Automating Meta-Analyses of Randomized Clinical Trials: A First Look

Matthew Michelson
InferLink Corporation
2361 Rosecrans Avenue, Suite 348
El Segundo, California 90245

Abstract

A "meta-study" or "meta-analysis" analyzes multiple medical studies related to the same disease, treatment protocol, and outcome measurement to identify if there is an overall effect or not (e.g., treatment induces remission or causes adverse effects). It's advantage lies in the pooling and analysis of results across independent studies, which increases the population size, mitigates some experimental bias or inconsistent results from a single study, etc. Meta-studies are important for understanding the effectiveness (or not) of treatment, influencing clinical guidelines and for spurring new research directions. However, meta-studies are extremely time consuming to construct by hand and keep updated with the latest results. This limits both their breadth of coverage (since researchers will only invest the time for diseases they are interested in) and their practically. Yet, high-quality medical research is increasing at a staggering rate, and there is an opportunity to apply automation to this increasing body of knowledge, thereby expanding the benefits of meta-studies to (theoretically) all diseases and treatment, as they are published. That is, we envision, long term an automatic process for creating meta-studies across all diseases and treatments, and keeping those meta-studies up-to-date automatically. In this paper we demonstrate that there is potential to perform this task, point out future research directions to make this so, and, hopefully, spur significant interest in this compelling and important research direction at the intersection of medical research and machine learning.

Introduction

A meta-analysis (or "meta-study") collects and analyzes the results from multiple studies that are all focused on the same disease, treatment and primary outcome to determine if there is an overall beneficial effect of some treatment (or not). Meta-studies can can confirm (or refute) the overall effect across the studies, lead to changes in clinical guidelines, or spur new directions for research.

However, meta-studies are currently constructed by hand. This is an extremely time consuming process that starts with a comprehensive search of the literature, followed by compiling and filtering the results, and lastly performing statistical analysis. In fact, the sheer scope of the manual effort involved causes two fundamental challenges in widely applying meta-analysis to medicine in general. First, many topics are left unexplored, either due to the lack of researcher interest or lack of time to produce the review. That is, the sheer scope of time required may outweigh the interest in every possible disease and outcome. Second, meta-analyses are often not updated to reflect the latest results and studies. Rather than being a dynamic report that changes with the results, they reflect only a snapshot in time, up to the point when the review was produced. To underscore the sheer commitment involved when producing high-quality reviews of multiple studies, the Cochrane Collaboration, a volunteer-based organization that publishes systematic reviews, leverages a volunteer workforce of over 25,000 people (as of 2011). Finally, a number of (unknown) biases, via subjective choices during the meta-study, may influence the results.

While daunting to produce, meta-analyses are nonetheless extremely important. Therefore, our long-term goal is to automate, as much as possible, the meta-analysis process. This should greatly reduce human bias; increase the dissemination of evidence, especially for diseases and interventions with less focused attention; and allow for the automatic updating of meta-studies as new results are published. We envision an automated, computational approach that generates meta-studies and keeps them up-to-date. The system will constantly scour the literature and clinical trial databases, pulling out the freshest results it can find, grouping them appropriately (while excluding those that dont seem to be high enough quality), and updating all of the appropriate meta-studies as necessary.

In fact, the key tasks in the meta-study correspond to well to different techniques in machine learning. The task of searching and aggregating the literature by disease, intervention and primary outcome is essentially a "clustering" task that should learn to automatically group the papers together based on the same sets of features (Xu, Wunsch, and others 2005). Compiling the results from the different papers corresponds to to "information extraction" (Sarawagi 2008) where a machine can learn (either by example or from statistics about the content) how to extract the outcomes of trials or the population sizes from the sentences in the paper that describe them. The final step in the meta-analysis is a statistical analysis, involving tasks such as funnel plot analysis or applying random-effects models, which are tasks well suited to computation.

In this paper we present some of our very early work towards these goals. Specifically, we focus on a particular type of meta-study: applying a random-effects model to the outcomes of randomized controlled trials (e.g., a "control" group's effect versus a population given treatment). We chose to focus on random-effects because these models can account for the heterogeneity when applied to understanding the effect across clinical trials (DerSimonian and Kacker 2007). The intuition here is that in each trial, the effect is measured in the control group versus the treatment, and the random-effects model determines, taking into account variability across the trials, whether this effect holds true across the trials. For this early work, we use the classic Paule-Mandel random effects model (Paule and Mandel 1982).

