Crowdsourcing a Comprehensive Clinical Trial Repository

Gert van Valkenhoef and **Joël Kuiper**
g.h.m.van.valkenhoef@rug.nl; joel.kuiper@rug.nl
Department of Epidemiology, University Medical Center Groningen
University of Groningen, The Netherlands

Abstract

We present the open problem of building a comprehensive clinical trial repository to remove duplication of effort from the systematic review process. Arguing that no single organization has the resources to solve this problem, an approach based on crowdsourcing supplemented by automated data extraction appears to be the most promising. To determine the feasibility of this idea, we discuss the key challenges that need to be addressed.

Introduction

Undertaking systematic reviews of randomized controlled trials to assess interventions' relative effectiveness (henceforth 'systematic reviews') is a costly affair. By the year 2000, a well-conducted systematic review with meta-analysis already cost up to a year of full-time work (Allen and Olkin 1999). Making matters worse, the standards for quality have since increased (Moher et al. 2009), the number of trials published per year is still increasing (Bastian, Glasziou, and Chalmers 2010), and network meta-analysis broadened reviews from tens to hundreds of trials (Cipriani et al. 2009; Leucht et al. 2013). The investment required to produce systematic reviews has become prohibitive, and this raises the question whether the systematic review enterprise will remain sustainable.

To illustrate: most systematic reviews are conducted in relative isolation by small teams of researchers. And, typically, only the final manuscript describing the included studies, results, and conclusions is made available. This means that valuable intermediate products such as screening decisions, data extractions, and quality assessments remain with the research team, or are sometimes lost completely. This causes a lot of duplication of effort, and it has long been recognized that a structured database of clinical trial data could eliminate most of that effort (Sim et al. 2000; Sim and Detmer 2005). However, no such repository has been established.

We believe that the systematic review data repository (SRDR) is an important step in the right direction, since it captures some of the intermediate products of systematic reviewing (Ip et al. 2012). However, it models the current review process and captures data in a review oriented, rather than a trial oriented, format. So the question remains: *why has a comprehensive clinical trials data repository not been established, despite its obvious advantages?*

Several explanations are available. First, capturing all the minutiae of clinical trials is a hard problem (Carini et al. 2009), and efforts to build a complete ontology of the domain have progressed slowly (van Valkenhoef et al. 2012). Second, new publications and clinical trials appear rapidly (Bastian, Glasziou, and Chalmers 2010) and any centralized database of trials will thus be perpetually out of date. And third, the publish-or-perish mindset that is prevalent in academia has resulted in data protectionism, meaning that researchers are often reluctant to share data with outsiders. Finally, the number of trials that have been conducted so far is likely to be well over a million (Dickersin and Rennie 2003), and therefore no single organization is likely to invest the resources to extract data on all of them.

In the next sections we elaborate on these explanations and discuss potential methods for overcoming the challenges involved. We propose to use *crowdsourcing* (Doan, Ramakrishnan, and Halevy 2011; Brabham 2013), leveraging the small contributions of many (the crowd), to create the comprehensive repository of clinical trials. In this case, the crowd is not necessarily the general public, but rather the broad research community involved in systematic reviews. The crowdsourced repository should aim to preserve the intermediate products of reviewing in a way that allows future reviews to build on top of them; reducing the duplication of effort inherent in the current process. We hope to start a constructive discussion on what we believe are the key challenges in realizing this concept.

Challenges

Many challenges need to be addressed to build a comprehensive, structured, machine-readable, trustworthy repository of clinical trials data. Some of the questions center on how the data should be represented, and these problems have been recognized for some time (van Valkenhoef et al. 2012; Sim et al. 2000; Fridsma et al. 2008; Carini et al. 2009; Kong et al. 2011). Ultimately the best representation may depend on how the data will be used (van Valkenhoef et al. 2013). By contrast, our focus is on the socio-technical forces that should shape the design of a community-driven clinical trials repository. These have scarcely been considered in pre-

vious proposals for such repositories.

