Systematic Reviews as an interface to the web of (trial) data: Using PICO as an ontology for knowledge synthesis in evidence-based healthcare research

Chris Mavergames[1], Silver Oliver[2], Lorne Becker[3]

[1]The Cochrane Collaboration, Director of Web Development, Freiburg, Germany
mavergames@web.cochrane.org
[2]Ontoba, Information Architect, London, United Kingdom
silver.oliver@ontoba.com
[3]Cochrane Innovations, Director, Tampa, USA
lornebecker@gmail.com

Abstract. Linked data and semantic technologies offer flexible and powerful solutions for connecting and synthesizing knowledge in the domain of healthcare research, in particular the area of evidence-based medicine (EBM). Systematic reviews of healthcare interventions, the primary methodological approach for evidence synthesis in EBM, involve a rigorous and time-consuming process of collecting and analyzing data from all the studies conducted around a particular clinical question. The number of primary studies reported each year is rising exponentially, and the process of producing systematic reviews that synthesize data from these studies is labor-intensive and keeping these reviews up to date a huge challenge. Currently, the reviews are primarily published in PDF format and much of the value is locked away from programmatic access. This position paper discusses the potential in using linked data technologies to improve discovery of knowledge in systematic reviews by using the PICO (Population, Intervention, Comparison, Outcome) framework as an ontology to aid in knowledge synthesis.

Keywords: linked data, systematic reviews, randomized controlled trials, Cochrane Collaboration, Cochrane Library, semantic web, Drupal, RDF, SPARQL, OWL, ontologies

1 Introduction

The evidence-based approach to health care decision-making calls for clinicians to find and apply the best and most up-to date results of medical research in their clinical decision making. This is no easy task, even if the search for evidence is restricted to the highest quality evidence (such as that provided by randomized controlled trials or RCTs). Thousands [1] of RCTs are completed each month and most result in several papers or reports which are often published in different journals and may not reference each another. Added to this, different RCTs addressing the same clinical question may have widely differing findings.

A relatively new type of approach, the Systematic Review, has been developed to address this problem. Systematic Reviews of healthcare interventions use rigorous methods to identify all of the studies that have investigated a particular clinically-relevant question, appraise their methods and combine their results to present a synthesis of the evidence for the question of interest. The PICO format is used to frame the clinical question as to the Population (patient with problem), the Intervention (drug or other) being given, the Comparison of that intervention with another intervention, no treatment or a placebo, and in relation to which Outcome(s) (symptoms relieved, etc.). Annotation of data and content by PICO forms a powerful framework for navigating and synthesizing the evidence and can help to create an interface onto the web of trial data available for analysis.

1.1 The Cochrane Collaboration and systematic reviews

The Cochrane Collaboration (http://www.cochrane.org), an international, non-profit research organization has developed much of the methodology underpinning the systematic review approach and publishes The Cochrane Database of Systematic Reviews (CDSR), which now contains more than 5,200 of these innovative articles. Although each Cochrane Review can be downloaded as a PDF, the CDSR was originally conceptualized as a database and has never been published as a paper journal. At its inception, almost 20 years ago, the CDSR was distributed on floppy disks, transferring to an online format once the World Wide Web came of age. Cochrane Reviews are continually updated and previous versions of the Reviews are available for download. Thus, one can track the "status" of the answer to a particular clinically-relevant question over time.

Making Cochrane reviews as accessible as possible is at the core of Cochrane's remit. Cochrane aims to help

people make well-informed decisions about health care by preparing, maintaining and promoting the accessibility of systematic reviews of the effects of healthcare interventions. [2] While Cochrane Reviews are internationally recognized for their rigor, validity and unbiased synthesis, they are long and complex documents that can be intimidating for naïve or even for somewhat experienced users. [2] There are Reviews that span 800-pages and ones with more than 750 analyses in the form of forest plots (http://en.wikipedia.org/wiki/Forest_plot).

Recently, Cochrane began a linked data project to investigate how semantic technologies could assist in helping users navigate the evidence in Cochrane Reviews, while also assisting in the process of producing them. The project hoped to build on the work already done in this space around creating ontologies for clinical research (OCRe - http://bioportal.bioontology.org/ontologies/1076) as well as work done to convert clinical trial reports into RDF (http://linkedct.org). The aim of the linked data prototype was to improve access to Cochrane Reviews through a demonstrator website orientated around the user stories and questions that user research showed clinicians and other end-users of Cochrane evidence are seeking to answer. Emphasis was placed on new access points for search and navigation and PICO emerged as the primary framework for tagging the content.

