments saying τ_1 is superior to τ_2, or τ_1 is equivalent to τ_2, or τ_1 is inferior to τ_2. Our general framework incorporates inference rules for generating arguments and counterarguments for claiming that one treatment is superior to another based on the available evidence, and preference rules for specifying which arguments are preferred. In this paper, we also present a new version of this framework that incorporates utility-theoretic criteria for defining specific preference rules over arguments.**

***Keywords*-Logical argumentation; Knowledge aggregation; Decision-support systems; Evidence-based medicine**

I. Introduction

The systematic use of evidence is already established in healthcare. However, the rapidly increasing amount of evidential knowledge on a subject means that it is difficult for a clinician or biomedical researcher to effectively and efficiently acquire and assimilate that evidence. Therefore, getting a quick, up-to-date review of the state of the art on treatment efficacy for a particular condition is not always feasible. This problem is exacerbated by the fact that the evidence is often conceptually complex, heterogeneous, incomplete and inconsistent. Hence, there is a need 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. Moreover, 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. As a first step to addressing these needs, our proposal in [3], [1], [2] presents a general framework for representing and synthesizing knowledge from clinical trials involving the *same* outcome indicator (e.g. overall survival, or disease-free survival).

In this paper, to further address these needs of aggregating evidence, we present a new framework for representing and synthesizing knowledge from clinical trials involving *multiple* outcome indicators. Our framework allows the construction of arguments on the basis of evidence as well as their syntheses, 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 favour preferred arguments in a conflict.

The input to a system based on our framework is a table of evidence comparing pairs of treatments. Each row in the table gives the pair of treatments, the kind of comparison (e.g. randomized clinical trial, meta-analysis, or network analysis), the outcome indicator (e.g. disease-free survival, or overall survival), the outcome, the statistical significance, etc. The output of the system is an overall comparison of a pair treatments τ_1 and τ_2, saying whether τ_1 is superior to τ_2, or τ_1 is equivalent to τ_2, or τ_1 is inferior to τ_2. The output is justified by the arguments and counterarguments used to reach this conclusion.

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

We proceed as follows: (Section II) We review an abstract model of argumentation that we will incorporate in our general framework; (Section III) We discuss how we can represent evidence in a tabular format; (Section IV) We show how we can represent arguments based on the available evidence; (Section V) We show how we can compare arguments based on the benefits promised by the evidence and the veracity of

that evidence; (Section VI) We show how we can use the framework to aggregate evidence to make recommendations; (Section VII) We present a case study with evidence taken from 21 meta-analyses concerning 5 treatment options for raised intraocular pressure (raised IOP), and we compare the results we obtain with those presented in the NICE Guideline for Glaucoma; (Section VIII) We conclude with a discussion of the proposal in this paper, and how it relates to the literature.

II. Abstract argumentation

In this section, we review the proposal for abstract argumentation by Dung [4]. The simplest way to formalize a collection of arguments consists of just naming arguments (so, in a sense, treating them as atomic) and merely representing the fact that an argument is challenged by another (and so not indicating what the nature of the challenge is). In other words, a collection of arguments can be formalized as a directed binary graph.

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

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

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

$$A_1 \leftrightarrows A_2 \leftarrow A_3$$

Arguments can form coalitions to attack other arguments and to defend their members from attack as follows.

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

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

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

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

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

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

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

The notion of admissible sets of arguments is the minimum requirement for a set of arguments to be accepted. The following provide further restrictions.

