Argumentation Devices in Reasoning About Health

Sally Jackson and Jodi Schneider
University of Illinois at Urbana-Champaign, sallyj@illinois.edu
University of Pittsburgh, jos188@pitt.edu

Abstract

Health controversies are infused with products of expert reasoning, often interpreted by non-experts. To understand these controversies, we must pay closer attention both to the field-dependent devices that characterize expert reasoning, and to how non-experts engage with experts' evidence and reasoning in their own argumentative practices. We describe two argumentation devices that have emerged in medical research and discuss the role of these devices within health controversies.

1 Introduction

Argumentation is a constantly evolving social practice, one that builds on thousands of years of human experience. The ubiquitous human practice of seeking advice from experts, for example, has very long historical roots, but it is also a basis for decision-making that is in constant flux as the grounds for expert opinion change. Expert fields do not just accumulate information; they also invent specialized ways of reasoning about information. Toulmin [1958] noted this fact and discussed at length the possibility that warrants (or backing for warrants) might justify the movement from data to claim only within particular fields. The Argument Interchange Format [Chesñevar *et al.*, 2006] acknowledges field dependence in argumentation by including context in the core model and assuming that context may include domain-specific argumentation rules that are direct counterparts of domain-independent schemes. Our goal in this paper is to explore field-dependent patterns of reasoning in health and medicine and to consider how these can be modeled.

Several examples drawn from a contemporary health controversy illustrate an important fact: As expert fields innovate in their own reasoning practices, arguments built by non-experts on the prior arguments of experts may take forms quite unlike the canonical form of argument from expert opinion. Each example mentions a conclusion drawn by an expert or group of experts, and at first glance, it would seem that each would pass all of the tests defined by standard lists of critical questions for the expert opinion scheme [Walton *et al.*, 2008, p. 15], including the "backup evidence question."

Example 1, Press Release from AutismSpeaks non-profit[1]

> In the largest-ever study of its kind, researchers again found that the measles-mumps-rubella (MMR) vaccine did not increase risk for autism spectrum disorder (ASD). This proved true even among children already considered at high risk for the disorder.
>
> In all, the researchers analyzed the health records of 95,727 children, including more than 15,000 children unvaccinated at age 2 and more than 8,000 still unvaccinated at age 5. Nearly 2,000 of these children were considered at risk for autism because they were born into families that already had a child with the disorder.
>
> The report appears today in *JAMA*, the Journal of the American Medical Association.

Example 2, The New York Times[2]

> According to Dr. Paul Offit, an infectious disease specialist at Children's Hospital of Philadelphia, young children readily handle the immune challenges of multiple vaccines. For example, studies have shown the five-in-one vaccine Pediarix against hepatitis B, polio, tetanus, diphtheria and pertussis is as safe and effective as giving each of these vaccines individually.

Example 3, The Guardian[3]

> The evidence of no link between MMR and autism is now extremely strong. In February 2012, the Cochrane Collaboration - which compiles gold-standard reviews of medical evidence - conducted a huge study into the safety of MMR. This mega-review brought together evidence from 54 difference(sic) scientific studies using a variety of methodologies and involving 14.7 million children from around the world.

These passages are typical of the appearance of expert knowledge in the public discussion of childhood vaccination. But the critical questions associated with the argument from expert opinion scheme will not provide the kind of searching evaluation that these examples require.

[1] https://www.autismspeaks.org/science/science-news/no-mmr-autism-link-large-study-vaccinated-vs-unvaccinated-kids

[2] http://well.blogs.nytimes.com/2015/08/10/not-vaccinating-children-is-the-greater-risk/?_r=0

[3] http://www.theguardian.com/society/2013/apr/25/measles-mmr-the-essential-guide

The basis for the expert opinion is in each case not only field-specific information ("backup evidence") but also some *field-dependent inference strategy*, applied directly by the expert source mentioned in Examples 1 and 3, and indirectly (by the expert's own expert sources) in Example 2. How should the differences among texts like these be represented, and what new critical questions do these arguments invite?

2 Field-dependent argumentation devices

Expert fields may build up repertoires of reasoning strategies over time, resulting in field-dependent inference rules. When any such new inference rule is proposed, other experts may challenge it, describing undercuts or rebuts to the strategy (as we will describe in 2.1 and 2.2). Iterative repair and critique continue, often over long periods of time, until the strategy is defeated, abandoned, or stabilized.

We will use the term *argumentation device* to describe a stable inference rule, currently accepted within a given field as a repeatable method for generating new, valid arguments within the field's domain. An argumentation device may contain material components that augment human reasoning in various ways and institutional components that underwrite their dependability.

