Automated Extraction of Reported Statistical Analyses: Towards a Logical Representation of Clinical Trial Literature

William Hsu[1], PhD, William Speier[1,2], MS, Ricky K. Taira[1], PhD
[1]Medical Imaging Informatics Group, Dept of Radiological Sciences
[2]Biomedical Engineering Interdepartmental Program
University of California, Los Angeles, CA

Abstract

Randomized controlled trials are an important source of evidence for guiding clinical decisions when treating a patient. However, given the large number of studies and their variability in quality, determining how to summarize reported results and formalize them as part of practice guidelines continues to be a challenge. We have developed a set of information extraction and annotation tools to automate the identification of key information from papers related to the hypothesis, sample size, statistical test, confidence interval, significance level, and conclusions. We adapted the Automated Sequence Annotation Pipeline to map extracted phrases to relevant knowledge sources. We trained and tested our system on a corpus of 42 full-text articles related to chemotherapy of non-small cell lung cancer. On our test set of 7 papers, we obtained an overall precision of 86%, recall of 78%, and an F-score of 0.82 for classifying sentences. This work represents our efforts towards utilizing this information for quality assessment, meta-analysis, and modeling.

Introduction

Randomized controlled trials (RCT) represent the most reliable source for elucidating causal relationships between treatments and outcomes by enforcing strict constraints on the study population and methodology. RCTs represent a significant source of evidence with over 26,000 papers being indexed on PubMed in 2011 alone[1]. While a significant amount of time and monetary resources have been spent conducting these trials, translating the derived information into knowledge that can be applied at the patient bedside remains a significant challenge. One important issue is the difficulty in placing the results of a single trial in the context of a global picture that incorporates all studies conducted for a given disease. A single study may contribute one causal relationship that needs to somehow integrate with a broader understanding of the complex causal chain and interactions of a disease process. Currently, no methodical approach exists for creating a knowledge source that integrates evidence from papers in a principled manner. Another issue relates to the inconsistent quality of conclusions being reported in literature. Ioannidis [1] argues that false findings are becoming more prevalent given the misuse of significance testing, inherent biases in the study population, and lack of statistical power. Several authors have also noted significant increases in retraction rates over the last decade [2, 3]. While journal editors and reviewers do their best to filter out studies with poorly described results or faulty experimental design, erroneous conclusions due to faulty statistical analyses continue to elude the review process and become published [4]. Assessing the validity and quality of the presented evidence requires the consideration of what variables are measured, how they are measured, what constraints were imposed on the experiment, estimated distributions of the data, and choice of statistical test and significance value. This information is primarily published as a free-text communication as part of the results section of a paper, which compounds the problem because free-text is inherently ambiguous and fraught with imprecision. Some papers summarize the results in a table, but the reader still needs to refer to the free-text to understand the content of what is being presented. Ultimately, the burden is on the reader to assess, weigh, and utilize available evidence to determine the best course of action for a patient's condition.

The overall objective of this work is to develop an automated pipeline that extracts information related to the statistical analysis from full-text RCT literature, mapping this information to a logical data model. Towards this objective, three goals are defined: 1) transform free-text descriptions of the hypothesis, statistical analysis, and clinical applicability into a logical representation; 2) build modules to identify and annotate variables, attributes, values, and their relationships from sentences; and 3) develop a workstation to assist in the validation of the generated results. This work is complementary to the multitude of prior and current efforts to create standardized reporting of clinical trial results: Global Trial Bank [5], a standardization of how clinical trial protocols are reported,

[1] Based on http://dan.corlan.net/medline-trend.html using query 'randomized and controlled and trial'

Consolidated Standards of Reporting Trials (CONSORT) statement [6], which defines a set of guidelines to aid RCT authors in deciding on what to report, ClinicalTrials.gov [7], a repository of semi-structured and standardized protocols of clinical trials, NeuroScholar [8], a framework for integrating neurology evidence captured from multiple papers, and EliXR [9], an automated system for extracting temporal constraints from eligibility criteria. Despite these efforts, little progress has been made towards formalizing the representation of statistical analyses reported in papers. While many systems attempt to summarize results at a high level (*i.e.*, classifying whether a sentence is pertinent to the result), few attempt to capture information from papers at a level of detail necessary to facilitate Bayesian analysis. Our system is a step towards automating the identification of key reported statistical findings that would contribute to the development of a Bayesian model of a complex disease.