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

1) Γ *is a* **complete extension** *iff* $\Gamma = \mathsf{Defended}(\Gamma)$
2) Γ *is a* **grounded extension** *iff it is the minimal (w.r.t. set inclusion) complete extension.*
3) Γ *is a* **preferred extension** *iff it is a maximal (w.r.t. set inclusion) complete extension.*

Example 2. *Continuing Example 1: The conflict free sets are* $\{\}$*,* $\{A_1\}$*,* $\{A_2\}$*,* $\{A_3\}$*, and* $\{A_1, A_3\}$*; The admissible sets are* $\{\}$*,* $\{A_1\}$*,* $\{A_3\}$*, and* $\{A_1, A_3\}$*; And the only complete set is* $\{A_1, A_3\}$*, and so this set is grounded and preferred.*

Example 3. *Consider the situation where we have just two arguments* A_4 *and* A_5 *that attack each other. We classify each subset of arguments as follows: The conflict free sets are* $\{\}$*,* $\{A_4\}$*, and* $\{A_5\}$*, and these are the admissible and complete sets; and The preferred sets are* $\{A_4\}$*, and* $\{A_5\}$*. However, there is only one grounded set which is* $\{\}$*.*

As can be seen from the examples, the grounded extension provides a skeptical view on which arguments can be accepted, whereas each preferred extension takes a credulous view on which arguments can be accepted.

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

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

III. Representing evidence

The types of evidence we consider in this paper are randomized clinical trial (RCT), meta-analyses (MA), network

TABLE I
FOUR RESULTS OBTAIN FROM THE NICE HYPERTENSION GUIDELINE (GC34, PAGES 36-37) CONCERNING ANGIOTENSIN-CONVERTING INHIBITORS (ACE) AND CALCIUM CHANNEL BLOCKERS (CCB)

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

analyses (NA), and cohort study (CS). Our focus will be on 2-arm superiority trials, i.e., clinical trials whose purpose is to determine whether, given two treatments, one is superior to the other (strictly speaking, such a trial tries to disprove the hypothesis that the two treatments are identical). This is an extremely common trial design.

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

- The **left** and **right** attributes signify the treatments compared in each item of evidence. In the table, these are angiotensin-converting inhibitors (ACE) for the left arm and calcium channel blockers (CCB) for the right arm.
- The **outcome indicator** is the specification of the particular outcome that is being considered when comparing the two treatments. In the table, in each row, it is the proportion of patients who have the event or condition (i.e. "mortality", "stroke", "heart failure" or "diabetes") within the period of the trial.
- The **value** of the outcome is the value obtained by the method applied to the outcome indicator. So for the first row, it is the proportion of patients who died during the trial taking ACE divided by the proportion of patients who died during the trial taking CCB.
- The **net outcome**, abbreviated by the column name Net, is a binary relation, denoted $>$ (superior), $\sim$ (equal), and $<$ (inferior), over the two treatments that is determined from the value of the outcome and an evaluation of whether the outcome indicator is desirable or undesirable for the patient class. For the first row, mortality is undesirable, and so a risk ratio value less than 1 means that the left arm is superior to the right arm, a risk ratio value equal to 1 means that the left arm is equal to the right arm, and risk ratio value greater than 1 means that the left arm is inferior to the right arm.
- The **statistically significant** attribute, abbreviated by the column name Sig, indicates whether the value is statistically significant. If the entry is "yes", then it 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 **evidence type**, abbreviated by the column name Type, specifies the type of study undertaken, e.g. randomized clinical trial (RCT), cohort study (CS), meta-analysis (MA), network analysis (NA), etc. It is an indicator of the quality of the evidence.

The set of attributes we have discussed here is only indicative. Often other attributes are useful for assessing and aggregating evidence (e.g. the number of patients involved in each trial, the geographical location for each trial, the drop-out rate for the trial, the methods of randomization for ensuring that patients and clinician do not know which arm a patient is in, etc). For a general introduction to the nature of clinical trials, and a discussion of a wider range of attributes, see [5].

