Argumentation Schemes for Clinical Interventions
Towards an Evidence-Aggregation System for Medical Recommendations

Olivia Sanchez-Graillet and Philipp Cimiano

Cognitive Interaction Technology Cluster of Excellence (CITEC)
Bielefeld University, Germany
Email: olivia.sanchez@uni-bielefeld.de

***Abstract*—The paradigm of evidence-based medicine requires that medical decisions are taken based on available, verified and high quality evidence. Such evidence has to be obtained from multiple relevant studies, considering their potential biases and shortcomings. Rationalizing and aggregating evidence from multiple studies is key to evidence-based decision making. Towards a system that is able to aggregate and summarize the evidence available in multiple studies, we have defined two argument schemes, which respectively provide reasons as to why a certain therapy may be regarded superior to another in terms of efficacy and safety. The argument schemes can be automatically instantiated via the semantic query language SPARQL from a knowledge base in which clinical studies have been formalized according to our own clinical trial ontology. The argument schemes are meant to be part of the framework of a configurable system that generates clinical recommendations by aggregating and summarizing the evidence from different clinical studies. We demonstrate the instantiation of the argument schemes in a study case on glaucoma and show that they are able to capture the reasoning behind determining such intervention superiority.**

Keywords–Argumentation in Medicine; Argumentation Schemes; Evidence-based Medicine; Summarization of Clinical Evidence

I. INTRODUCTION

Medical decisions are taken based on available, verified and high quality evidence. Randomized Clinical Trials (RCTs) are considered as the gold-standard for clinical research [1]. There are thousands of such controlled clinical studies and related publications available in open-access databases, such as PubMed [2]. Since clinicians are interested not only in obtaining effective intervention outcomes, but also in that the outcomes rely on high quality evidence, it is necessary to collect evidence from multiple relevant studies, whilst considering the possible presence of bias and the shortcomings of those studies [3]. The summarization and comparison of the aggregated information is normally done in the form of systematic reviews and meta-analyses.

Rationalizing and aggregating evidence from multiple clinical trials are crucial tasks for evidence-based decision making [4]–[7]. Criteria for grading the level of evidence have been already developed (e.g., The Grading of Recommendations Assessment, Development and Evaluation GRADE [8], [9]). However, applying such criteria to aggregate and summarize the evidence available in the vast number of relevant publications, requires an extensive manual effort. As a part of a system for generating medical recommendations based on clinical trial evidence, in this paper we present argument schemes that provide reasons as to why a certain therapy is regarded as superior to another – in terms of efficacy and safety – by aggregating the evidence found in multiple studies. The studies are formalized in a knowledge base structured according to our own Clinical Trial Ontology (C-TrO) [10]. The arguments are automatically instantiated from the knowledge base by extracting evidence via the SPARQL query language [11]. The argument schemes can be, in principle, used to rationalize the evidence for therapies of any health condition or disease. In this paper, we show the use of the schemes through a study case on glaucoma.

The remainder of this paper is structured as follows. Section II describes the proposed argumentation schemes and Section III the knowledge base defined on basis of C-TrO. The exemplary use of the schemes on glaucoma is presented in Section IV. Our conclusions and plans for future work are given in Section V.

II. ARGUMENT SCHEMES FOR AGGREGATING AND SUMMARIZING CLINICAL TRIAL EVIDENCE

In the definition of the argument schemes presented in this paper, we followed the model proposed by Walton et al. [12], in which argument schemes are seen as defeasible inference patterns that make explicit how a certain conclusion follows from a set of premises or assumptions. Thus, such arguments consist of a set of premises (assumptions), a conclusion, and critical questions that could invalidate the conclusion if satisfied. In our case, the conclusion of an argument corresponds to the claim that a certain therapy is superior to another given the evidence available in the form of multiple studies. Such type of arguments represent tools to systematically reason about the available evidence and thus support decision making. The arguments are defeasible reasoning patterns in the sense that the conclusion that one therapy is superior to another may be challenged and even invalidated by additional information (e.g., biases in the publications, lack of significance of size of effect, etc.).