Our results demonstrate the potential in automating these steps. While the work is very early, we hope we can ignite research interest in this important and interesting topic, especially as it blends the worlds of medical research and machine learning. We note that the process of automating meta-studies has received attention lately (see Tsafnat et al. (2014) for recent survey). However, as the survey shows, while individual pieces (such as extraction) received attention for automation, we intend to spur interest in automating the whole, integrated process and our results are toward that end (though clearly not there yet).

Automating Meta-Studies

Overall, the intuition behind our process is to turn papers (abstracts) describing the results of clinical trials into structured data that we can then analyze using the Paule-Mandel random-effects model. This model will then tell us if there is an appreciable overall effect for the treatment or not. Therefore, our main tasks in automating meta-studies fall into two categories. First, we take the natural language description of the results from the papers, and perform information extraction to turn this into structured data that can be processed. Second, we then group together all of the studies that apply the same treatment to the same disease for the same primary outcome. This grouped set of structured data then becomes the input to the random effects model which outputs whether there is a meaningful effect (or not) across the trials. Figure 1 shows this process.

We now describe each of these higher level tasks in more detail.

Information extraction to structure results

We start by describing how information extraction can turn the results described in a technical article into structured data. The random effects model we automate here is based on a log-odds ratio between the control population and the treatment population. Therefore, the goal of our extraction module is to gather the results from the papers and convert them into a format that is appropriate for this type of computation. This means that data for (at least) two groups must be gathered: one set of data for the treatment group and one set of data for the control group. For each of these groups, we need to extract the overall population size and the size that demonstrated the effect of the treatment. Finally, we need to be sure that the effect is for the primary outcome.

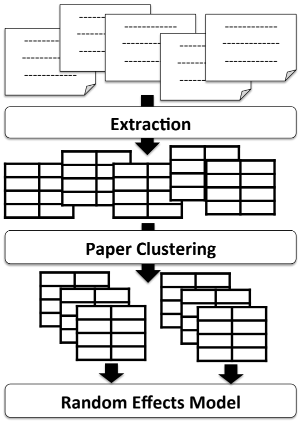

Figure 1: Architecture for meta-study automation

This task is a combined extraction and classification task. We do not explicitly demarcate which sentences contain "results." Rather, data is extracted and then a small classifier is used to determine whether (a) the sentence involves a primary outcome (e.g., it is a result) and (b) whether it pertains to the control group or the treatment group. Sample results are given in Table 1.

To begin the process, we extract values from sentences using a number of patterns that identify results for odds ratio computations. For instance, we looked for patterns of the form "X of (the) Y," where "X" is the numerator in a reported ratio, "the" is an optional word, and "Y" is the denominator. Other patterns include "X/Y" (as shown in the table), or "X% (n=Z)" which then requires that we compute the number in an affected group based on this inference.[1] For instance, if we see a sentence "reported 33% effect (N=12)" then we can infer that the affected group has a size of 4. Once a paper has been processed in this manner, we end up with records as shown in the table.

We then assigned the extracted results to either the control group or the treatment group, by analyzing textual clues around the extracted results. To do this, we look for the words "control" or "placebo" and then assign them to the result which they are closest to. In Table 1, in the first row, the result 9 of 36 is closer to "placebo" than the treatment name and so it is assigned to that value. As with the above extraction, this is currently a heuristic process, but we intend to eventually learn this assignment.

Combining similar results using clustering

Once the results are extracted, the papers are then grouped together such that studies for the same treatment, for the

[1] Note, for this early work, we only include the case where the value of n is in the same sentence as the reported percentage. Handling the harder cases, where the n in a sentence must be linked to a previous sentence are currently part of our future work.

Table 1: Example sentences and extracted results

	Treatment Group		Control Group	
Sentence	**# Affected**	**Total**	**# Affected**	**Total**
At the end of the trial, 9 of 36 patients administered placebo and 14 of 48 administered mesalamine were in remission	14	48	9	36
For the study using 15 mg/week of oral methotrexate 33% (5/15) of methotrexate patients failed to enter remission compared to 11% (2/18) of placebo patients	5	15	2	18

same disease and outcome can be compared within the meta-analysis. For this study, we employed a simple clustering method called greedy clustering, using Jaccard similarity as our metric. Greedy clustering defines within a cluster any two members that share some value for the similarity metric that is above a threshold. The Jaccard similarity is computed as the ratio between the words in common over the unique set of all words across the two studies (Jaccard 1901). We set our threshold at 0.1252, which is the similarity value one standard deviation above the mean, as computed across all pairs of abstract titles. Therefore, if the similarity between two titles is at least one standard deviation from the mean, we group the titles together to build clusters. While we mined this threshold directly from the data (and therefore likely over fit it such that it would not generalize well outside of this study), such an approach is less biased for these small studies than choosing and tuning a value by hand.