The patient class is an important attribute that can be captured about an item of evidence. In the above table, the patient class is people with "persistent raised blood pressure of 160/100 mmHg or more". In our previous work, we showed how the patient class may involve a conjunction and/or disjunction of terms from a medical ontology and description logics can be used to provide inferencing (see [6]). Similarly, treatments presented in the left arm and right arm can be composed for a conjunction and/or disjunction of terms from an ontology. Again, medical ontologies cater for this by providing categories and relationships on treatments, substances used, and other characteristics. See [7], [3] for proposals for using a medical ontology in argumentation about clinical trials.

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

IV. REPRESENTING ARGUMENTS

Here we present a general framework for evidence aggregation that involves constructing and comparing arguments from items of evidence where the evidence involves multiple outcome indicators.

We start with a set of evidence EVIDENCE = $\{e_1, .., e_n\}$. Each item in EVIDENCE is a result from an RCT, an MA, a CS, or an NA, represented as a row in a table of evidence (as described in the previous section).

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

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

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

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

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

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

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

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

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

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

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

$$\langle \{e_3\}, \text{ACE} > \text{CCB} \rangle \qquad \langle \{e_1\}, \text{ACE} < \text{CCB} \rangle$$
$$\langle \{e_4\}, \text{ACE} > \text{CCB} \rangle \qquad \langle \{e_2\}, \text{ACE} < \text{CCB} \rangle$$
$$\langle \{e_3, e_4\}, \text{ACE} > \text{CCB} \rangle \qquad \langle \{e_1, e_2\}, \text{ACE} < \text{CCB} \rangle$$

In the example, we see intuitively that the arguments with differing claims conflict. We capture this relationship with the following definition. Note that this definition is symmetric, i.e., if A_i conflicts with A_j, then A_j conflicts with A_i.

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

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

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 $\tau_1 > \tau_2$, those arguments with claim $\tau_1 \sim \tau_2$, and those arguments with claim $\tau_1 < \tau_2$.

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

	Left	Right	Outcome indicator	Value	Net	Sig	Type
e_{71}	τ_1	τ_2	mortality	0.80	$>$	yes	RCT
e_{72}	τ_1	τ_2	palpitations	1.15	$<$	yes	NA

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

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

We use the preference rules chosen by the user in breaking the symmetry present in the conflict relation, and capture the attack relation as follows.

Definition 10. *For any pair of arguments* A_i *and* A_j, A_i **attacks** A_j *iff* A_i *conflicts with* A_j *and* A_i *is preferred to* A_j *and it is not the case that* A_j *is preferred to* A_i.

The motivation here is that if A_i and A_j conflict with each other and A_i is preferred to A_j then A_j's conflict with A_i is cancelled. However, this wording leads to problems when A_i is preferred to A_j according to a preference rule and A_j is preferred to A_i according to a preference rule. In this case, cancelling both attacks will give the misleading impression that A_i and A_j 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 11. *Given a pair of treatments* τ_1 *and* τ_2, *and a set* EVIDENCE *concerning these treatments, an* **argument graph** *is a graph where the set of nodes is the set of arguments formed by Definition 7 and the set of arcs is the attacks relation given by Definition 10.*

We leave the formalization of specific preference rules until the next section. In the meantime, we illustrate the use of an informally defined preference rule to get the following argument graph by applying the preference rule to the arguments in our running example.

Example 6. *Continuing Example 5, suppose we prefer the argument with statistically significant evidence showing superiority for the outcome indicator of "mortality" over other*

arguments. As a result, we get the following argument graph.

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

We can directly use the dialectical semantics given by Dung [4] (i.e. Definition 5) to decide extensions of argument graphs. Here, there is one grounded and preferred extension and it contains just the argument $\langle\{\{e_{71}\}, \tau_1 > \tau_2\rangle$.

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

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

V. Comparing arguments

In this section, we introduce the ideas behind comparing arguments based on the benefits promised by the evidence and on the veracity of that evidence. We start with the table given in Example 5. Here, we see that e_{71} promises that τ_1 is superior to τ_2 because the relative risk of mortality is 0.8, whereas e_{72} promises that τ_1 is inferior to τ_2 because relative risk of palpitation is 1.15. The first is a substantial benefit for the left arm whereas the second is a modest benefit for the right arm. So to aggregate the evidence, we require a method to take account of these relative benefits.