Background

Evidence-based medicine (EBM) requires clinicians to make "conscientious, explicit, and judicious use of the current best evidence" in their everyday practice [10]. However, understanding how to apply EBM to answer a clinical question may not be straightforward. One mnemonic developed to assist in applying EBM is: Patient, Intervention, Comparison, and Outcome (PICO) [11]. Physicians are first asked to characterize the individual or population being examined and are then asked to define the intervention or therapy being considered. Next, they are asked to consider alternative treatments and the desired clinical outcome. To facilitate the utilization of PICO, several groups have examined how to divide a clinical question into its component parts and then apply machine-learning techniques to perform information retrieval to answer the question [12]. Researchers have studied the task of automating the extraction of various trial characteristics such as eligibility criteria, sample size, enrollment dates, experimental and control treatment: Hansen et al [13] describe an approach for extracting the number of trial participants from the abstract by classifying each sentence using a binary supervised classification based on a support vector machine algorithm achieving an F-score of 0.86. Chung et al [14] utilize conditional random fields to classify abstract sentences as being related to either intervention or outcome, achieving an accuracy of 95% and F-score from 0.77 to 0.85 depending on task. Boudin et al [15] use an ensemble approach to classify PICO elements, achieving an F-score of 0.863 for participant-related sentences, 0.67 for interventions, and 0.566 for outcomes.

While these initial efforts have demonstrated the ability to automatically classify information from literature to facilitate information retrieval, several limitations exist. First, these works focused on processing Medline abstracts, but as Blake [16] notes, only 8% of scientific claims made in a paper are done so in the abstract, emphasizing the need to examine the entire paper rather than relying on the abstract alone. Second, while sentence classification is useful to assist in identifying relevant papers, interpreting this information and determining the quality of the information is still left to the reader. None of these works have developed a data model for the information being captured to ensure that all necessary contexts for interpreting the information are captured. Finally, most works remain at the sentence level; contents of these sentences such as variables are not explicitly characterized and linked with the analyses performed. In this paper, we attempt to not only classify sentences related to the statistical analyses, but also characterize the values reported in these sentences to populate the data model. This allows the computer to assist in assessing the validity of reported information and enables this information to be used for meta-analysis and probabilistic disease modeling. In the following sections, we describe the development of our logical representation for capturing reported hypothesis, statistical analyses, and conclusions. We then present our framework for extracting and annotating information from RCT papers.

We demonstrate our system in the domain of non-small cell lung cancer (NSCLC), which is the most common form of lung cancer. NSCLC was chosen as the domain given its phenomenological complexity, the numerous physical properties that have been reported (*e.g.*, radiologic, genetic, molecular pathways, clinical findings), and the numerous interventions that have been used to treat the disease.

Modeling Published Literature

This project is a step towards our overall objective to develop a process model that formally captures key information from full-text RCT papers such the hypothesis, experimental design, data collection process, analyses performed, and conclusions drawn. The goal is to create an expressive, robust, and flexible system for capturing all forms of evidence reported in scientific literature. Ultimately, we desire a representation that permits: 1) traditional information retrieval tasks such as returning studies based on primary outcome measures; 2) critical appraisal of the paper by clearly delineating reported (and missing) information; and 3) aggregation of results from multiple trials by accounting for differences between study populations. In this paper, we focus primarily on creating a logical representation of the statistical analysis and how it relates to the study hypothesis and clinical conclusions. The representation is comprised of seven classes: 1) paper (*e.g.*, author names, article title, MeSH terms), 2) hypothesis

(*i.e.*, statement conveying the objectives of the study), 3) arm (*e.g.*, description of each arm of the study, sample size of each participant arm/group) 4) statistical method (*e.g.*, the hypothesis test used, power level); 5) result (*e.g.*, statement describing statistical significance, p-level); 6) interpretation (any generalizable statements based on study results that can be applied to clinical practice); and 7) variable (*e.g.*, outcome measure). Figure 1 depicts the logical representation of the aforementioned classes using an entity-relationship diagram.