Through the empirical analysis of published clinical trials of different types, and meta-analyses and systematic reviews on different health conditions, we identified the basic forms of argument schemes for inferring superiority of interventions in terms of efficacy and safety. The schemes can be applied in different configurations. For example, considering a given type of population, patient preconditions, country, etc. The different scenarios can be formed from the information contained in C-TrO. These schemes allow to structure available evidence as a basis to reason about the superiority of a certain treatment over another one. Each scheme states a major premise that is

 ISBN: 978-1-61208-759-7

Major premise: For people who suffer a given disease/health-disorder, it is desirable that a certain outcome indicator (or measurement) related to that disease/health-disorder changes, either increasing or decreasing.
Minor premise: It has been shown in a bigger number of clinical trials that $T1$ changes (either increasing or decreasing) a given disease/health-disorder indicator from the baseline in terms of an aggregation method in greater magnitude than $T2$.
Conclusion: $T1$ is a more effective medication treatment compared to $T2$ for changing the given disease/health-disorder indicator in the desired direction.

Critical Questions:
CQ1: Is the change (either increasing or decreasing) of the given disease/health-disorder indicator statistically significant (p-value)?
CQ2: Is the size of effect of $T1$ bigger than the one of $T2$?
CQ3: Are $T1$ and $T2$ applied to a similar number of patients across the different studies?

Figure 1. Scheme for superiority in terms of efficacy.

assumed to hold independently of the current level of evidence. Whilst the minor premise summarizes the current level of evidence as supporting the conclusion. The critical questions proposed pretend to challenge the validity of the conclusion based on the available information. The proposed argument schemes are described in what follows.

A. Argument Scheme for Superiority Based on Efficacy

In this argument scheme, the major premise expresses the general objective of the primary outcome of the intervention in question, and the minor premise considers the magnitude of the differences between the intervention results. The first critical question considers the statistical significance of the results; the second one refers to the size of the population that receives the intervention, which is important to consider since the p-value may vary according to this size; and the third one is about the absolute size of effect, i.e., the magnitude of the difference between groups. An intervention in which both size of effect and statistical significance are reported, tends to be more convincing than one in which only the size of effect is mentioned.

Figure 1 presents the corresponding argument scheme and its critical questions, where $T1$ and $T2$ are different drug treatments. In this scheme, it is implied that when there is a smaller number of clinical trials in which the outcome indicator in $T1$ changes in a bigger magnitude than in $T2$, the conclusion would be that $T1$ is less effective than $T2$. If the number of clinical trials is the same (or very similar), then $T1$ and $T2$ would be considered as being equally effective.

B. Argument for Superiority Based on Safety

In this argument scheme, the major premise expresses the general objective of the intervention outcome relative to safety. The minor premise considers the magnitude of the adverse effect that can be expressed in different ways, such as:

- Absolute magnitude, which refers to the number of people affected by a given adverse effect (e.g., "The most significant side effect of latanoprost was increased pigmentation of the iris which was observed in 15 patients").

Major premise: For people who suffer a given disease/health-disorder and who are under a medication treatment, it is desirable not to suffer any adverse effect.
Minor premise: It has been shown in a number of clinical trials that the administration of $T1$ leads to less incidence of adverse effects compared to the administration of $T2$.
Conclusion: Therefore, $T1$ is superior to $T2$ in terms of its safety profile, leading to less cases of the adverse effects.

Critical Questions:
CQ1: Are the adverse effects statistical significant?
CQ2: Are the size of effect of the adverse effects bigger for $T2$ than for $T1$?

Figure 2. Scheme for superiority in terms of safety.

- Relative magnitude, which refers to the percentage of people affected by a given adverse effect (e.g., "The most frequent drug adverse event was reported in 0.5% patients") .
- Uncertain magnitudes that denote uncertainty about the presence of an adverse effect (e.g., "The presence of an adverse effect was suspected").
- Modal words that indicate the degree of affection (e.g., "slightly affected") or expressions like "bigger degree (or amount)".

The corresponding argument scheme and its critical questions are presented in Figure 2. The critical questions are related to the statistical significance of the observed adverse effects.

C. Critical Questions Relative to the Quality of the Evidence

The following are the critical questions that apply to both schemes.

CQ3: How reliable is the evidence from these studies?

- ***CQ3.1*** Is there a risk of bias?
- ***CQ3.2*** Is the study randomized?
- ***CQ3.3*** Is the study blind?
- ***CQ3.4*** Is the study multi-center?
- ***CQ3.5*** Is the study intention-to-treat?