The way we can deal with this is by adopting a **benefit function** B from sets of evidence into the reals (i.e. $B : \wp(\text{EVIDENCE}) \mapsto \mathbb{R}$) based on the outcome indicators and values appearing in the evidence. This function gives a measure of the evidence based on how useful the outcomes (as promised by the evidence) are for the patient. We can think of a benefit function in monetary units, and so it could indicate how much a patient is prepared to pay for the combination of benefits promised by the evidence. For instance, we could adopt a benefit function for this example where $B(\{e_{71}\}) = 100$ and $B(\{e_{72}\}) = 1$ which reflects our discussion of the benefits above.

A benefit function is defined in terms of a **utility function** over attributes appearing in the evidence table. For this paper, we assume it is in terms of the outcome indicator and value attributes. If INDICATORS is the domain for the outcome indicator attribute and REALS is the domain for the value attribute, then we assume that we can define a utility function over sets of the tuples from INDICATORS $\times$ REALS or at least over those relevant for the arguments being considered. So, for the above example, let the utility function be U, defined as follows.

$$U(\{(\text{mortality}, 0.80)\}) = 100$$
$$U(\{(\text{palpitations}, 1.15)\}) = -1$$

We define the benefit function using the utility function as follows.

Definition 12. *Let X^* be the set of the tuples from* INDICATORS $\times$ REALS *appearing in X.*

1) *For $X \subseteq$* SUPERIOR*, $B(X)$ is $U(X^*)$*
2) *For $X \subseteq$* EQUITABLE*, $B(X)$ is 0*
3) *For $X \subseteq$* INFERIOR*, $B(X)$ is $-U(X^*)$*

Hence, for $X = \{e_{71}\}$, we have $X^* = \{(\text{mortality}, 0.80)\}$, and for $X = \{e_{72}\}$, we have $X^* = \{(\text{palpitations}, 1.15)\}$. In this way, for this example, the benefit function above (i.e. $B(\{e_{71}\}) = 100$ and $B(\{e_{72}\}) = 1$) is obtained by the utility function above.

So utility is the money the patient would pay in order to get the advantages of τ_1 instead of τ_2. We assume that the utility function conforms to the usual axioms of utility theory as regard preferences over outcomes. For instance, we assume that all sets of benefits can be ranked (i.e. all subsets of INDICATORS $\times$ REALS can be assigned by a utility), and we assume that transitivity holds so that if an outcome indicator o_1 with value v_1, denoted (o_1, v_1) is preferred to (o_2, v_2), and (o_2, v_2) is preferred to (o_3, v_3), then (o_1, v_1) is preferred to (o_3, v_3). We also assume that the utility function is monotonic in the membership of the subset of INDICATORS $\times$ REALS when there are not multiple occurrences of the same outcome indicator (i.e. when no outcome indicator occurs in more than one tuple in a set).

Example 7. *Consider the following evidence table containing fictional evidence comparing use of the contraceptive pill (CP) with no contraception (NC).*

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

For this table, a user of contraceptive pills may be trading the benefit of a substantial reduction in risk of pregnancy against a small increased risk of breast cancer and thrombosis. Though, there is also small positive effect for contraceptive users who get a reduced risk of ovarian cancer. We can use the following utility function, where oc *is ovarian cancer,* preg *is pregnancy,* bc *is breast cancer, and* th *is thrombosis.*

$$U(\{(\text{oc}, 0.99)\}) = 2$$
$$U(\{(\text{preg}, 0.05)\}) = 20$$
$$U(\{(\text{oc}, 0.99), (\text{preg}, 0.05)\}) = 22$$
$$U(\{(\text{bc}, 1.04)\}) = -2$$
$$U(\{(\text{th}, 1.02)\}) = -1$$
$$U(\{(\text{bc}, 1.04), (\text{th}, 1.02)\}) = -3$$