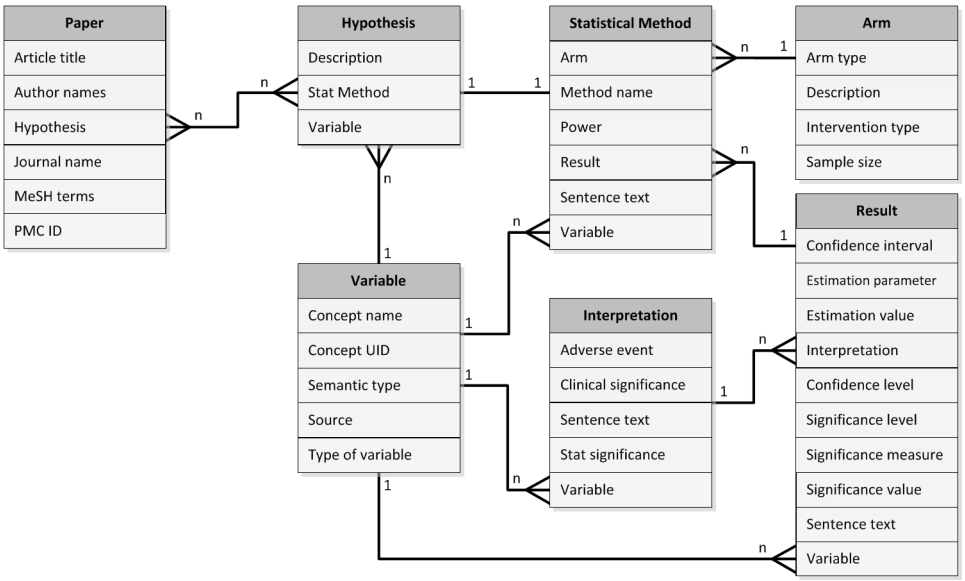

Figure 1: The logical representation depicted as an entity-relationship diagram. Relationships between classes are denoted as 1:1 (one-to-one), 1:n (one-to-many), or n:n (many-to-many).

System Architecture

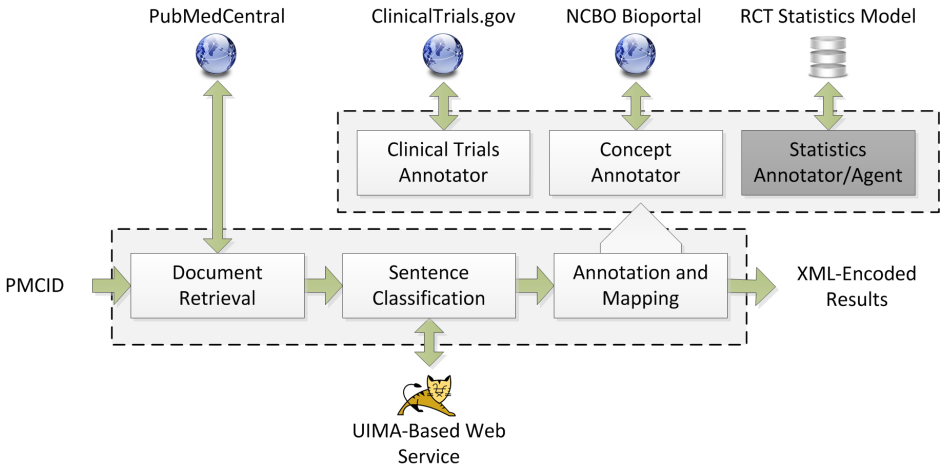

Figure 2: The overall architecture of the system. The user provides the input (PubMedCentral identifier, PMCID) and reviews the output through the validation workstation. The dotted lines represent components that are implemented in the Automated Sequence Annotation Pipeline. Grey rectangles represent annotators implemented as unique plans; the dark gray rectangle represents an annotator implemented as both a plan and agent.

The system process and information flow are summarized in Figure 2. Briefly, the process involves retrieving the body of the paper, identifying relevant sentences and variables, annotating variables and incorporating other

metadata, and mapping the extracted and annotated information to the logical representation. A validation workstation has been implemented that allows users to view the paper with identified sentences and annotations overlaid. The following sections describe each component in detail.

Document corpus

The process begins when a user selects a paper of interest though the validation tool. As it would be an intractable task to consider all NSCLC-related literature, we limited our initial analysis to full-text papers of RCT studies related to NSCLC and chemotherapy treatments. Utilizing the Open Access subset of PubMedCentral, we performed an initial search of relevant papers using keywords "non-small cell lung cancer OR NSCLC", "chemotherapy", "randomized controlled trial", and "NOT review", resulting in 51 papers. The abstract of each paper was then manually inspected to ensure that the paper was indeed reporting results of an RCT study. A total of 42 papers were considered. The corpus was then split into two parts: 35 papers (~80%) were used to create our representation and classifiers and the remaining 7 papers (~20%) were set aside for the evaluation.