In many respects, argumentation devices resemble argumentation schemes. Schemes, though, are generally assumed to be domain-independent and stable over long periods of time [Chesñevar *et al.*, 2006, p. 297], while the inventions we call argumentation devices are deeply entwined with the state of knowledge in a given domain. They work like schemes (as rules that justify drawing a conclusion from data); and like schemes, they have specifiable critical questions. However, the critical questions needed to evaluate the output of an argumentation device need to be *discovered* for each such device, often by seeing how the device fares in actual debate among experts, and then again, in larger contexts (like public debate) where the output of the device may be used as evidence for some further conclusion. They may change in response to change in the substantive knowledge of the field, as when some newly discovered fact about the phenomena exposes a previously undetectable way for the device to go wrong.

Argumentation devices can be extremely complex, incorporating material and institutional components that simply do not figure in ordinary schemes. For domains advancing high-stakes claims, like medical research, there are many different motivations for critical scrutiny (scientific commitment to empirical adequacy, pragmatic interest in quality of health care, patient concern for safety, financial interest in health care products, and more), and any of these motivations can lead either to the discovery of new critical questions or to the invention of new strategies for disarming them. In the next two sections, we introduce two argumentation devices that have emerged over the past half-century and co-evolved rapidly, supported by significant investment in material and institutional resources.

2.1 Randomized controlled trials

Establishing and defending claims about medical treatments is central to health science and practice. Although the problem has existed throughout human history, our standards for defense of such claims have changed dramatically in the last century, with the invention of the Randomized Controlled Trial (RCT). RCTs combine three features: (1) a comparison of a treatment of interest with a control condition (or with an alternative treatment); (2) random allocation of patients to treatment conditions; and (3) "blinding" of patients and researchers to the treatment any given individual receives.

Meldrum [2000] provides an illuminating account of the emergence of RCTs, documenting the series of innovations that, when combined into a single experimental design, became the standard against which all other medical evidence has come to be compared. We summarize her account here to highlight the fact that specific innovations (like random allocation) serve specific argumentative functions, so much so that their omission is said to make the experiment invalid as evidence for a conclusion about the effect of a treatment.

Prior to the 1900's, controlled experiments in human health were rare, and according to Meldrum, even more rarely conducted on treatments that could be administered to individual patients. Medical practitioners engaged in careful observation and sharing of results, and the literature was filled with case reports of what had worked in individual cases, but without procedural controls needed for strong inference from these observations.

Proliferation of treatments – particularly drugs and patent medicines – led to the formation of assessment agencies in the early 1900's, including the American Medical Association's Council on Pharmacy and Chemistry, and the first U.S. federal bureau empowered to review "the extravagant claims" made by the pharmaceutical industry of the time [Meldrum, 2000, p749]. Of central importance to our treatment of RCTs as an argumentation device is the role agencies played in challenging these extravagant claims.

To understand RCTs as an argumentation device, it is important to understand how profoundly doubt, disagreement, and error have affected the elaboration of this device over time. Scientists working with human subjects had to *discover* the need for randomization in the assignment of patients or other subjects to experimental conditions; the general superiority of comparisons based on randomly assigned groups is counterintuitive, but is nowadays universally acknowledged to be the best defense against bias or suspicion of bias. Other innovations like double-blinding were added as standard features of experiments on human subjects, not because logic requires them, but because of the practical *discovery* that patients' and experimenters' expectations could affect health outcomes, leading to novel criticisms of experiments for falling prey to "the placebo effect." RCTs with various forms of blinding are the present standard for evidence in medicine, but they achieved their present status only slowly,

and only incrementally. At each stage of development, it has been a device meant to disarm known objections to the conclusions drawn from a set of observations.

RCTs stabilized into a standardized, widely accepted form only in the late 1950's [Meldrum, 2000, p754], about ten years after the first large-scale trials were initiated (1946 in the US, 1947 in the UK). A decade later RCTs gained institutional status. In the wake of thalidomide-associated birth defects, the U.S. Food and Drug Administration began to investigate new approaches for reviewing drugs for safety [Meldrum, 2000]. This led to a 1970 regulation enshrining the RCT in U.S. law.

RCTs are not by any means a secure defense for a claim about a treatment effect. A series of RCTs, each competently executed, can come to different conclusions about a treatment. And each one remains vulnerable to subtle counterarguments that only expert researchers are likely to discover—previously unknown confounds, for example. However, RCTs handily defeat most other forms of evidence that might be advanced for the same class of claims. They are a "package deal" of evidence for a claim and evidence against a standard set of possible rebuttals, creating a strong but still defeasible conclusion.