Hence, we get the benefit function as follows.

$$B(\{e_{81}\}) = 2 \qquad B(\{e_{82}\}) = 2$$
$$B(\{e_{84}\}) = 1 \qquad B(\{e_{83}\}) = 20$$
$$B(\{e_{81}, e_{84}\}) = 3 \quad B(\{e_{82}, e_{83}\}) = 22$$

Normally, there is some doubt about the veracity of the evidence, and this should be used to qualify the use of the

benefit function. For this, we assume a probability function V, which we call the **veracity function**, over sets of evidence (i.e. $V : \wp(\text{EVIDENCE}) \mapsto [0, 1]$) which captures the belief the user has in the benefits reported in the evidence being true given the quality of the evidence. Note, the empty set always has veracity 1.

Example 8. *Returning to Table I, we may choose to take a very simple approach capturing a difference in veracity between items of evidence that are statistically significant and items of evidence that are not statistically significant. So here, only e_1 is not statistically significant.*

$$V(\{e_3\}) = 0.7 \qquad V(\{e_1\}) = 0.1$$
$$V(\{e_4\}) = 0.7 \qquad V(\{e_2\}) = 0.7$$
$$V(\{e_3, e_4\}) = 0.65 \quad V(\{e_1, e_2\}) = 0.1$$

We have $V(\{e_3, e_4\})$ lower than either $V(\{e_3\})$ or $V(\{e_4\})$ since the probability that both items of evidence turn out to be correct is lower than either individually even if they are both statistically significant.

Example 9. *Consider the following fictional evidence table concerning treatments for breast cancer, tamoxifen (TAM) and methotrexate with flouracil and tamoxifen (MFT).*

	Left	Right	Outcome indicator	Value	Net	Sig	Type
e_{91}	MFT	Tam	mortality	0.89	$>$	yes	RCT
e_{92}	MFT	Tam	mortality	0.87	$>$	yes	MA
e_{93}	MFT	Tam	mortality	0.91	$>$	yes	RCT

Each item of evidence confirms the benefits promised by the other items of evidence. So we would expect that the veracity rises monotonically with respect to set membership for the evidence used as follows.

$$V(\{e_{91}\}) = V(\{e_{92}\}) = V(\{e_{93}\}) = 0.70$$
$$V(\{e_{91}, e_{92}\}) = V(\{e_{91}, e_{93}\}) = V(\{e_{92}, e_{93}\}) = 0.71$$
$$V(\{e_{91}, e_{92}, e_{93}\}) = 0.72$$

In general, the veracity function is a subjective probability, (i.e., we consider how we would bet whether a set of evidence would indeed turn out to be true with respect to the benefits promised) that conforms to various constraints such as expressed for the examples above. However, we could also consider learning the veracity function from past data concerning similar kinds of evidence (though not necessarily for the same treatments or disorders). For instance, if we look at the literature 5 years ago, identify the evidence used to make recommendations, and then reconsider that evidence in the light of today's literature to see whether it subsequently was confirmed or rebutted. This then gives us an objective probability. In this paper, we have restricted the set of attributes used in the evidence table, but other attributes are likely to be useful for this. For instance, whether the study was funded by a drug company, the size of the study, the randomization techniques used, the confidence interval, etc.

VI. Aggregating evidence

Now, we consider how we harness the ideas introduced so far for systematically aggregating evidence. We start by defining our preference rules which use the utility and veracity measures, and then show how we can use argumentation to aggregate the evidence.