Automated Sequence Annotation Pipeline

The automated sequence annotation pipeline (ASAP) provides an interface for querying biomedical knowledge sources and integrating the results [17]. The system provides two methods for accessing this data: 1) plans, which wrap external data sources accessed through the web, and 2) agents, which create a local copy of a data source that is updated periodically. Users can then create custom plans that successively query these data sources and link their results in a pipeline. Although ASAP was originally designed for genetic sequencing data, the framework is flexible enough to extend to other clinical or research data sources. Custom plans and agents are written in Perl and integrated into the ASAP project. New plans are then able to access databases created by agents and also have the freedom to call other existing plans. Programmatic interaction with ASAP is achieved through integration with the distributed annotation system (DAS) [18]. ASAP can be queried using standard web request protocols, automatically creating and executing the desired job. Output values and intermediate files generated from querying external sources are encoded in eXtensible Markup Language (XML) based on the DAS protocol [19].

In this work, custom plans and agents for ASAP have been written to retrieve information from data sources such as PubMedCentral and map extracted terms to relevant databases and biomedical ontologies. Depending on the type of research article being analyzed, different plans can be executed dynamically. For example, a paper on the molecular mechanisms of a chemotherapy drug would be mapped to knowledge sources such as Gene Ontology versus a paper on the imaging correlates of treatment response, which would be mapped to sources such as RadLex. Here, we demonstrate ASAP's ability to serialize multiple plans into a pipeline: the input (*e.g.*, PMCID) is passed to the parent plan, which references additional plans to perform the requested processing steps; the results of each referenced plan are then aggregated by the parent plan and outputted as a uniform XML representation.

Document retrieval

We have written an ASAP plan that interfaces with the PubMedCentral Open Access web service: given a PMCID, the plan submits a request to the web service and receives a response with information about the paper encoded in XML format. While the XML contains information about the entire paper (*e.g.*, text, tables, references), for the scope of this work, only text related to the body of the paper is examined.

Sentence classification

Once the XML representing the full-text paper is returned, a second ASAP plan is executed that calls a web service for our classification module. The module is built upon the Unstructured Information Management Application (UIMA) framework originally developed by IBM [20] and wrapped as a servlet. UIMA is becoming broadly utilized for natural language processing tasks in the biomedical domain, having recently been implemented in systems such as Mayo Clinic's cTAKES [21]. The classification module is comprised of a sentence boundary detector, regular expressions, parts-of-speech tagger, and a dictionary lookup module. To address the variability in how papers are organized, a rule-based approach is used to reclassify the original headings into one of five section categories defined by CONSORT: abstract, introduction, methods, results, and discussion. Each section is then further tokenized into sentences. A set of regular expressions is used to further classify sentences into specific topics based on their content. For example, a subset of sentences in the abstract or introduction sections can be categorized as being part of the study objectives or hypotheses. A sentence in the methods section could be categorized as describing statistical methods. Based on the topic, a sentence may be processed using one or more additional annotators to extract variables, attributes, and values. For example, sentences categorized under the topic 'outcomes

and estimation' will be processed using regular expression patterns that identify p-values, confidence intervals, and statistical interpretation (*e.g.*, no significant difference). Other sentences, such as ones categorized under the topic 'outcomes' are processed using the OpenNLP implementation of a maximum entropy-based parts-of-speech tagger to extract noun phrases, which are then tagged for additional processing in the annotation step. The output of the classification module is encoded in XML and contains the original sentence, section, topic, phrases/values that are tagged for annotation, and the name of the annotator to execute. A summary of topics and related annotators are given in Table 1.

Section	Topic	Annotator
Paper metadata	---	Clinical trials annotator (retrieve trial metadata)
Introduction	Hypothesis or objective	---
Methods	Outcomes	Parts of speech tagger (identify noun phrases) Concept annotator (identify biomedical concepts)
	Sample size	Regular expression (extract number of patients)
	Statistical methods	Parts of speech tagger (identify noun phrases) Statistics annotator (identify statistical tests)
Results	Outcomes and estimation	Regular expression (extract statistical interpretation, p-value, confidence intervals, comparison groups)
	Harms	Parts of speech tagger (identify noun phrases) Concept annotator (identify medical problems)
Discussion	Generalizability	Regular expression (classify sentences related to clinical applicability)

Table 1: A summary of section headings and topics. For each topic, one or more annotators is used to parse the sentence and extract information that is mapped to the logical representation.

Annotation and mapping

The results returned by the sentence classifier are encoded in an XML that specifies the category in which the sentence or token is classified. ASAP then executes additional plans based on the nature of the sentence or token to be annotated. For example, if the sentence category is hypothesis, the biomedical concept annotator is executed to identify potential variables. Below, we elaborate on each of the implemented plans.