1.2 PICO

The PICO framework has long been held as a key mechanism for information retrieval in evidence health care. [3] PICO stands for Population (patients with a condition), Intervention, Comparison and Outcome(s). For example, is drug A (Intervention) effective for the relief of B condition in C Population in Comparison with X drug (or placebo) for Y Outcome(s) (symptoms relieved, etc.). Cochrane reviews use the PICO framework extensively at each stage of formulating a review: question, searching, screening, analysis and publication. PICO was identified as a key method of providing access points and facets for browsing in the Cochrane linked data demonstrator.

Research into using PICO for information retrieval [3,4] has affirmed its usefulness and explored options for identifying these elements in corpora of documents via manual annotation and natural language processing techniques. The following aspects of the PICO framework from this research [3,4] were drawn upon in the design of the demonstrator:

- Weakness of expressivity in terms of relating PICO elements
- Overloading of P (condition, aspects of the population)
- Overriding benefits of P and I (as opposed to C and O) to retrieval.
- Works better for therapeutic questions

Thus, the prototype looked to focus on the following:

- Creating a Cochrane Review ontology that also covered the studies that are included in the reviews.
- Modeling PICO within this ontology of reviews and studies by breaking the elements of PICO into more specific classes and their relationships.
- Drawing on linked data sets for populating the PICO elements with instances from controlled vocabularies such as SNOMED CT.

2 The ontology

A prototype ontology (Fig. 1) that partially models Cochrane Reviews, the studies that are included in them, and the PICO framework was created. The purpose of this ontology was to fulfil the specific set of user stories drawn from user research into the kinds of questions people were trying to answer using Cochrane systematic reviews and to drive improvements to navigation and discoverability of content within the documents. The prototype ontology was developed from a “product” (published Cochrane Review) perspective and thus the design was not driven by the underlying methodology of doing systematic reviews.

While care was taken to ensure that it is methodologically sound in its design, its purpose was to demonstrate the advantages of using a linked data approach to describing, storing and managing our content and processes. The next iteration of the ontology will be methodology driven and will likely result in the creation of a suite of ontologies to power evidence synthesis.

