Title
Contextualized Semantic Maps for Retrieval and Summarization of Biomedical Literature

Author
Garcia-Gathright, Jean Garcia-Gathright

Publication Date
2016

Peer reviewed|Thesis/dissertation

University of California
Los Angeles

Contextualized Semantic Maps for Retrieval and Summarization of Biomedical Literature

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Biomedical Engineering

by

Jean Imelda Garcia-Gathright

2016

Abstract of the Dissertation

Contextualized Semantic Maps for Retrieval and Summarization of Biomedical Literature

by

Jean Imelda Garcia-Gathright

Doctor of Philosophy in Biomedical Engineering

University of California, Los Angeles, 2016

Professor Denise R. Aberle, Chair

As the volume of biomedical literature increases, it can be challenging for clinicians to stay up-to-date on this massive store of knowledge. Graphical summarization systems condense knowledge into a more tractable form via "concept maps" – networks of nodes (concepts) and edges (relations between concepts). In existing graphical summarization systems, the context of the extracted relations (such as study design and study population) is omitted. However, context is crucial for capturing the full meaning of a relation. With context, the user may pose more detailed queries than those accommodated by traditional, context-free maps.

This dissertation describes Casama, a system for creating "contextualized semantic maps" to represent the current state of scientific knowledge in the domain of non-small cell lung cancer (NSCLC). A formalism for contextualized semantic maps is presented, including targeted relations, study design context, and study population context. An annotated gold standard conforming to this representation is produced, and methods for extracting these contexts are developed. Contextualized semantic maps are evaluated in an information retrieval task and a summarization usability study, showing significant improvement over PubMed and SemRep.

The dissertation of Jean Imelda Garcia-Gathright is approved.

Alex Anh-Tuan Bui

Michael G. Dyer

Edward B. Garon

Ricky Kiyotaka Taira

Denise R. Aberle, Committee Chair

University of California, Los Angeles

2016

To my father

Table of Contents

List of Figures

List of Tables

Acknowledgments

I'd like to thank my committee: Deni, for being my role model and guide; Alex, for your experience and insight; Eddie, for your generosity and expertise; Ricky, for helping me see the big picture; and Professor Dyer, for supporting me exactly when I needed you.

Many thanks to my annotators, relevance judges, and evaluators: Andrea Oh, Phillip Abarca, Mary Han, William Sago, Marshall Spiegel, Brian Wolf, Nicholas Matiasz, James Carroll, Alice Li, Jennifer Strunck, Carlos Adame, Karthik Sarma, Lauren Sauer, and Nova Smedley. This work would not have been possible without your efforts!

Thank you to the Medical Imaging Informatics lab: to our teachers and mentors (Frank, Will, Corey, Jim Sayre, Suzie El-Saden, Craig Morioka, and the late Hooshang Kangarloo); to those who graduated before me (Emily, Eugenio, Neda, Brian, Bill); to the freshly-minted graduates who took this journey with me (Mary, Kyle, Anna, Maurine); to the next generation of PhD candidates (Johnny, Shiwen, Nick, Edgar); to all the students who are finding their way in this wild adventure (Pavitra, Simon, Nova, Panayiotis, Karthik, Jiayun, Tianran, Justin); and to the hard-working current and former research and administrative staff (Isabel, Audrey, Lew, Shawn, Carlos, Brian, Denise, Weixia, Patrick, Hamid, Noah). Stay curious, my friends.

Thank you to the UCLA Graduate Writing Center, the GWC writing consultants, and my boot camp peers for your many helpful reviews.

To my dearest friends from the dumpling gang (Nerissa, Emily, Gin, Melissa), the dinner party club (Melissa, Jeremy, Micah, Tiffany), the Friendsgiving crew (Jeremy, Luis, Cathy, Cing, Andy, Jill, Tim, Mariann, Mark, Myra, Scott, Shirley, Tom, Alice, Wilson, Raina, Ha, Tommy, and the children), and Noelle, who defies categorization – I think a celebratory meal is in order.

Thank you to everyone I've made music with during my graduate studies, especially the beautiful people from the parish choir of St. Thomas the Apostle Hollywood, Mandolin Ensemble LA, Roar of the Current, Voices of Reason, Storytown, UCLA Early Music En-

semble, UCLA Chorale, UCLA Bluegrass and Old-Time String Ensemble, LA Intermediate Irish Session, Andy and Barbara Cameron's Old-Time and Latin America jam, Pierre Baldi and the Students, Meredith Monk and Vocal Ensemble (my participation made possible by the amazing Rebecca Lord and Meryl Friedman), Together in Music, and the Jean Garcia Trio – music self-played is happiness self-made.

To my JPL/Keck friends: Mark Colavita, Andrew Booth, Michelle Thaller, Mark Swain, Melanie Swain, Armin Kleinboehl, Olga Pikelnaya, Greg Dubos, Thomas Kurosu, and Amy Vogel née Tumminello – thank you for showing me how to win at work and life.