Process versus community

Systematic reviewing is grounded in formal processes for verifying that information is accurate (Higgins and Green 2009). For example, publication screening is often done in duplicate or triplicate by independent reviewers, after which inter-rater agreement is assessed using statistical methods (Orwin and Vevea 2009), and data are extracted using standardized extraction forms designed specifically for the review at hand. For any data repository to be used in the systematic review process, there must be a high level of trust in the accuracy of the data it provides. From this perspective, an obvious approach would be to strictly enforce that all contributions are made only as part of a thorough systematic review process. On the other hand, building a *comprehensive* repository of clinical trials is a large task, and from this perspective one would like to encourage contributions from as many users as possible. This warrants the design of a system in which contributions are as small as possible, even a single correction to a minor factual statement (microtasking). It would appear that these forces are in direct opposition, but existing systems have successfully balanced reliability with a low barrier to entry. A notable example is Wikipedia, which has managed to create a self-policing community of users that has built not just a text-based repository of knowledge, but also structured data (available through `http://dbpedia.org`) (Doan, Ramakrishnan, and Halevy 2011). Research projects have aimed to use both information extraction and user contributions to increase the information that is represented as structured data (Weld et al. 2008), or even to build fully fledged knowledge bases using crowdsourcing (Richardson and Domingos 2003).

Another social obstacle is that the high cost of assembling systematic review datasets may make researchers reluctant to share their data indiscriminately. Therefore, if the system is to capture all intermediate results of the systematic review process, it needs to offer a clear incentive for being so open. In part, this could be achieved by seeding the repository with data from sources such as ClinicalTrials.gov, or by using machine learning methods for screening (Cohen et al. 2006; Bekhuis and Demner-Fushman 2010; Wallace et al. 2010; 2011), data extraction (Cohen and Hersh 2005; de Bruijn et al. 2008; Boudin et al. 2010; Kiritchenko et al. 2010; Kim et al. 2011) and scoping (Shemilt et al. 2014). In addition, receiving clear credit for contributions made is often a motivation in itself, so the system should aim to give credit where it's due. Two related approaches proposed for genetics could be leveraged to this end: (1) microattribution, attributing small scholarly contributions such as data sets to a particular author and subjecting them to the same citation indices as journal articles, and (2) nanopublications, making single machine-readable assertions attributable (Giardine et al. 2011; Mons et al. 2011; Patrinos et al. 2012).

Dealing with change

As extractions of older trials are constantly being added and improved and as new trials continue to be conducted, a repository of clinical trials would be in a constant state of change. However, systematic review and meta-analysis need to work from a stable knowledge base, and be able to justify the decisions that were made based on the knowledge as it was at that point in time. Moreover, it is important to keep track of the *provenance* of data: to what extent can the data be traced back to its original sources? Ideally, when statements are derived from semi-structured sources such as ClinicalTrials.gov, a representation of how the data have been transformed should be available. These transformations, and the rules for deciding when they are applied, should then also be made available for review. Indeed, every user should be able to audit how each data element has been derived from its source, regardless of whether that source is a database or an unstructured PDF document. Techniques for schema and ontology alignment (Rahm and Bernstein 2001; Kalfoglou and Schorlemmer 2003; Shvaiko and Euzenat 2005) could help automate this process.

Keeping a repository up-to-date with external sources introduces the problem of detecting and incorporating changes to those external sources, and each source could intentionally or unintentionally disappear or fundamentally change without notice. This touches on a fundamental problem in distributed systems: how to reason about data that is essentially ephemeral. Distributed version control systems address this concern by requiring every user to keep a copy of the entire data set, including its history of changes (Ram 2013). They assign addresses (e.g. hashes) to each previous state of the data, and keep a record of each change that is made, to create a stable basis to reason from as well as an audit trail. Unfortunately requiring every user to keep a copy of the entire database introduces substantial complexity in establishing an authoritative source, as multiple (potentially unlimited) conflicting versions may exist. Many open-source software projects address the problem of authority by introducing a hierarchy of experts that decide whether changes should be incorporated in the 'master' version. However this model might be inappropriate for crowd sourcing a clinical trials repository: there is a lot of heterogeneity in subject areas and it might make the barrier to entry too steep. Therefore, a more centralized model based on event sourcing (Fowler 2005) or convergent replicated data types (Shapiro et al. 2011) may be more appropriate, because these techniques also deal with continuous change and allow previous states of the system to be reconstructed. Even if a consistent view of the claims made by each of the disparate data sources and users of the repository can be constructed, the problem remains that these claims may be more or less reliable, and that they may conflict. This problem is discussed more extensively in the next section.