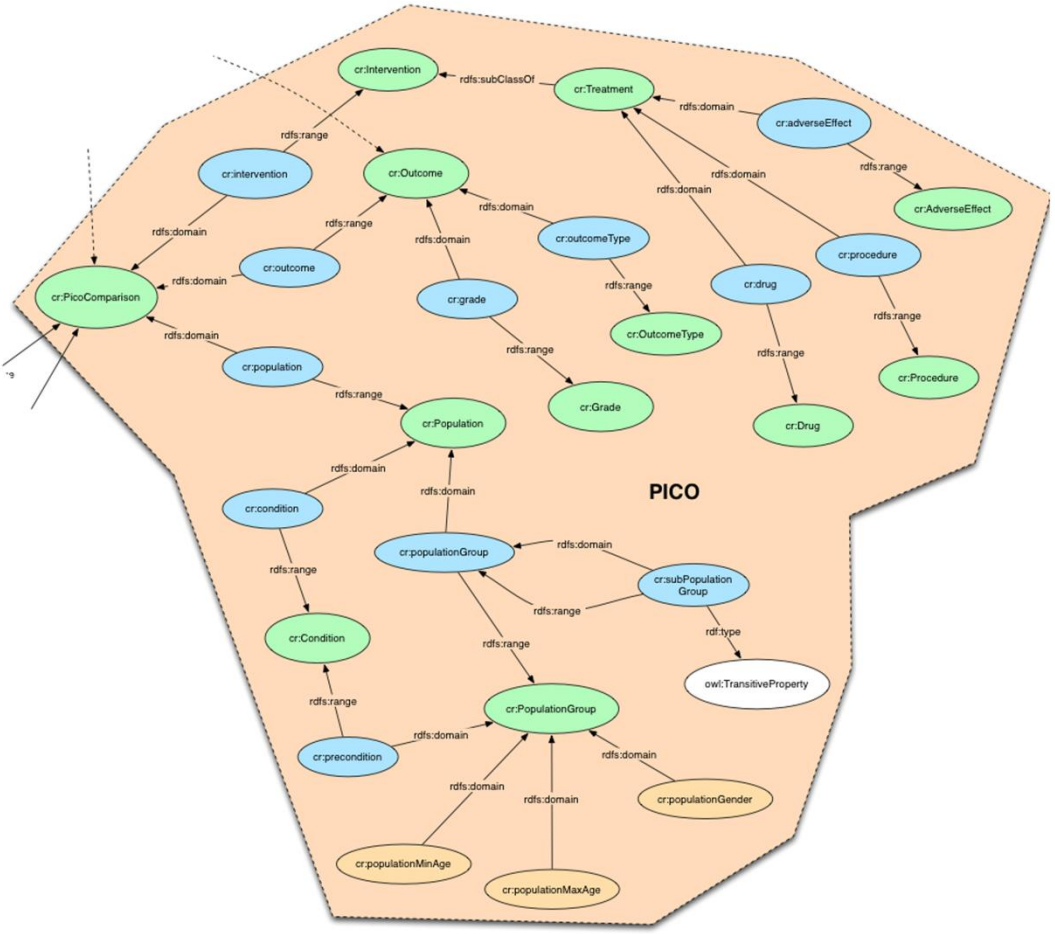

Fig 1. PICO portion of the Cochrane Review prototype ontology

3 The demonstrator

3.1 Information architecture

For the demonstrator [5], we focused on a small corpus of Cochrane Reviews on Asthma that addressed PICO questions where the interventions assessed were drugs. The information architecture of the site aimed to create:

- **Page (URI) per condition, intervention, study, review.** The demonstrator exposed the PICO model as part of the information architecture of the site. For core classes of concepts, each had a specific URI and links to related things in the graph. This encouraged browsing through rich user journeys. This was also considered beneficial from a search perspective exposing the key classes to be landing pages for external search.

- **Searching Reviews by drug name.** Currently, there is no cross-indexing against variant names of drugs in Cochrane Reviews. We have linked to Drugbank (http://www.drugbank.ca/) which includes most of the variants of drug names including the different brand names and generic names used in different countries. We created a "semantic search" that allows users to type any name for an asthma medication and find the relevant Cochrane Reviews. This functionality would greatly improve the discoverability of Cochrane content in The Cochrane Library (http://www.thecochranelibrary.com) as, for example, if you search for "Prozac" you get zero results, but if you search for "fluoxetine" you get 30 results.

- **Displaying selected portions of reviews.** Clicking on any title on the "List of Reviews" page in the demonstrator takes you to a custom view of that review that could eventually be customized based on

the query or route of browsing by PICO, once the ontology is fully populated with instances for tagging these elements.

- **Linking out to selected content.** In addition to linking to Drugbank as noted above, we have linked to SIDER (http://sideeffects.embl.de/), a linked data set that includes information on side effects from FDA label information, as well as linking to PubMed records via PubMed identifiers supplied by the Cochrane Register of Studies (CRS), Cochrane's "studified" register of RCTs published in CENTRAL (Cochrane Central Register of Controlled Trials), the largest collection of RCTs in the world.

3.2 Building the demonstrator site

The following steps were taken in building the demonstrator site [5]:

- **Development of user stories.** The project began with a relatively detailed analysis of key user stories. So all features built upon a sound understanding of business and user needs.
- **Development of a model.** A RDF model was developed to support the prototype. Here the majority of the focus was the interplay between the PICO elements and associating them with the right section of the review.
- **Extract Transform Load (ETL).** Scripts (Java) were created for transforming the XML review documents into RDF. Taking the XML and structure and minting URIs where necessary.
- **Annotation of reviews.** The reviews were hand annotated against the URIs populating PICO classes for I and C for Asthma drugs and interventions. This was done in a spreadsheet and converted to RDF using Open Refine with the Deri RDF plug-in. The annotations were at the question level of the review.
- **Build the "views".** The open-source content management system Drupal was used as the prototyping engine for the purpose of creating a demonstrator site. Drupal "plays nicely" with the semantic web stack including an RDFx module with a module called SPARQL Views which allows for SPARQL queries to be constructed within the core Drupal Views system. OWLiM was used as the triple store software and Drupal connected to this repository via SPARQL Views and queries and results were generated and rendered within the Drupal website. This allowed us to quickly create a working website that can be quickly styled and made functional using Drupal's built-in theming and templating system.