- Concept annotator. The purpose of the concept annotator is to identify variables from the entire set of noun phrases. This annotator is implemented as an ASAP plan that interfaces with the National Center for Biomedical Ontology's BioPortal web service. Using BioPortal, noun phrases identified in the classification step are mapped to standardized concepts from one or more biomedical ontologies. Given our domain of interest, NSCLC, we initially map phrases to the National Cancer Institute Thesaurus (NCIt); however, other ontologies such as Gene Ontology may be added depending on the nature of the papers being analyzed. The result of the concept annotator is a list of matched concepts, their concept unique identifiers, and semantic types.
- Statistics annotator. The statistics annotator focuses on matching a noun phrase to a list of analysis techniques, such as hypothesis testing and survival analysis. This annotator is comprised of three components: an agent, a local database, and a plan. We created a list of possible statistical analysis methods by manually reviewing concepts in existing research ontologies such as the Ontology for Biomedical Investigations and Ontology of Clinical Research. The spreadsheet contains columns representing the first word of the method name, the full method name, a unique identifier, and a mapping to the source ontology. Using the agent, the spreadsheet is loaded as a table in the ASAP database. A plan was then written to search the database using the noun phrase as input using a dictionary lookup approach. Presently, the annotator returns the unique identifier associated with the statistical method; however, given additional information about each statistical method (*e.g.*, nature of independent/dependent variables, number of dependent variables), annotations could help determine whether the appropriate statistical method was used to analyze the data.
- Clinical trials annotator. The goal of the clinical trials annotator is to retrieve contextual information about the study from a public resource such as ClinicalTrials.gov. As of July 2012, over 129,000 trials conducted in 179 countries are indexed on the site. Each trial includes semi-structured information pertaining to study objectives, patient recruitment, interventions, primary outcomes, and results, whenever available. The annotator is implemented as an ASAP plan that interfaces with the ClinicalTrials.gov web service. The input for the plan is

the national clinical trial identifier that is typically provided in the DataBank element of the PubMed XML and/or abstract text. The output is information about the study encoded in XML.

Results from each plan are combined into an XML schema specified by DAS containing all of the sections, sentences, and annotations generated by the pipeline.

Validation Workstation

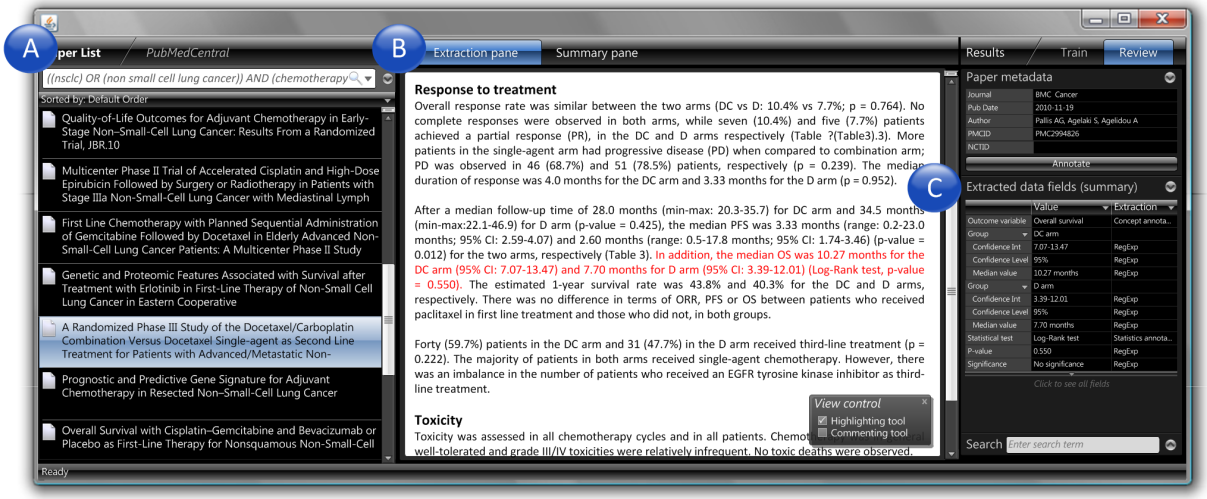

Figure 3: The validation workstation user interface. (A) A search pane provides functions to search for papers of interest and retrieve selected papers from PubMedCentral. (B) The text body of the paper is presented in the extraction pane. Results of the ASAP pipeline are shown as highlighted sentences, which can be selected to reveal extracted values. (C) A results pane presents a structured view of the extracted values.