To the surprising and hilarious individuals from UCLA Linguistics: Meaghan Fowlie, Michael Tseng, Kathleen Chase O'Flynn – thank you for your unique charms and perspectives.

To the pub quiz killers of RAND Corporation: Andrew Naber, Lisa Rand, Will Frankenstein, Jonathan Welburn, Amanda Stype, Tanguy Hubert, Charlotte Greenan, Gabrielle Filip-Crawford, George Wilcoxon, Ajax Peris, Josh Embree, Kevin Heins, Asya Spears, Darlene Kiloglu, Amber Jaycocks, and Fritz – go easy on us when you're taking over the world.

To George Kirkman, Bing Jiang, and the staff and students of Rolling Robots – thank you for always making me feel valued, for giving me the opportunity to teach, and for Botball.

Deepest gratitude to my sister Melissa, who reminded me that I always finish what I start, and to my mother and father, who taught me that no one can take away my education. I'd also like to thank my extended family (the Garcias, the Tans, and the Gathrights) for being a constant source of joy and love.

To my dear husband John, thank you for supporting me through seven hot Los Angeles summers. Take to the sea!

This work was supported by NLM T15-LM007356, NIH/NLM R01-LM009961, and the UCLA Department of Radiological Sciences.

Chapters Four and Five were adapted from material published in:

Garcia-Gathright, Jean I., Andrea Oh, Phillip A. Abarca, Mary Han, William Sago, Marshall L. Spiegel, Brian Wolf, Edward B. Garon, Alex A.T. Bui, and Denise R. Aberle. "Representing and extracting lung cancer study metadata: Study objective and study design." *Computers in Biology and Medicine* 58 (2015): 63-72. doi: 10.1016/j.compbiomed.2015.01.004.

Garcia-Gathright, Jean I., Nicholas J. Matiasz, Edward B. Garon, Denise R. Aberle, Ricky K. Taira, and Alex A.T. Bui. "Toward patient-tailored summarization of lung cancer literature." In: *Proceedings of the IEEE International Conference on Biomedical and Health Informatics*. February 2016, Las Vegas, Nevada. doi: 10.1109/BHI.2016.7455931

Andrea Oh, Phillip A. Abarca, Mary Han, William Sago, Marshall L. Spiegel, Brian Wolf, and Nicholas J. Matiasz annotated the Casama document set and helped revise the annotation guidelines. Edward B. Garon and Denise R. Aberle contributed their lung cancer expertise to the design of the Casama representation. Alex A.T. Bui and Ricky K. Taira assisted with experiment design and interpretation of results.

Vita

2002	B.S. (Computer Science), California State Polytechnic University, Pomona.
2002-2008	Software Engineer, Jet Propulsion Laboratory.
2008-2012	National Library of Medicine Pre-Doctoral Trainee.
2012-2014	Graduate Student Researcher, Department of Radiological Sciences, UCLA.
2013	RAND Corporation Summer Associate.

Publications and Presentations

Garcia-Gathright, Jean I., Nicholas J. Matiasz, Edward B. Garon, Denise R. Aberle, Ricky K. Taira, and Alex A.T. Bui. "Toward patient-tailored summarization of lung cancer literature." In: *Proceedings of the IEEE International Conference on Biomedical and Health Informatics.* February 24-27, 2016. Las Vegas, Nevada. doi: 10.1109/BHI.2016.7455931

Garcia-Gathright, Jean I., Andrea Oh, Phillip A. Abarca, Mary Han, William Sago, Marshall L. Spiegel, Brian Wolf, Edward B. Garon, Alex A.T. Bui, and Denise R. Aberle. "Representing and extracting lung cancer study metadata: Study objective and study design." *Computers in Biology and Medicine* 58 (2015): 63-72. doi: 10.1016/j.compbiomed.2015.01.004.

Garcia-Gathright, Jean I., Frank Meng, and William Hsu. "UCLA at TREC 2014 Clinical Decision Support Track: Exploring Language Models, Query Expansion, and Boosting."

In NIST Special Publication: SP 500-308. The Twenty-Third Text REtrieval Conference Proceedings (TREC 2014).

Garcia-Gathright Jean I., Corey W. Arnold, Alex A.T. Bui. "Automatic extraction of patient characteristics from clinical reports." Presented at: Radiological Society of North America Annual Meeting, Chicago, Illinois; November 2013.

Garcia-Gathright, Jean I., Ricky K. Taira, Denise R. Aberle. "Learning Context-Sensitive Association Models." Presented at: National Library of Medicine Training Conference, Madison, Wisconsin; June 2012.

Singleton Kyle W., Jean I. Garcia-Gathright, Brian Burns, Krupa Rocks, J. Eugenio Iglesias, Alex A.T. Bui, Denise R. Aberle. "Semi-automated medical text and image selection for multimedia presentation at tumor board reviews." Presented at: Radiological Society of North America Annual Meeting, Chicago, IL; November 2009.