4 Discussion

4.1 PICO as an ontology

Though the work was only a proof of concept, and the development of the model is still evolving, the work to date has raised a number of significant benefits of capturing PICO as a rich RDF model.

- RDF lends itself well to rapid prototyping: the flexibility of an RDF data store and the ability to play with a variety of linked data sets to see how they enrich features. For example including RDF data from SIDER without having to do any additional modelling.

- Inference: using transitive taxonomical relationships in the drug class hierarchy to address issues of intervention queries at different hierarchy depths. For example Salmeterol having a child relationship with respect to Long-Acting Beta 2 Agonists. The relationship that associates the role of a drug in an intervention was transitive-over the hierarchy and thus this reasoning insured any query for Long-Acting Beta 2 Agonists would include PICO's referencing specific drugs such as Salmeterol.

- Rich queries: SPARQL queries taking advantage of the rich graph of PICO data showed great potential. Queries that would ask for a drug (including brand named drugs via Drugbank at any depth of the drug hierarchy) that has been compared to any other intervention for a particular condition go significantly further than equivalent information retrieval approaches that use PICO as a source of identifiers for classifying a document.

- Ability to map out to standard models and upper-level ontologies: RDF lends itself well to providing a mechanism to express the specific needs of the product in hand and map out to existing models for interoperability moving forward.

For the purposes of the prototype, PICO was only associated with the question level of the review and only I and C were used for annotation thus far. In addition, we identified at least three key places for the annotation of PICO with relation to Cochrane systematic reviews that should be explored and implemented going forward:

- Study level: PICOs being reported in the trial reports
- Question level: The PICO question posed by the systematic review
- Findings level: The results from analyses and meta-analyses for a particular PICO question in the review

And, further development of a stand-alone, PICO ontology will model, in detail, the facets of the various components of the PICO framework, in order to capture the richness and granularity present, especially with regards to P (population) and O (outcomes).

Each of these levels of metadata would provide increasingly more value to end-users in terms of discoverability, navigation and traversing the graph of available information and analyses. For example, users would be able to restrict searches for reviews to those that actually find results for a given question as opposed to simply where this question has been asked but without result. There is clearly a cost-benefit analysis to be done here, but machine-supported annotation has been explored for the labour intensive job of annotating at the study level, where studies for a given review might extend to hundreds.

5 Conclusion

Our experience with the Cochrane linked data project to date has convinced us that it has potential to become an "enabling technology" for the Collaboration that could allow us to do more with our data in terms of synthesizing evidence, but also to enable better discoverability and presentation. However, there are a number of issues that should be explored as we decide on how best to leverage linked data and semantic technologies both within the Cochrane technology stack. Also, there is possibility of moving toward an "operating system" for evidence-based healthcare research by leveraging semantic technologies.

The systematic review is an excellent example of the article as an interface to the underlying web of trial data. The use of PICO as an upper-level ontology for annotating studies and systematic reviews holds great promise in moving toward "living systematic reviews", whereby the notion of the systematic review as a document slowly fades in favour of "high quality online evidence summaries that are dynamically updated as new evidence becomes available". [6]

References

1. Bastian H, Glasziou P, Chalmers I (2010) Seventy-Five Trials and Eleven Systematic Reviews a Day: How Will We Ever Keep Up? PLoS Med 7(9): e1000326. http://www.plosmedicine.org/article/info%3Adoi%2F10.1371%2Fjournal.pmed.1000326
2. Mavergames, C. Sustainability and Cochrane Reviews: How Technology can Help Plenary talk by Chris Mavergames from UK Contributors' Meeting in Loughborough, March 2012 http://www.slideshare.net/mavergames/sustainability-and-cochrane-reviews-how-technology-can-help-12207716.
3. Huang, X., Lin, J., Demner-Fushman, D. Evaluation of PICO as a Knowledge Representation for Clinical Questions. AMIA Annu Symp Proc. 2006; 2006: 359–363. PMCID: PMC1839740
4. Boudin, F., Nie, J., Dawes, M. Clinical information retrieval using document and PICO structure In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (2010), pp. 822-830.
5. Cochrane Linked Data Demonstrator, http://demonstrator.dev.cochrane.org
6. Elliott, J., Mavergames, C., Becker, L., Meerpohl, J., Thomas, J., Gruen, R., Tovey, D. Achieving high quality and efficient systematic review through technological innovation. BMJ Rapid Response. January 2013. http://www.bmj.com/content/346/bmj.f139/rr/625503.