A validation workstation has been implemented to assist researchers with viewing, annotating, and validating statistical information extracted from the paper. The workstation, illustrated in Figure 3, consists of the following features:

1. Search pane (Fig. 3A): The user first searches for a collection of papers in the area of interest (*e.g.*, non-small lung cancer, chemotherapies). A list of matching PMCIDs are returned and displayed as a selectable list.
2. Extraction pane (Fig. 3B): Once the user selects a specific paper to annotate, the body of the paper appears in the viewing pane. Metadata drawn from PubMed such as article title, authors, journal name, and MeSH terms are displayed alongside the selected paper. Manual annotation tools such as highlighting and commenting are provided.
3. Results pane (Fig. 3C): When the user selects "Annotate", the ASAP pipeline shown in Figure 2 is executed. Results are returned, parsed, and presented in the results pane. Extracted values are presented in a tabular view. Alongside the value is the rule or annotator that was used to extract the information. Users can correct the extracted values, if needed. This pane reflects the information captured in the logical representation for the selected paper.

Evaluation

Inter-rater agreement

Two informatics researchers were asked to manually annotate papers from the training set. Both individuals had previous experience performing systematic reviews of clinical literature and only needed minimal training on the types of information being sought in the data model. The annotators were asked to identify the minimal number of sentences that answer the following questions:

- What are the objectives or hypotheses of the study?
- What are the statistical methods? What are the primary/secondary endpoints? What is the total sample size? What are the sample sizes for each group? What statistical tests are used? What variables are measured? Are significance levels and/or confidence intervals reported?
- What conclusions (statistical/clinical significance) are reported about the study?

We examined the overall agreement between the two annotators and obtained a Cohen's kappa coefficient of 0.89, indicating that consistency between the two raters overall was achieved. While the vast majority of the paper could be reliably classified, the clinical conclusions annotation had the least amount of agreement, achieving a kappa score of 0.12. We believe the poor agreement between the two annotators spawns from the inherent ambiguity in defining what a conclusion was in the context of a paper. One annotator highlighted any sentence that was in the conclusion section of the paper while the other highlighted sentences that pertained specifically to the study's clinical relevance. Moving forward, we intend to utilize the CONSORT statement's definition of 'generalizability' to guide annotators with this part of the annotation task. We also plan to involve clinicians as part of the evaluation process to provide perspective on their information needs while reading scientific literature.

Classification

We performed an evaluation of the sentence classifier using precision, recall, and F-score as measures of performance. The test set consisted of seven papers with a total of 753 sentences. We primarily evaluated the sentence classification and value extraction tasks. For sentence classification, we validated that sentences classified being a hypothesis, statistical method, result, or discussion. For the value extraction task, we present results for extracting specific values related to the statistical test name and reported significance value. The results of the classification task are summarized in Table 2.

Topic	Example	Precision	Recall	F-Score
Hypothesis	…this phase III trial which was conducted in order to determine whether the combination of docetaxel/carboplatin provides any therapeutic benefit compared to single-agent docetaxel.	83%	91%	0.86
Statistical method	The primary endpoint of the study was to assess the overall response rate (ORR)...	95%	76%	0.84
Outcomes and estimation	…the median PFS was 3.33 months (range: 0.2-23.0 months; 95% CI: 2.59-4.07) and 2.60 months (range: 0.5-17.8 months; 95% CI: 1.74-3.46) (p-value = 0.012; Figure 1)…	93%	88%	0.90
Generalizability	The results of the current study demonstrate that second-line combination treatment with docetaxel/carboplatin offers a statistically significant therapeutic benefit compared to docetaxel monotherapy, in terms of PFS in patients with NSCLC…	63%	55%	0.59

Table 2: Summary of results after evaluating the test set of seven papers. Sentences in the example column are drawn from [22].

Our overall precision, recall, and F-score were 86%, 78%, and 0.82, which are comparable to systems performing similar tasks. While categories such as hypothesis and result had relatively limited number of variations in the way the information was expressed, the discussion category proved to be most difficult given the large variability seen in sentence structure and vocabulary used. As previously discussed, manual annotators had poor agreement with categorizing sentences in this section. Errors generated at the initial stages of the pipeline propagate through the subsequent tasks; thus, we are exploring robust, non rule-based approaches such as conditional random fields and support vector machines.