2.2 Cochrane Reviews

As noted briefly above, RCTs on a particular treatment may accumulate within a scientific literature, each reporting some measurement of the effect of the treatment. Despite the widely acknowledged value of RCTs for evaluating treatment effects, expertise in interpretation is still necessary. One of the things experts know is that random variability is always present in the results of any series of identically designed experiments on human subjects. This creates an opportunity for confirmation bias to operate as readers cherry-pick results that support their beliefs and ignore or discount results that do not. Accompanying the rise of RCTs in medicine is another important invention, the systematic research review designed to aggregate evidence from many individual studies into a statement of what the research *as a whole* may be taken to support. Over just the past three decades, a highly standardized form of systematic review has emerged, known as the Cochrane Review.

Cochrane Reviews are named for Archie Cochrane, a Scottish doctor and epidemiologist, who championed the use of RCTs for guidance of clinical practice. In 1989, the publication of a 2-volume work on pregnancy and childbirth marked what Cochrane regarded as "a real milestone in the history of randomised trials and in the evaluation of care" [Chalmers *et al.*, 1989; and Cochrane's Foreword]. This was the first major systematic review in health science, a massive undertaking involving ten years of effort to review over 3000 controlled trials published since 1950 [Review, 1990]. A Cochrane Review is a review of literature conducted using very well-defined procedures outlined in an official handbook.[4] These procedures include exhaustive search for relevant studies; use of scoring rubrics for evaluation of the relevance and strength of evidence in each study; prescribed methods for combining information quantitatively; preferred methods for presentation of findings; and more

Unlike RCTs, systematic reviews do not generate new observations. They assemble evidence that already exists in a scientific literature and draw inferences from this evidence in a highly disciplined way. Evidence that would be considered inconsistent from a common-sense point of view is taken as input to the review, and interpreted in light of what experts know about variability. A Cochrane Review treats study-to-study variation in findings from multiple RCTs as normal and unremarkable, and because all relevant evidence is included, it offers good defense against any charge of cherry-picking. New reviewing standards emerge in response to problems noticed in the quality of argumentation produced by a review. For example, the Cochrane handbook includes cautions against "common mistakes" made in reviewing, such as concluding that there is evidence of no effect of an intervention when all that is really justified by the literature is that there is no evidence of an effect.[5] Against a charge that the Cochrane Review is only as good as the body of primary research available for aggregation, the Cochrane Collaboration (more than 37,000 contributors from over 130 countries) has adopted a formal practice of "grading" the strength of the evidence base itself.

Although systematic review methods are still in a period of rapid methodological innovation, the Cochrane Review has already achieved the status of a trusted argumentation device, largely because its procedures are so explicitly linked to critical questions on which earlier styles of research synthesis regularly failed. The methodical search procedures required for a Cochrane Review make it hard for a critic to object that evidence was assembled to fit the reviewer's own hypothesis. Counter-arguing individual studies (a once-common practice in narrative reviews of literature) is replaced with careful and explicit coding decisions applied impartially to the entire corpus of potentially relevant studies. Reviewer bias is further minimized through highly structured reporting methods: For example, if the review includes meta-analysis (a technique for transforming results of each individual study into a quantitative effect size measure), the results must be displayed as a "forest plot" that allows readers to inspect results on a study-by-study basis.

3 Modeling the role of argumentation devices

In a very preliminary way, we want to consider the challenges of including argumentation devices like these in formal models. Argumentation devices resemble schemes in most respects; they serve as reusable links between different collections of data and conclusions drawn from these data. They are *applied* to data, and although devices do not need defense in each application, they do have a context-independent defense that can be attacked either in the particular occasion of use or in a general critique of all

[4] http://handbook.cochrane.org

[5] Cochrane Handbook Part 2 section 12.7.2

arguments using the device. In an argument network, they would be better represented as a scheme node than as an information node. In a Toulmin diagram, the device is the warrant for conclusions drawn from data. The field-dependence of a argumentation device will commonly be most apparent in what appears in the backing for the device.

An argumentation device gains its status through incorporation of various assurances of its own ability to deliver reliable conclusions, including new institutional resources that underwrite the device as a whole. The Cochrane device is a particularly clear example, since it depends very openly on the growth of institutional resources to assure that a conclusion from a Cochrane Review is based on the most exhaustive search possible for relevant evidence. Although a machine-searchable database of medical research literature (MEDLINE, for the US National Library of Medicine) has been available since the 1960s, the Cochrane Collaborative has created a specialized database specifically for controlled trials, known as CENTRAL (Cochrane Central Registry of Controlled Trials), that includes both a subset of MEDLINE entries and other items retrieved from a variety of sources, including manual search of conference programs by members of the Cochrane Collaboration. Reviewers are expected to search both MEDLINE and CENTRAL to identify every possible relevant item, and to examine each item for whether it meets inclusion criteria. A typical Cochrane Review will identify thousands of potentially relevant items and winnow these to a few dozen studies that actually provide relevant data.