Opinion versus fact

Systematic reviews aim to identify objective facts, but the process for doing so is heavily reliant on expert opinion to interpret and assess text-based reports. As systematic reviews are currently performed, the facts are established through independent assessment followed by discussion of differences to finally arrive at consensus. If statements made as part of past reviews or as independent contributions by users of the system are to be reused in a new systematic re-

view, when should such statements be considered fact, and when should they be considered mere opinion?

In addition, statements do not necessarily originate only from human experts, but could also be made by computational agents. For example, one such agent might be responsible for keeping the repository up-to-date with new and revised records on ClinicalTrials.gov, and another for matching intervention names against a standardized vocabulary.

In such an open ended system, a formal model for determining which statements are to be deemed trustworthy is required, since it creates the opportunity for mistaken and malicious contributions. For example, the reliability and utility of statements can be determined using a combination of machine learning and user feedback (Richardson and Domingos 2003). The repository can also be viewed as a multi-agent system, and the reliability of statements can then (in part) be derived from formal measures of trust or reputation of agents in such systems (Ramchurn, Huynh, and Jennings 2004). Most likely, a successful model would allow for feedback between both agent reputation and statement reliability and utility, and rely heavily on the user community as a source of these measures. Such a model would allow a collection of trustworthy facts, an *intersubjective consensus*, to be distilled from the full set of statements made, for example using a reliability threshold.

Teams performing systematic reviews may want to build from such a consensus fact base, but make their own additions and corrections, preferring these over the consensus. However, any changes made to the consensus fact base should be clearly presented, so that manipulations do not go undetected.

Summary

In this paper, we discussed the following key challenges that need to be addressed to enable the crowdsourcing of a comprehensive and machine-readable repository of clinical trials:

1. How can reliability be balanced with a low barrier to entry?
2. Can microtasks be made possible in this complex domain?
3. What are to be the incentives for reviewers to share their intermediate results openly?
4. Can the repository be kept up-to-date with constantly changing data sources?
5. How can stable systematic reviews be built on top of a constantly changing repository?
6. How should data provenance be tracked and made visible?
7. When do statements transition from opinion to fact, and can we enable reviewers to override consensus opinions in their projects?
8. How can we enable third parties to build computational agents that contribute to the repository, without compromising its integrity?

We did not consider how the data should be structured so that useful analyses can be built on top of that data, as this problem has been discussed at length elsewhere. We also did not outline how such a repository could be financed and operated, nor the ethico-legal concerns that arise from the tension between intellectual property and the "social good". Rather, we hope that this problem statement will spark a productive debate about the socio-technical problems posed by crowdsourcing a comprehensive trials repository, eventually leading to a solution for the systematic review problem.

Discussion

We would argue that the comprehensive trials repository we propose is both desirable and cost-effective from a societal perspective, but we acknowledge that there may be alternative solutions and that future developments may reduce its usefulness. The current trend is towards more open and complete publication of clinical trial results, especially for trials supporting marketing authorization decisions. Hopefully, the results of most trials will eventually be published in systems similar to ClinicalTrials.gov, eliminating the need to extract data from text-based publications on these trials. However, the repository we propose would still be relevant for a number of reasons: (1) many older trials are likely to remain relevant for the foreseeable future and are not available from a structured data source; (2) indexing the structured data sources in a comprehensive repository adds value by removing barriers to data identification and integration; (3) researchers will still want to annotate the structured data sets with appropriate meta-data; and (4) providing all data through a consistent interface enables automated reasoning and knowledge discovery.

Previous and future developments in artificial intelligence and semantic web technologies will be instrumental to the success of the repository we propose. For example, tightly integrated machine learning methods to speed up the abstract screening task would provide an incentive for reviewers to use the system. Automated methods for the extraction of data and meta-data from publications is an area that needs further research. Moreover, machine learning algorithms could have greater impact if they were implemented in agents that both consumed and contributed to an evolving knowledge base. How human and computational agents can interact to build such a knowledge base is an active area of research in multi-agent systems. New approaches may be needed to blend uncertain information on the trustworthiness of agents with existing semantic web technologies for knowledge representation and reasoning. Finally, approaches to the automated detection of data extraction errors could be developed to safeguard the overall quality of the data repository.

References

Allen, I. E., and Olkin, I. 1999. Estimating time to conduct a meta-analysis from number of citations retrieved. *Journal of the American Medical Association* 282(7):634–635.