Once we have the extracted values, and they are clustered by disease, treatment and outcome, we then pass the structured results through our Paule-Mandel random effects model. Again, we emphasize that while our work is quite early, focusing mostly on heuristic methods, there is significant room to improve the approach using rather standard methods from machine learning. For instance, we intend to learn the extraction models, rather than define them manually, and we will improve clustering using more advanced techniques. However, in this paper we intend to push the idea forward by demonstrating that pieces are at least possible, so this important task can begin to benefit from the automation that machine learning can provide.

Experiments and Results

As we mentioned above, the work is quite preliminary, and our motivation in this paper is to present the possibility that meta-studies can be automated to help spur research interest in this important area. To that end, we ran a small pilot study, using our approach, and report those results here.

For our study we chose a specific meta-study that focused on treatment for Inflammatory Bowel Disease (Camma et al. 1997), with the intention that if these studies were all grouped together (e.g., clustering works perfectly), would we be able to extract the data from them to build the meta-study. The meta-study references 15 other studies, 13 of which contained freely available online abstracts that a computer could access (to mimic an automatic harvesting process).[2]

Using our information extraction approach we were able to extract the results from four of the abstracts (26.7% of the original 15). All of the extractions were manually verified as correctly reflecting the reported treatment ratios (yielding an accuracy of 100%), though only on four studies. In the metrics of information extraction, for this (albeit) tiny study, our precision is 100% and our recall is 26.7%. The studies where we failed to extract the data used other patterns, such as presenting the results as percentages with an associated population value in different sentences. This motivates a machine learning approach that will learn how to generate such patterns, covering many different types of representations in the text.

Two interesting points came out of even this small study. The endpoint focus of the meta-analysis was remission for Crohn's disease. Of the four papers for which we could extract results, three reported the results as proportions of patients in remission. However, one of the studies reported the results as the proportion that relapsed, rather than remission. Therefore, the system will need to learn that if the endpoint is remission, and the results are reported as relapse, then they cannot be directly grouped together for analysis without some transformation (e.g., perhaps treating relapse as the inverse of remission). That is, there is some subtlety required to extracting the pertinent results and outcomes. Again, this points to overall difficulty of this problem.

Another interesting result that points in this direction is that for one of the papers, the system extracted two sets of results, both correct. However, each focused on a slightly different population in the study. One reported the results for the whole population (our desired results), while the other was only for Crohn's patients with ileal-based disease. Again, this points to the notion that the system will need to have some sense of which results to group together when it generates a meta-analysis. Further, it motivates our idea that researchers should be able to change and interact with the data themselves, to mitigate such subtleties that may be beyond the systems capabilities.

Next, we performed a simple study to determine whether it would be possible to automatically group these relevant abstracts together from a larger set. Here, we took the 15 Crohn's studies and combined them with 15 studies focused on lupus (again remission was the endpoint). In this study,

[2] This study focused on processing abstracts, rather than whole papers, since it made the extraction easier.

the classes are balanced and the goal is to determine that once we mix together all of the abstract titles, the system can separate them again. This is a proxy study for a true clustering experiment.

Indeed, with our clustering approach, using only the abstract titles, we end up with 3 clusters of studies. One cluster contains all of the Crohn's studies, and only the Crohn's studies. This is an ideal cluster. The other clusters are more interesting. One cluster contains 14 of the 15 lupus studies, while the third cluster only contains one of the lupus papers. Upon further analysis the paper not included in the large lupus cluster has a title that is 36.7% longer (in number of words), than the other two clusters titles average length of words. This is known deficiency with the simple Jaccard metric: it can be sensitive to the length of the input text it receives, and we are seeing this behavior here. As with the extraction study, although this was a small, focused study it motivates more sophisticated clustering methods that should be able to appropriately select and group the studies, even based solely on the titles.