These critical questions are based on the following reasons:

- Intention-to-treat (ITT) studies are more realistic and unbiased than pre-protocol studies because they include all the patients in the results, while pre-protocol studies exclude patients who deviated from the protocol.
- Multi-center studies are more inclusive than single-center studies.
- Blind (or double-blind) studies are more objective than those of different type.
- There might be a risk of bias when a conflict of interest exists.

The context in which the argument schemes and critical questions are applied can be constrained by considering further

 ISBN: 978-1-61208-759-7

information. For example, the population's country, gender and age range. This evidence is available in the knowledge base, which is described in the next section.

III. The Knowledge Base and C-TrO

As part of the system for generating medical recommendations, we have developed the C-TrO ontology [10] that describes clinical studies with the adequate level of formalization and granularity for instantiating the proposed argument schemes and for defining different contexts of interest in which the argument schemes could be used.

The knowledge base follows the C-TrO structure and is described in Resource Description Framework (RDF) triples [13]. Below is an abbreviated example of the RDF triples corresponding to a clinical trial, one of its arms and one of the arm's interventions.

```
:CT_3 rdf:type :ClinicalTrial ;
   :hasObjectiveDescription "Latanoprost, a ..."
   :hasArm Arm_31, Arm_32 ;
   :hasPopulation CT3_Population ;
   :hasCTDesign DoubleBlind, Randomized .
:Arm_31 rdf:type ctro:Arm ;
   :hasNumberPatients 134 ;
   :hasIntervention :CT3_Intervention1 .
   :hasPrimaryOutcome :CT3_A1_OC1 ;
:CT3_Intervention1 rdf:type ctro:Intervention ;
   :hasFrequency "Once_at_evening";
   :hasInterval "Daily" ;
   :hasDuration "3 months";
   :hasMedication :CT3_I1_M1 .
```

IV. Study Case on Glaucoma

Glaucoma is a disease that damages the optic nerve and can lead to permanent visual loss. The damage of the optic nerve usually occurs when the Internal Ocular Pressure (IOP) increases. Therefore, the reduction of IOP is a desired outcome in an intervention for glaucoma.

For our study case, we formalized the clinical trials included in the meta-analysis carried out by Zhang et al. [14] which compares the efficacy and safety of the drugs timolol and latanoprost. The quality of the evidence was assessed by considering the design characteristics of the clinical trials such as masking, randomization, etc. The main outcome indicators studied were the percentage of IOP reduction for efficacy, and the relative risk for side effects (e.g., hyperaemia, conjunctivitis, etc.). The meta-analysis suggested that latanoprost was more effective than timolol in lowering IOP.

Figure 3 shows the instantiation of the argument scheme for the treatment superiority in terms of efficacy. This instantiation is carried out by executing a SPARQL query that represents the argument scheme with the information referent to glaucoma over the knowledge base. This query retrieves the appropriate evidence.

In this scheme, the major premise states the main objective of medical interventions for glaucoma, which refers to the greatest reduction of the diurnal IOP mean from baseline. The minor premise mentions that in the eleven studies compared, it was found that latanoprost interventions were more efficacious in reducing the diurnal IOP mean than timolol interventions. The table below this premise contains the evidence that supports this assertion (only five out of eleven studies are shown). Each row is a pairwise comparison of the interventions of each clinical trial. The first column contains the trial identifiers, the second column contains the publications that describe the clinical trials, and the remaining columns display the reduction of IOP mean (in $mmHG$ units) by the latanoprost and the timolol treatments, respectively.

Given this evidence, the conclusion in this argument scheme states that the latanoprost treatments are more effective than the timolol treatments under the specified conditions. The critical question is related to the statistical significance (p-value) of the IOP mean reduction from the baseline. Since the p-values of the interventions of some clinical trials were not reported, the conclusion of the argument can be weakened.

Major premise: For people who suffer glaucoma it is desirable that the *diurnal mean IOP* is reduced.
Minor premise: It has been shown in eleven clinical trials that *latanoprost* treatments reduced the *diurnal mean IOP* from baseline in a greater magnitude than the *timolol* treatments.

Evidence (Mean IOP Reduction (mmHg))			
CT_Id	**Reference**	**Latanoprost**	**Timolol**
CT_1	Alm A et al,1995	7.8	6.7
CT_2	Aquino MV et al.,1999	11.1	9.1
CT_3	Camras CB et al.,1996	6.7	4.9
CT_4	Diestelhorst M et al.,1998	4.9	2.1
CT_5	Mastropasqua L et al,1999	4.8	4.6

Conclusion: *latanoprost* treatment is a more effective medication treatment compared to *timolol* treatment for reducing the diurnal mean IOP.