Definition 13. *The* **benefit preference rules** *over arguments A_i and A_j are as follows where $\pi \in [0, 1]$ is a threshold for veracity, V is a veracity function, B is a benefit function, X_i is the support of A_i, and X_j is the support of A_j.*

1) *If $V(X_i) > \pi$ and $B(X_i) > B(X_j)$, then A_i is preferred to A_j.*
2) *If $V(X_i) > \pi$ and $V(X_j) \leq \pi$, then A_i is preferred to A_j.*

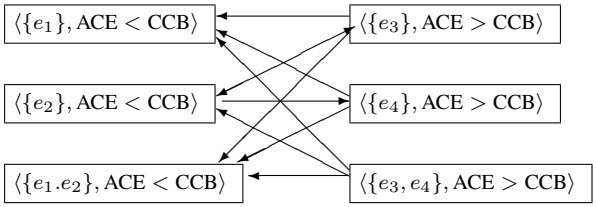
So for arguments based on evidence with a veracity above an acceptable threshold (i.e. $V(X) > \pi$ where X is the support of the argument and π is say 0.6), then we prefer the arguments with evidence with greater benefit.

With these preference rules, we see that when we construct all the arguments from the evidence, an argument with any empty support can only affect the outcome of the argumentation (i.e. it can only affect the overall aggregated result) if there is no argument with the same claim and a non-empty support. So to simplify our presentation, we will take advantage of this redundancy by not considering arguments with an empty support if there is an argument with the same claim and a non-empty support.

Example 10. *For the evidence in Table I, we use the following benefit and veracity functions.*

X	$B(X)$	$V(X)$
$\{e_1\}$	10	0.1
$\{e_2\}$	9	0.7
$\{e_1, e_2\}$	16	0.1
$\{e_3\}$	9	0.7
$\{e_4\}$	4	0.7
$\{e_3, e_4\}$	10	0.65

For the argument graph (for simplicity we have excluded the arguments with empty support) there is a grounded extension containing just the arguments with the claim ACE $>$ CCB.



We use the following criteria for interpreting an argument graph that has been generated from an evidence table, and thereby show how we obtain an aggregation of that evidence.

- If there is a non-empty grounded extension, and ϵ is the claim of the arguments in the extension (note, all arguments in a grounded or preferred extension will have the same claim), the result of the aggregation is ϵ.
- If there is an empty grounded extension, then there are multiple preferred extensions (say $E_1, ..., E_n$), and so the result of the aggregation is ϵ_1 or ... or ϵ_n where ϵ_1 is the claim of the arguments in E_1 and ... and ϵ_n is the claim of the arguments in E_n.

Note, if the result of the aggregation is a disjunction, then either there is insufficient evidence to determine which of the disjuncts holds (and this can be determined by the support for all the arguments being below the threshold π) or the utility of the evidence is the same for the claims.

VII. Case study

In this section, we report on a case study we undertook concerning treatments for raised intraocular pressure (raised IOP), which is raised pressure in the eye, where the evidence was obtained from the NICE Guideline for Glaucoma (available from www.nice.org.uk). The evidence is presented in Table II. Each item is an MA generated by the guideline authors as presented in the appendix of the guideline. The medications considered are no treatment (NT), beta-blocker (BB), prostaglandin analogue (PG), sympathomimetic (SY), and carbonic anhydrase inhibitor (CA). The Net column gives an interpretation of the value with respect to the type of outcome indicator: For the outcome indicator "change in IOP", if the value is negative, the left arm is superior, otherwise it is inferior. For the outcome indicator "acceptable IOP", which is a desirable outcome for the patient, if the value is greater than 1, the left arm is superior, otherwise it is inferior. For each of the remaining outcome indicators (i.e. for "respiratory problems", "cardiovascular problems", "allergy problems", "hyperaemia", "convert to COAG", "visual field progression", "IOP $>$ 35mmHg", and "drowsiness"), which are undesirable for the patient, if the value is less than 1, then the left arm is superior, otherwise it is inferior. Note, "hyperaemia" means redness of eyes, "convert to COAG" means the patient develops chronic open angle glaucoma, "visual field progression" means that there is damage to the retina and/or optic nerve resulting in loss of the visual field and "IOP $>$ 35mmHg" means that the intraocular pressure is above 35mmHg (which is very high).