Annotation

We also started to evaluate how well values are being extracted from sentences. Table 3 provides examples of how sentences are parsed, annotated, and mapped to the logical representation. Values such as the calculated p- or r-value and confidence intervals have high precision and recall rates (92% and 95%, respectively), given that the variability in how these values are expressed is small. On the other hand, capturing the variability in how independent variables, dependent variables, and hypothesis tests are reported has been challenging. While certain variables such as outcome measures are well defined (*e.g.*, response rate, overall survival, progression-free survival), less standardized variables have been difficult to consistently identify (*e.g.*, abbreviations, variations of drug names). The mapping process identifies a large number of false positives, particularly when common words (*e.g.*, treatment) and acronyms are present in the sentence. Utilizing a part-of-speech tagger helps narrow the scope of words that are under consideration. Nevertheless, our current dictionary lookup approach performs poorly and is unable to handle comparisons between words with subtle differences.

Sentence	**Class**	**Attribute**
From July 2003 to December 2007, 561 patients were enrolled onto the study… 268 (95%) and 275 patients (98%) received celecoxib and placebo, respectively.	Arm	Total enrolled: 561 Arm 1: Celecoxib, n=268 Arm 2: Placebo, n=275
Median progression-free survival was 4.5 months (95% CI, 4.0 to 4.8) for the celecoxib arm and 4.0 months (95% CI, 3.6 to 4.9) for the placebo arm (hazard ratio [HR], 0.8; 95% CI, 0.6 to 1.1; P = .25).	Statistical Method	Comparison arm 1: Celecoxib Comparison arm 2: Placebo Method name: log rank test
	Result	Estimation parameter: Hazard ratio Estimation value: 0.8 Confidence interval: [0.6, 1.1] P-value: 0.25 Significance level: 0.05
	Variable	Progression-free survival
	Interpretation	Stat significance: Not significant

Table 3: Examples of how sentences from [23] are parsed, annotated, and mapped to the logical representation.

Discussion

We present an automated system for extracting, annotating, and mapping information related to statistical methods from RCT papers into a logical representation. While the focus of this paper is on representing statistical information, the overarching goal of our project is to create a logical representation of the entire trial, including the experimental design, data collection process, analysis methods, and conclusions drawn. Ultimately, we seek to create a representation that captures the relationships between each section of an RCT paper. For example, the way patients are included/excluded from the study population will influence the generalizability of the study's findings in clinical practice. We wish to capture the dependencies between each section in a process model, where the study population, interventions, outcome measures, and flow of events are explicitly represented on a timeline. Fragments of evidence, such as measurements taken at different time points reported in the paper, are captured and associated with the variables that they measure. Statistical information that has been extracted (*e.g.*, study arms, hypothesis test, significance level) can be associated with the outcome variables, reported results, and interpreted statistical and/or clinical significance. A distinction should be drawn in regards to statistical significance and clinical significance [24]. Even if a hypothesis is found to have statistical significance based on a null hypothesis test that yields a p-value lower than a predefined significance criterion (*e.g.*, $p < 0.05$), the finding may not be clinically significant. Furthermore, the intuition of p-values has been widely debated [25]. Clinical significance is influenced by factors such as magnitude of effect, prevalence of the disease, practicality of applying such approach in practice, and the tradeoff between cost and benefit. We believe a logical representation such as the one described helps capture important context related to a statistically significant finding. This information can facilitate a Bayesian interpretation of findings, providing a meaningful way for clinicians to judge how reported findings are applicable to their individual patients.

Our efforts towards automating the identification of statistical information are complementary to existing efforts to model the RCT domain. We are standardizing the values that are being inputted into our logical representation using concepts specified within the Ontology for Biomedical Investigations or Ontology of Clinical Research. Our use of the ASAP framework demonstrates its flexible architecture for adapting to different domains. We were able to wrap various knowledge sources and processing tools (*i.e.*, UIMA) as web services with which ASAP could interface. We have followed a hybrid top-down, bottom-up strategy towards creating our logical representation: we started with a basic model created from existing ontologies and formalisms of statistical knowledge. Then, we manually examined the training set, adding new concepts and attributes to the model based on our observations. We acknowledge the need to expand the breadth of our work into other domains (*e.g.*, all lung cancers) to capture further variations in reported statistical analysis. We believe the approach described in this paper can be generalized to facilitate the information extraction of other sections of the paper such as the interventions and experimental protocol.

Acknowledgements

The authors would like to thank Dr. Michael Ochs of the Johns Hopkins University for allowing us to adapt the ASAP framework for this application. We would also like to acknowledge Dr. James Sayre for his feedback to this work. This work is supported in part by the National Library of Medicine through grants 5R01LM009961 and 5T15LM007356.