Future Research Directions

As our results and approach demonstrate, while it may be feasible to automate meta-studies, there is still a significant amount of research to be done in this area. Here we outline some challenging future research directions that are relevant to this topic, aligned with different aspects of the meta-study process.

Improved information extraction

We initially demonstrate that linguistic patterns can extract some data and structure into a more natural format for comparison (including normalizing the data from percent to value, etc.). There is absolutely value in pursing this path of research. We believe that unsupervised approaches to extraction could perform well in this domain, where simple patterns can be used to bootstrap data collection and then those can be used as input for machine learning algorithms to train themselves. This will also help deal with the scale problem associated with the massive, and growing, sets of published medical literature.

While core extraction of results is an interesting path, there are also other new challenges to turn medical studies into structured data for processing. Specifically, beyond the numerics in the results of the study, there are certain pieces of information about a study, such as participant information, that also need to be extracted and normalized in order to be compared across studies. For instance, the information about the cohort participants, such as the age range, demographics, medical history (e.g., smokers versus non-smokers) also needs to be extracted and normalized. Perhaps more difficult than the extraction itself will be the ability to normalize this type of information into standard categorical values.

Improved clustering of results

As we improve the clustering, there are a number of improvements required to group the results into appropriate sets for meta-analysis. First and foremost, the challenge of scale will need to be addressed. Once we can perform extraction from all of the published studies, grouping them by similarity will be a massive challenge for scale and efficiency. Further, general similarity will need to be improved, for instance, as alluded to above, resolving some of the more subtle differences in outcomes of studies that may seem similar (e.g., remission versus symptom re-emergence).

However, a potentially even more important task involves grouping the studies together using deeper levels of sophistication than presented here. Specifically, building upon the ability to extract information about the study participants, the clustering should take these differences and similarities across populations into account. For instance, some meta-studies only focus on adverse events in pediatric use. Therefore, the system should be able to group all adverse effect studies together, and further, build a sub-cluster for analysis that focuses on adverse events in pediatric use or for users with specific medical histories (e.g., those with diabetes and those without). Along these lines, meta-studies often look at similarities in the "methods" section for including studies in the meta-analysis (or not). For instance, differences or similarities among the populations or their inclusion criteria may determine which papers should be clustered together or not. This requires more sophisticated extraction and clustering and is an important future research direction.

Improved meta-analysis

Finally, we believe there are a number of interesting avenues to pursue with respect to core analysis. For instance, once the results are consumable by a machine, multiple meta-analyses could be run simultaneously to re-affirm (or dispute) one another. For instance, some techniques are better than others for certain cases in the data, and therefore multiple tests, which might be difficult to do by hand, could be done in volume. A machine could even learn to pick which type of statistical analysis is most appropriate given the underlying study data it is being passed.

Conclusion

Our aim in this paper is spark research interest in the automation of meta-studies. Meta-studies are an increasingly important tool for medical researcher to uncover valuable insights across the myriad of studies that are related, but independent from one another. Here we outline the tasks, which if accomplished, can begin to automate this manual and complicated task. We showed early possibility that some of this work has potential for automation, but there is still much work to do. Yet, if we, as machine learning researchers, can automate this process, bridging it to the scale of medical literature, we can potentially uncover novel therapies personalized to a group, improve the timeliness of clinical guidelines and even spur new directions for medical research.

References

Camma, C.; Giunta, M.; Rosselli, M.; and Cottone, M. 1997. Mesalamine in the maintenance treatment of crohn's disease: a meta-analysis adjusted for confounding variables. *Gastroenterology* 113(5):1465–1473.

DerSimonian, R., and Kacker, R. 2007. Random-effects model for meta-analysis of clinical trials: An update. *Contemporary Clinical Trials* 28:105–114.

Jaccard, P. 1901. Etude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Societe Vaudoise des Sciences Naturelles* 37:547–579.

Paule, R. C., and Mandel, J. 1982. Consensus values and weighting factors. *J Res Natl Bur Stand* 87:377–385.

Sarawagi, S. 2008. Information extraction. *Foundations and trends in databases* 1(3):261–377.

Tsafnat, G.; Glasziou, P.; Choong, M.; Dunn, A.; Galgani, F.; and Coiera, E. 2014. Systematic review automation technologies. *Systematic Reviews* 3(1):74.

Xu, R.; Wunsch, D.; et al. 2005. Survey of clustering algorithms. *Neural Networks, IEEE Transactions on* 16(3):645–678.