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

For our study we classify some outcomes as major (namely, "visual field prog" when value $<$ 0.95, "change in IOP" when value $<$ -1, "acceptable IOP" when value $>$ 1.2, "convert to COAG" when value $<$ 0.95, and "IOP $>$ 35mmHg" when value $<$ 0.9) and the remaining outcomes we classify as minor. For a set of evidence X, we let x be the number of major outcomes appearing in X^*, we let y be the number of minor outcomes appearing in X^* where the value is less than 1, and we let z be the number of minor outcomes appearing in X^* where the value is greater than 1. We then defined the utility function $U(X^*)$ to be $(5x) + y - z$. For veracity, if all evidence in X is statistically significant, we let $V(X) = 0.8$, otherwise we let $V(X) = 0.2$. We illustrate them with some of the arguments we generated.

TABLE II
The evidence table for the case study. Each row is a meta-analysis from the NICE Glaucoma Guideline (CG85 Appendix pages 213-223) for the class of patients who have raised intraocular pressure (i.e. raised pressure in the eye) and are therefore at risk of glaucoma with resulting irreversible damage to the optic nerve and retina.

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

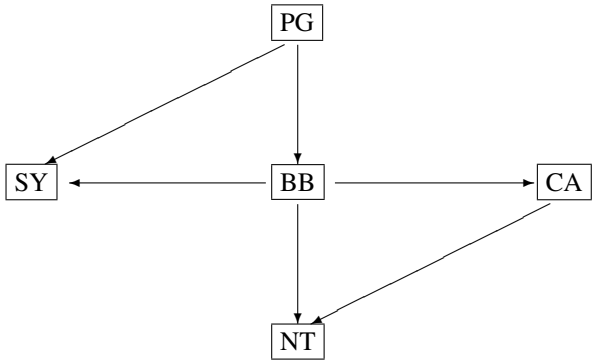


Fig. 1. Summary of the conclusions from the argumentation with the glaucoma case study where a directed arc from τ_1 to τ_2 denotes τ_1 is superior to τ_2 and an undirected arc from τ_1 to τ_2 denotes τ_2 is superior or equivalent or inferior to τ_2.

Argument	X	$B(X)$	$V(X)$
$\langle\{e_{05}, e_{06}, e_{07}\}, \text{PG} > \text{BB}\rangle$	$\{e_{05}, e_{06}, e_{07}\}$	*10*	*0.8*
$\langle\{e_{09}, e_{10}\}, \text{PG} < \text{BB}\rangle$	$\{e_{09}, e_{10}\}$	*2*	*0.2*
$\langle\{e_{10}\}, \text{PG} < \text{BB}\rangle$	$\{e_{10}\}$	*1*	*0.8*
$\langle\{e_{14}, e_{15}, e_{16}\}, \text{CA} > \text{NT}\rangle$	$\{e_{14}, e_{15}, e_{16}\}$	*15*	*0.2*
$\langle\{e_{18}, e_{19}\}, \text{SY} > \text{BB}\rangle$	$\{e_{18}, e_{19}\}$	*10*	*0.2*

From the argumentation, we got the aggregations of the evidence concerning the treatment options which we summarise in Figure 1. By contrast, the NICE guideline just selects two

options, namely prostaglandin analogues and beta-blockers, as superior to the rest with prostaglandin analogues being the preferred treatment for the patients with thinner central corneal thickness, a group of patients who are more at risk of developing glaucoma as a result of having raised IOP. We therefore claim that our qualitative framework provides a result that indicates the adequacy of our framework for capturing the underlying argumentation inherent in the guideline.