References

1. Ioannidis JPA. Why Most Published Research Findings Are False. PLoS Med. 2005;**2**(8):e124.
2. Cokol M, Ozbay F, Rodriguez-Esteban R. Retraction rates are on the rise. EMBO Rep. 2008;**9**(1):2-.
3. Marcovitch H. Is research safe in their hands? BMJ. 2011;**342**:d284.
4. Harris AHS, Reeder R, Hyun JK. Common statistical and research design problems in manuscripts submitted to high-impact psychiatry journals: What editors and reviewers want authors to know. Journal of Psychiatric Research. 2009;**43**(15):1231-4.
5. Sim I, Detmer DE. Beyond trial registration: a global trial bank for clinical trial reporting. PLoS Med. 2005;**2**(11):e365.
6. Schulz K, Altman D, Moher D, Group tC. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. BMC Medicine. 2010;**8**(1):18.
7. Zarin DA, Tse T, Williams RJ, Califf RM, Ide NC. The ClinicalTrials.gov Results Database — Update and Key Issues. N Engl J Med. 2011;**364**(9):852-60.
8. Burns G, Cheng W-C. Tools for knowledge acquisition within the NeuroScholar system and their application to anatomical tract-tracing data. Journal of Biomedical Discovery and Collaboration. 2006;**1**(1):10.
9. Weng C, Wu X, Luo Z, Boland MR, Theodoratos D, Johnson SB. EliXR: an approach to eligibility criteria extraction and representation. J Am Med Inform Assoc. 2011;**18**(Suppl 1):i116-24.
10. Sackett DL, Rosenberg WMC, Gray JAM, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ. 1996;**312**(7023):71-2.
11. Sackett DL. Evidence-based medicine : how to practice and teach EBM. Edinburgh; New York: Churchill Livingstone; 2000.
12. Huang X, Lin J, Demner-Fushman D. Evaluation of PICO as a knowledge representation for clinical questions. AMIA Annu Symp Proc. 2006. p. 359-63.
13. Hansen MJ, Rasmussen NØ, Chung G. A method of extracting the number of trial participants from abstracts describing randomized controlled trials. J Telemed Telecare. 2008;**14**(7):354-8.
14. Chung G. Sentence retrieval for abstracts of randomized controlled trials. BMC Med Inform Decis Mak. 2009;**9**(1):10.
15. Boudin F, Nie J-Y, Bartlett J, Grad R, Pluye P, Dawes M. Combining classifiers for robust PICO element detection. BMC Med Inform Decis Mak. 2010;**10**(1):29.
16. Catherine B. Beyond genes, proteins, and abstracts: Identifying scientific claims from full-text biomedical articles. J Biomed Inform. 2010;**43**(2):173-89.
17. Kossenkov A, Manion FJ, Korotkov E, Moloshok TD, Ochs MF. ASAP: automated sequence annotation pipeline for web-based updating of sequence information with a local dynamic database. Bioinformatics. 2003;**19**(5):675-6.
18. Dowell R, Jokerst R, Day A, Eddy S, Stein L. The Distributed Annotation System. BMC Bioinformatics. 2001;**2**(1):7.
19. Speier W, Ochs MF. Updating annotation with the distributed annotation system and the automated sequence annotation pipeline. Bioinformatics. Forthcoming.

20. Ferrucci D, Lally A. UIMA: an architectural approach to unstructured information processing in the corporate research environment. Nat Lang Eng. 2004;**10**(3-4):327-48.
21. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010 September 1, 2010;**17**(5):507-13.
22. Pallis A, Agelaki S, Agelidou A, et al. A randomized phase III study of the docetaxel/carboplatin combination versus docetaxel single-agent as second line treatment for patients with advanced/metastatic Non-Small Cell Lung Cancer. BMC Cancer. 2010;**10**(1):633.
23. Groen HJM, Sietsma H, Vincent A, et al. Randomized, Placebo-Controlled Phase III Study of Docetaxel Plus Carboplatin With Celecoxib and Cyclooxygenase-2 Expression As a Biomarker for Patients With Advanced Non–Small-Cell Lung Cancer: The NVALT-4 Study. J Clin Oncol. 2011;**29**(32):4320-6.
24. Friedman LM. Clinical Significance versus Statistical Significance. Encyclopedia of Biostatistics: John Wiley & Sons, Ltd; 2005.
25. Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. Ann Intern Med. 1999;**130**(12):995-1004.