CQ1: Is the reduction of the diurnal mean IOP statistically significant? Only some p-values were reported.

Figure 3. Instantiation of the scheme for efficacy.

The argument schema in terms of safety for glaucoma is instantiated as shown in Figure 4.

Major premise: For people who suffer glaucoma and who are under a medication treatment it is desirable not to suffer any adverse effect.
Minor premise: It has been shown in eleven clinical trials that the administration of the *timolol* treatment leads to less incidence of *Conjunctival_hyperemia* than the *latanoprost* treatment.

Latanoprost		Timolol	
Adverse Effect	**No.**	**Adverse Effect**	**No.**
ConjunctivalHyperemia	7	ConjunctivalHyperemia	2
IncreasedPigmentation	2	ReducedHeartRate	2
IrisPigmentationChange	1	ReducedBloodPreasure	2
		ChangeBloodVelocity	1
		Smarting	1
		IrisPigmentationChange	1

Conclusion: The *timolol* treatment is superior to the *latanoprost* treatment in terms of its safety profile, leading to less cases of the adverse effect *ConjunctivalHyperemia*.

Figure 4. Instantiation of the scheme for safety.

The major premise states that the ideal outcome of the medical interventions is that they do not cause any adverse effect. The minor premise states that across the eleven studies it was found that the presence of *conjunctival hyperemia* occurred more times in the latanoprost interventions than in

 ISBN: 978-1-61208-759-7

the timolol interventions. The table below this premise contains the evidence that supports that assertion (i.e. seven mentions in the case of latanoprost and only two mentions in the case of timolol).

Given this evidence, the conclusion in this scheme states that the timolol treatment is safer than the latanoprost treatment relative to conjunctival hyperemia. This schema can be applied to any other adverse effect of interest. In this study case, we have analyzed glaucoma and only the effect of two drugs. However, the scheme can consider any other disease and drugs that are contained in the knowledge base.

V. CONCLUSIONS

We have presented basic argument schemes for deciding the superiority of medical interventions in terms of efficacy and safety based on the aggregation of clinical trial evidence. From these schemes, more specific argument schemes could be derived. A study case on glaucoma was used as a proof-of-concept of our argument schemes, showing the feasibility of their use.

Other approaches that rely on argument schemes are those for identifying argument schemes in genetic research articles [15] and in letters for genetic counselling [16]. More recently, Mayer et al. [17] presented a method for recognizing evidence in the form of premises and conclusions in RCT abstracts, and Mayer et al. [18] developed a method for classifying the type of evidence found in clinical trials. Furthermore, several argumentation-based approaches for medical and health-care decision support have been proposed [19]–[23]. Hunter and Williams [20] developed an argument-based framework to aggregate clinical trial evidence, in which inductive arguments are generated from a set of evidence. Afterwards, the superiority of the interventions is determined according to given preference criteria across the generated arguments. In contrast to this approach, we have defined argumentation schemes supported by an ontology to rationalize the decision as to which intervention is superior, as well as to aggregate the corresponding level of evidence. Furthermore, our knowledge base is populated with evidence extracted from published clinical trials, while in the Hunter and William's method, the input evidence is taken from partially aggregated and synthesized information contained in medical guidelines.

As future work, we plan to validate our argument schemes via user studies. We want to extend our approach so that efficacy and safety are not considered as independent dimensions, but that they can be weighted against each other in the form of a meta-argument with an internal structure that resolves the trade-off depending on the relative weight given to safety over efficacy or vice versa. Finally, we are currently developing an information extraction system that automatically encodes the study results and clinical evidence in a knowledge base following the C-TrO ontology.

ACKNOWLEDGMENT

This work has been funded by the Deutsche Forschungsgemeinschaft (DFG) within the project Rationalizing Recommendations (RecomRatio), as part of the Priority Program "Robust Argumentation Machines (RATIO)" (SPP-1999).

REFERENCES

[1] A. Akobeng, "Understanding randomised controlled trials," Archives of Disease in Childhood, vol. 90, 2005, pp. 840–844.

[2] "PubMed", URL: https://www.ncbi.nlm.nih.gov/pubmed [Last access: August, 2019].