The resources that are required for an argumentation device to operate at all need some presence in any graph, diagram, or other formal representation of an argument from expertise that is itself an argument from some field-dependent device—arguments like those presented in Examples 1, 2, and 3. These resources are meant as strengtheners of the expert argument, but they are also a system of delegations in which responsibility for the validity of any one conclusion has been spread throughout a huge collective of participants. The individual performers of Cochrane Reviews take responsibility for faithful adherence to Cochrane procedures, but responsibility for the exhaustiveness of the search is delegated to databases; the responsibility for what is available to be retrieved is delegated to funding agencies that set research priorities; and the responsibility for establishing hierarchies of evidence is delegated to trusted working groups within the Cochrane Collaboration. These delegations are themselves an interesting fact about contemporary argumentation [Jackson, 2015a] that could be better understood if they were explicitly included in formal models of argumentation. Figure 1 illustrates how these delegations might be incorporated in a Toulmin diagram, as forms of backing for the Cochrane Review procedure.

The most distinctive differences between argumentation devices and familiar argumentation schemes are their field-specificity and their openness to redesign [Jackson, 2015b]. The primary purpose of an argumentation device is to provide convincing evidence for a conclusion *to people who understand the workings of the device and have confidence in it*. Both RCTs and Cochrane Reviews share a well-defined context consisting of an audience of medical experts, a pre-existing literature, and other features whose argumentative relevance is as yet unclear. Both have developed iteratively from critique within the field, and both are still being elaborated to eliminate vulnerabilities in their conclusions. Argumentation devices demand consideration of context: not only the community within which they emerge but also the state of play within that community.

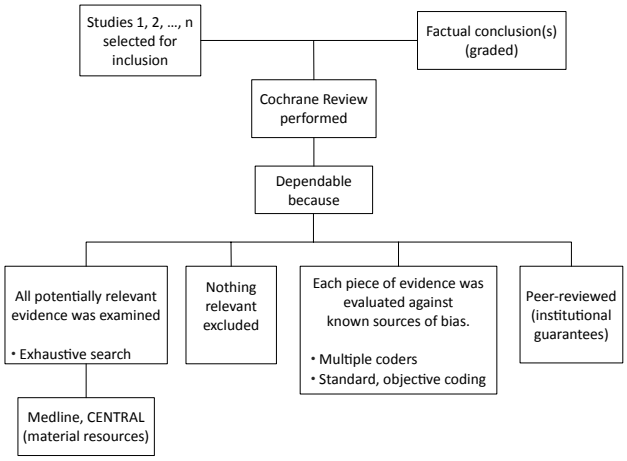

Figure 1. General form of a Cochrane Review's argument, with delegations of responsibility in the backing for the warrant.

4 Critical questions about devices

A Cochrane Review is organized around both presentation of data and response to critical questions about the gathering and interpretation of the data. In other words, much of the text of a Cochrane Review consists of explicit answers to the questions other experts would be presumed to have. An enormous advantage that comes with use of an established argumentation device is that the device itself does not need defense for each occasion of use. It can function as a warrant for many specific conclusions, each of which has its own unique body of evidence.

Although an argumentation device may be applied in a completely uncontroversial way within an expert field, that is no protection against questions or challenges from beyond the field. The fact that a device has earned the confidence of a group of experts is not quite sufficient to earn trust from other potential audiences. The testing ground for any new argumentation device is argumentation itself. The device must earn its status by withstanding critique. We end by considering what kinds of questions might arise, reasonably or even unreasonably, as devices like Cochrane Reviews enter new testing grounds.

To begin with, critical questions relevant to arguments supported by Cochrane Reviews share some similarities with critical questions for arguments from expert opinion. The accuracy of an arguer's understanding of expert opinion

is always relevant. Consider again Example 3—the Guardian's appeal to a Cochrane Review of 54 studies as evidence against any link between autism and MMR. The review [Demicheli *et al*., 2012] did in fact look at 54 studies, but only 10 included autism as an outcome variable, and by the reviewers' assessments of quality, it does not appear that they would agree that the 10 studies relevant to this particular claim provide "extremely strong" evidence. (None of the ten were RCTs, and none individually offered a strong design for detecting a link between MMR and autism. Reviewers classified all ten of the autism-related studies as containing either "high" risk of bias or "moderate/unknown" risk of bias.) Where the Guardian has gone wrong here is in assuming that a "gold standard" procedure can produce "extremely strong" evidence from a research literature that is inadequate, a failure to understand that any limitations of the primary research literature are inherited by the review.