VIII. Conclusion

The problem of conflicting information is a general issue in handling knowledge and it arises in virtually all real-world domains. To address this, computational models of argumentation which aim to reflect how human argumentation uses conflicting information to construct and analyze arguments, are being developed (for reviews see [8], [9]).

In this paper, we have drawn on argumentation techniques (in particular influenced by assumption-based argumentation [10]) to provide a general framework for taking evidence involving multiple outcome indicators and aggregate it in terms of arguments. In this framework, we instantiate abstract argument graphs with arguments generated by inference rules applied to the evidence, and attacks relationships obtained via the preference rules. For any application of our framework, a specific set of inference rules and preference rules needs to be given. Given an evidence table, the algorithms for generating arguments and the argument graph are simple, and there are existing argumentation engines [11], [12] that can be used to calculate the different extensions.

As well as presenting the general framework, we have presented a new specific version of it with a simple and complete set of inference and preference rules, and we have evaluated this specific version with respect to a case study with evidence taken from 21 meta-analyses concerning 5 treatment options for raised intraocular pressure (raised IOP), and we have shown the results we obtained corresponded closely with those presented in the NICE Guideline for Glaucoma.

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

Little work exists that aims to address the problem in focus in this paper. Medical informatics and bioinformatics research does not address the reasoning aspects inherent in the analysis of evidence of primary nature, especially from clinical trials. Previous interesting work ([13], [14], [15], [16] and others) exists that uses argumentation as a tool in medical decision support, but as such, assumes the existence of a hand-crafted knowledgebase.

In future work, we aim to developed generic utility functions based on an ontology of outcome indicators, and a generic veracity function based on calculus of evidence quality. We also aim to develop theoretical tools for effectively and efficiently acquiring and representing functions based on lattice theory and/or logical constraints.

References

[1] N. Gorogiannis, A. Hunter, V. Patkar, and M. Williams, "Argumentation about treatment efficacy," in *Knowledge Representation for Healthcare (KR4HC)*, ser. LNCS, vol. 5943. Springer, 2010, pp. 169–179.

[2] A. Hunter and M. Williams, "Qualitative evidence aggregation using argumentation," in *UCL Dept of Computer Science Technical Report*, 2010.

[3] N. Gorogiannis, A. Hunter, and M. Williams, "An argument-based approach to reasoning with clinical knowledge," *International Journal of Approximate Reasoning*, vol. 51, no. 1, pp. 1 – 22, 2009.

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

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

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

[7] 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)*, vol. 2. IEEE Computer Society Press, 2007, pp. 254–261.

[8] T. Bench-Capon and P. Dunne, "Argumentation in artificial intelligence," *Artificial Intelligence*, vol. 171, no. 10-15, pp. 619–641, 2007.

[9] Ph.. Besnard and A. Hunter, *Elements of Argumentation*. MIT Press, 2008.

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

[11] M. South, G. Vreeswijk, and J. Fox, "Dungine: A java dung reasoner," in *Computational Models of Argument (COMMA'08)*. IOS Press, 2008, pp. 360–368.

[12] U. Egly, S. Gaggl, and S. Woltran, "Aspartix: Implementing argumentation frameworks using answer-set programming," in *Proceedings of the Twenty-Fourth International Conference on Logic Programming (ICLP'08),*, ser. LNCS, vol. 5366. Springer, 2008, pp. 734–738.

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

[14] V. Patkar, C. Hurt, R. Steele, S. Love, A. Purushotham, M. Williams, R. Thomson, and J. Fox, "Evidence-based guidelines and decision support services: A discussion and evaluation in triple assessment of suspected breast cancer," *British Journal of Cancer*, vol. 95, no. 11, pp. 1490–1496, 2006.

[15] P. Tolchinsky, U. Cortés, S. Modgil, and F. C. A. López-Navidad, "Increasing human-organ transplant availability:argumentation-based agent deliberation," *IEEE Intelligent Systems*, vol. 21, no. 6, pp. 30–37, 2006.

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