[3] D. L. Sackett, W. M. C. Rosenberg, J. A. M. Gray, R. B. Haynes, and W. S. Richardson, "Evidence based medicine: what it is and what it isn't," BMJ, vol. 312, no. 7023, 1996, pp. 71–72.

[4] G. Tsafnat, A. Dunn, P. Glasziou, and E. Coiera, "The automation of systematic reviews," BMJ, vol. 346, 2013, p. f139.

[5] A. M. OConnor, G. Tsafnat, S. B. Gilbert, K. A. Thayer, and M. S. Wolfe, "Moving toward the automation of the systematic review process: A summary of discussions at the second meeting of international collaboration for the automation of systematic reviews (ICASR)," Systematic Reviews, vol. 7, no. 1, 2018, p. 3.

[6] M. H. Murad and V. M. Montori, "Synthesizing evidence: Shifting the focus from individual studies to the body of evidence," JAMA, vol. 309, no. 21, 2013, pp. 2217–2218.

[7] H. Bastian, P. Glasziou, and I. Chalmers, "Seventy-five trials and eleven systematic reviews a day: How will we ever keep up?" PLoS Medicine, vol. 7, no. 9, 2010, p. e1000326.

[8] H. Balshem et al., "GRADE Guidelines: 3. Rating the quality of evidence," J Clin Epidemiol, vol. 64, no. 4, 2011, pp. 401–406.

[9] G. H. Guyatt et al., "What is quality of evidence and why is it important to clinicians?" BMJ, vol. 336, no. 7651, 2008, pp. 995–998.

[10] O. Sanchez-Graillet, P. Cimiano, C. Witte, and B. Ell, "C-TrO: An ontology for summarization and aggregation of the level of evidence in clinical trials," in The Joint Ontology Workshop (JOWO) 2019: Ontologies and Data in the Life Sciences (To appear).

[11] "The SPARQL Query Language", URL: https://www.w3.org/TR/rdf-sparql-query [Last access: August, 2019].

[12] D. Walton, C. Reed, and F. Macagno, Argumentation Schemes. Cambridge University Press, 2008.

[13] "The RDF Working Group", URL: https://www.w3.org/RDF/ [Last access: August, 2019].

[14] W. Y. Zhang, A. L. W. Po, H. S. Dua, and A. Azuara-Blanco, "Meta-analysis of randomised controlled trials comparing latanoprost with timolol in the treatment of patients with open angle glaucoma or ocular hypertension," Br J Ophthalmol., vol. 85, no. 8, 2001, pp. 983–990.

[15] N. Green, "Towards creation of a corpus for argumentation mining the biomedical genetics research literature," in Proc. of the 1st Workshop on Argumentation Mining, 2014, pp. 11–18.

[16] N. L. Green, "Annotating evidence-based argumentation in biomedical text," in IEEE Intl. Conf. on Bioinformatics and Biomedicine (BIBM), 2015, pp. 922–929.

[17] T. Mayer, E. Cabrio, M. Lippi, P. Torroni, and S. Villata, "Argument mining on clinical trials," in Proc. of COMMA 2018, vol. 305. IOS Press, 2018, pp. 137–148.

[18] T. Mayer, E. Cabrio, and S. Villata, "Evidence type classification in randomized controlled trials," in Proc. of the 5th Workshop on Argumentation Mining, 2018.

[19] N. Gorogiannis, A. Hunter, and M. Williams, "An argument-based approach to reasoning with clinical knowledge," Intl. J. of Approx. Reasoning, vol. 51, no. 1, 2009, pp. 1–22.

[20] A. Hunter and M. Williams, "Aggregating evidence about the positive and negative effects of treatments," Art. Int. in Medicine, vol. 56, no. 3, 2012, pp. 173–190.

[21] R. Craven, F. Toni, C. Cadar, A. Hadad, and M. Williams, "Efficient argumentation for medical decision-making," in Proc. of the 13th Intl. Conf. on Principles of Knowledge Representation and Reasoning (KR), 2012, pp. 598–602.

[22] L. Longo and L. Hederman, "Argumentation theory for decision support in health-care: a comparison with machine learning," in Proc. of the Intl. Conf. on Brain and Health Informatics, 2013, pp. 168–180.

[23] M. A. Grando, L. Moss, D. Sleeman, and J. Kinsella, "Argumentation-logic for creating and explaining medical hypotheses," Art. Int. in Medicine, vol. 58, no. 1, 2013, pp. 1–13.

 ISBN: 978-1-61208-759-7