But in addition to questions similar to those relevant to assessment of argument from expert opinion, any device of this kind will be vulnerable to challenges specific to the device. A significant feature of the current design of the Cochrane Review is that it aggregates evidence from scientific literature (sometimes including unpublished data, but mostly from reports published in some form and included in a database). By design, a Cochrane Review ignores evidence that could, in principle, be relevant. This includes the very wide range of evidence types that can be supplied by ordinary people paying attention to their own health and their own reactions to treatments. For the vaccination controversy, this includes evidence that is highly credible to many members of the public (first-hand parent observations of adverse reactions to vaccines); the fact that no serious effort has been made to systematically review these reports is a reason for those affected to question the credibility of the institutions that back the Cochrane device. So one class of critical questions have to do with whether there are forms of evidence the device does not (or cannot) ingest.

Another class of critical questions have to do with biases built into the device. The device is always designed to answer some set of questions but not others, and to assume those things that its expert users assume. To illustrate, a common notion within anti-vaccination discourse is that the institutions responsible for the production of the primary research have so strong an interest in mass immunization that they conceal or suppress evidence of serious risks—characterized as conspiracy thinking by Oliver and Wood [2014]. While no one seriously expects scientists to respond to conspiracy theories, it is certainly reasonable to ask what interests and assumptions shared within an expert community might make the community blind to certain evidence or deaf to certain arguments.

Seeing argumentation devices as an encapsulation of how the expert community reasons, questions can be asked not only about the individual use of the device in one argument, but also about the assumptions the device encapsulates. This is an important shift of scale that involves questions that may need to be asked to correct an unsuspected bias. Such questions can sometimes be formulated more easily by non-experts than by the experts themselves, by coming from a perspective with its own biases, but different ones.

In health controversies where much is at stake, both experts and non-experts will fully explore the possible grounds for disagreement with conclusions drawn from experts' argumentation devices, and the devices themselves will improve in order to better withstand critique. An important goal in modeling argumentation devices is to expose avenues for productive examination of the devices by non-experts, and to assist experts in responding productively to even the most skeptical critique.

Acknowledgments

The second author was supported by training grant 5T15LM007059-29 from the National Library of Medicine and National Institute of Dental and Cranio-facial Research.

References

[Chalmers *et al.*, 1989] Iain Chalmers, Murray Enkin, and Marc J.N.C. Keirse. *Effective care in pregnancy and childbirth: Pregnancy*. Oxford University Press, 1989.

[Chesñevar *et al.*, 2006] Carlos Chesñevar *et al.* Towards an Argument Interchange Format, *The Knowledge Engineering Review* 21(4): 293–316, December 2006.

[Demicheli *et al.,* 2012] Vittorio Demicheli, Alessandro Rivetti, Maria Grazia Delabine, and Carlo di Pietrantonj. Vaccines for measles, mumps and rubella in children. *Cochrane Database of Systematic Reviews 2012* Issue 2. Art. No.: CD004407.

[Jackson, 2015a] Sally Jackson. Deference, distrust, and delegation: Three design hypotheses. *Reflections on Theoretical Issues in Argumentation Theory* (pp. 227-243). Springer International Publishing, August 2015.

[Jackson, 2015b] Sally Jackson. Design thinking in argumentation theory and practice. *Argumentation,* 29(3): 243-263, August 2015.

[Meldrum, 2000] Marcia L. Meldrum. A brief history of the randomized controlled trial: From oranges and lemons to the gold standard. *Hematology/oncology clinics of North America*, 14(4), 745-760, August 2000.

[Oliver and Wood, 2014] J. Eric Oliver and Thomas Wood. Medical conspiracy theories and health behaviors in the United States. *JAMA Intern Med, 174*(5):817-818, 2014.

[Review, 1990] Book review of [Chalmers *et al.*, 1989]. *Birth*, 17(1): 55–62, March 1990.

[Toulmin, 1958] Stephen E. Toulmin. *The uses of argument*. Cambridge University Press, 1958.

[Walton *et al.*, 2008] Douglas Walton, Chris Reed, and Fabrizio Macagno, *Argumentation Schemes*, 1st edn, Cambridge University Press, August 2008.