Bastian, H.; Glasziou, P.; and Chalmers, I. 2010. Seventy-five

trials and eleven systematic reviews a day: How will we ever keep up? *PLoS Medicine* 7(9):e1000326.

Bekhuis, T., and Demner-Fushman, D. 2010. Towards automating the initial screening phase of a systematic review. *Studies in Health Technology and Informatics* 160:146–150.

Boudin, F.; Nie, J. Y.; Bartlett, J. C.; Grad, R.; Pluye, P.; and Dawes, M. 2010. Combining classifiers for robust PICO element detection. *BMC Medical Informatics and Decision Making* 10:29.

Brabham, D. C. 2013. *Crowdsourcing*. MIT Press.

Carini, S.; Pollock, B. H.; Lehmann, H. P.; Bakken, S.; Barbour, E. M.; Gabriel, D.; Hagler, H. K.; Harper, C. R.; Mollah, S. A.; Nahm, M.; Nguyen, H. H.; Scheuermann, R. H.; and Sim, I. 2009. Development and evaluation of a study design typology for human research. In *AMIA Annual Symposium Proceedings 2009*, 81–85.

Cipriani, A.; Furukawa, T. A.; Salanti, G.; Geddes, J. R.; Higgins, J. P. T.; Churchill, R.; Watanabe, N.; Nakagawa, A.; Omori, I. M.; McGuire, H.; Tansella, M.; and Barbui, C. 2009. Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. *The Lancet* 373(9665):746–758.

Cohen, A. M., and Hersh, W. R. 2005. A survey of current work in biomedical text mining. *Briefings in Bioinformatics* 6:57–71.

Cohen, A.; Hersh, W.; Peterson, K.; and Yen, P.-Y. 2006. Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association* 13(2):206219.

de Bruijn, B.; Carini, S.; Kiritchenko, S.; Martin, J.; and Sim, I. 2008. Automated information extraction of key trial design elements from clinical trial publications. In *AMIA Annual Symposium Proceedings 2008*, 141–145.

Dickersin, K., and Rennie, D. 2003. Registering clinical trials. *Journal of the American Medical Informatics Association* 290(4):516–523.

Doan, A.; Ramakrishnan, R.; and Halevy, A. Y. 2011. Crowdsourcing systems on the world-wide web. *Communications of the ACM* 54(4):86.

Fowler, M. 2005. Event sourcing. In *Development of Further Patterns of Enterprise Application Architecture*.

Fridsma, D. B.; Evans, J.; Hastak, S.; and Mead, C. N. 2008. The BRIDG project: A technical report. *Journal of the American Medical Informatics Association* 15(2):130–137.

Giardine, B.; Borg, J.; Higgs, D. R.; Peterson, K. R.; Philipsen, S.; Maglott, D.; Singleton, B. K.; Anstee, D. J.; Basak, A. N.; Clark, B.; and et al. 2011. Systematic documentation and analysis of human genetic variation in hemoglobinopathies using the microattribution approach. *Nature Genetics* 43(4):295–301.

Higgins, J., and Green, S., eds. 2009. *Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.2 [updated September 2009]*. The Cochrane Collaboration. Available from http://www.cochrane-handbook.org.

Ip, S.; Hadar, N.; Keefe, S.; Parkin, C.; Iovin, R.; Balk, E. M.; and Lau, J. 2012. A web-based archive of systematic review data. *Systematic reviews* 1:15.

Kalfoglou, Y., and Schorlemmer, M. 2003. Ontology mapping: the state of the art. *The Knowledge Engineering Review* 18(1):1–31.

Kim, S.; Martinez, D.; Cavedon, L.; and Yencken, L. 2011. Automatic classification of sentences to support evidence based medicine. *BMC Bioinformatics* 12(Suppl 2):S5.

Kiritchenko, S.; de Bruijn, B.; Carini, S.; Martin, J.; and Sim, I. 2010. ExaCT: automatic extraction of clinical trial characteristics from journal publications. *BMC Medical Informatics and Decision Making* 10:56.

Kong, Y. M.; Dahlke, C.; Xiang, Q.; Qian, Y.; Karp, D.; and Scheuermann, R. H. 2011. Toward an ontology-based framework for clinical research databases. *Journal of Biomedical Informatics* 44(1):48–58.

Leucht, S.; Cipriani, A.; Spineli, L.; Mavridis, D.; rey, D.; Richter, F.; Samara, M.; Barbui, C.; Engel, R. R.; Geddes, J. R.; Kissling, W.; Stapf, M. P.; Lssig, B.; Salanti, G.; and Davis, J. M. 2013. Comparative efficacy and tolerability of 15 antipsychotic drugs in schizophrenia: a multiple-treatments meta-analysis. *The Lancet* 382(9896):951–962.

Moher, D.; Liberati, A.; Tetzlaff, J.; Altman, D. G.; and The PRISMA group. 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Medicine* 6(7):e1000097.

Mons, B.; van Haagen, H.; Chichester, C.; 't Hoen, P.-B.; den Dunnen, J. T.; van Ommen, G.; van Mulligen, E.; Singh, B.; Hooft, R.; Roos, M.; and et al. 2011. The value of data. *Nature Genetics* 43(4):281–283.

Orwin, R. G., and Vevea, J. L. 2009. Evaluating coding decisions. In Cooper, H.; Hedges, L. V.; and Valentine, J. C., eds., *The handbook of research synthesis and meta-analysis*. New York, NY, USA: Sage, 2nd edition. 177–203.

Patrinos, G. P.; Cooper, D. N.; van Mulligen, E.; Gkantouna, V.; Tzimas, G.; Tatum, Z.; Schultes, E.; Roos, M.; and Mons, B. 2012. Microattribution and nanopublication as means to incentivize the placement of human genome variation data into the public domain. *Human Mutation* 33(11):1503–1512.

Rahm, E., and Bernstein, P. a. 2001. A survey of approaches to automatic schema matching. *The VLDB Journal* 10(4):334–350.

Ram, K. 2013. Git can facilitate greater reproducibility and increased transparency in science. *Source code for biology and medicine* 8(1):7.

Ramchurn, S. D.; Huynh, D.; and Jennings, N. R. 2004. Trust in multi-agent systems. *The Knowledge Engineering Review* 19(1):1–25.

Richardson, M., and Domingos, P. 2003. Building large knowledge bases by mass collaboration. In *Proceedings of the 2nd international conference on Knowledge capture*, 129–137.

Shapiro, M.; Preguiça, N.; Baquero, C.; and Zawirski, M. 2011. A comprehensive study of convergent and commutative replicated data types. Rapport de recherche RR-7506, INRIA.

Shemilt, I.; Simon, A.; Hollands, G. J.; Marteau, T. M.; Ogilvie, D.; OMara-Eves, A.; Kelly, M. P.; and Thomas, J. 2014. Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews. *Research Synthesis Methods* 5(1):31–49.

Shvaiko, P., and Euzenat, J. 2005. A survey of schema-based matching approaches. In Spaccapietra, S., ed., *Journal on Data Semantics IV*, volume 3730 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg. 146–171.

Sim, I., and Detmer, D. 2005. Beyond trial registration: A global trial bank for clinical trial reporting. *PLoS Medicine* 2(11):1090–1092.

Sim, I.; Owens, D. K.; Lavori, P. W.; and Rennels, G. D. 2000. Electronic trial banks: A complementary method for reporting randomized trials. *Medical Decision Making* 20(4):440–450.

van Valkenhoef, G.; Tervonen, T.; de Brock, B.; and Hillege, H. 2012. Deficiencies in the transfer and availability of clinical evidence in drug development and regulation. *BMC Medical Informatics and Decision Making* 12:95.

van Valkenhoef, G.; Tervonen, T.; Zwinkels, T.; de Brock, B.; and Hillege, H. 2013. ADDIS: a decision support system for evidence-based medicine. *Decision Support Systems* 55(2):459–475.

Wallace, B. C.; Small, K.; Brodley, C. E.; and Trikalinos, T. A. 2010. Active learning for biomedical citation screening. *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 10*.

Wallace, B. C.; Small, K.; Brodley, C. E.; Lau, J.; and Trikalinos, T. A. 2011. Deploying an interactive machine learning system in an evidence-based practice center: Abstrackr. In *ACM SIGHIT International Health Informatics Symposium (IHI)*.

Weld, D. S.; Wu, F.; Adar, E.; Amershi, S.; Fogarty, J.; Hoffmann, R.; Patel, K.; and Skinner, M. 2008. Intelligence in wikipedia. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, 1609–1614.