CHAPTER 1

Introduction

In the domain of clinical practice, the primary medium of new knowledge is "natural language": free, unstructured text in the form of research articles, clinical trial reports, and clinical guidelines. With the volume and accessibility of published biomedical literature increasing at an unprecedented rate, it is challenging for a clinician to stay up-to-date on this massive amount of knowledge. Aggregating and summarizing the current state of knowledge in a disease domain can help inform a clinician's thinking on disease processes and the effectiveness of treatment strategies. Ultimately, the goal of summarization is to improve patient care by providing accurate, structured, and tractable knowledge, thereby illuminating pathways between disease, therapies, and patient outcomes.

Summarization systems such as UpToDate provide manually-curated overviews of clinical topics. However, given the expense associated with expert curation, utilizing natural language processing techniques for automatic summarization is an attractive alternative. One approach to automatic summarization is relation extraction, the process of automatically mining natural language text for entities of interest (such as treatments and outcomes) and the semantic relationships that exist between them (such as "treatment X improves outcome Y"). Relation extraction has been relatively well-studied. Current relation extraction systems omit the context of the extracted relations. In other words, if a relation such as "treatment X improves outcome Y" is detected, this association is considered "true" regardless of the context in which the relation was found. However, context is crucial for capturing the full meaning of a relation.

In this dissertation, the notion of "context" is characterized at two levels, informed by two current perspectives in medicine: evidence-based medicine and precision medicine. The first

level of context is the study level, which describes experimental conditions such as study design and outcome measures. In doing so, the resulting summaries inform the clinician regarding strength of evidence, thus facilitating clinical decision making based on evidence from compelling, well-designed studies. The second level of context is the patient/population level, which captures properties of the study population. In accordance with the efforts of precision medicine, leveraging this type of context assists the clinician in retrieving studies whose study populations are similar to an individual patient, customizing clinical decisions to an individual's personal features.

Figure 1.1 illustrates an example of a clinical use case and how contextualization can address problems inherent to uncontextualized summarization. Consider a clinician seeking information on non-small cell lung cancer (NSCLC) harboring EGFR mutation (EGFR+). The clinician has a limited amount of time to review all the literature on NSCLC. A graphical summary based on relations mined from published literature can save time. However, even a small collection of documents will produce a large, potentially conflicting set of extracted relations. Furthermore, these relations each exist in separate contexts. Some relations were found in clinical trials; others were discovered in less compelling retrospective studies. Also, some study populations were unlike the clinician's current patient, while other study populations matched exactly.

The clinician may wish to answer the following questions: What treatments are available for this disease? What is a likely prognosis for this disease and treatment strategy? This dissertation describes a framework that aims to assist a clinician in answering these questions, in addition to more complex queries concerning the strength of the claims found in the literature and the applicability of those claims to an individual patient. For instance, is this treatment safe and appropriate to prescribe to the patient, given his or her overall health status? Is this treatment likely to be effective, based on compelling, consistent evidence in the literature?

These efforts are demonstrated in the domain of clinically-oriented knowledge of driver mutations and targeted therapies in NSCLC. The Lung Cancer Mutation Consortium, the National Cancer Institute's effort to characterize driver mutations in lung cancer, found that

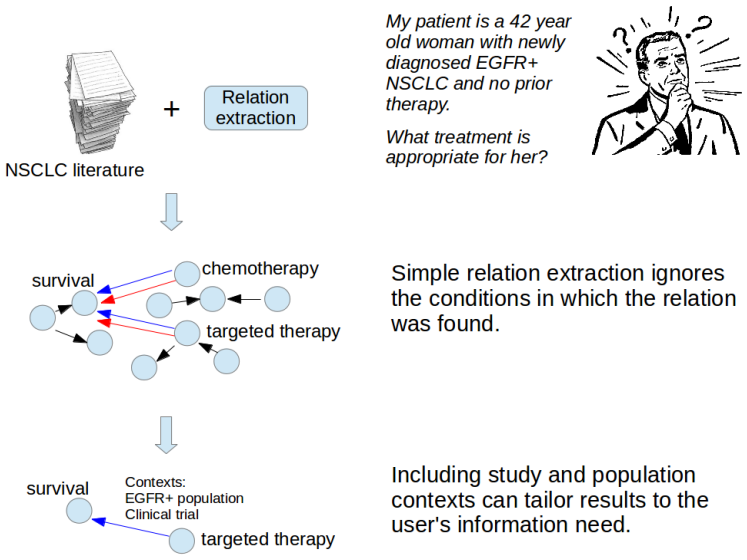


Figure 1.1: Without context, every relation is considered "true" regardless of differing study and population contexts. Contextualized relations can help tailor summarized knowledge.
Circles represent concepts of interest; arrows are the relations between these concepts. Red and blue arrows indicate that conflicting relations were found. Context, represented as the attributes of a relation, can reduce the size of the network and resolve conflicting information.

driver mutations were present in 64% of lung adenocarcinomas, and that patients who were treated with targeted therapy lived longer than those who did not receive such treatment [KJB14]. Currently, targeted treatments approved by the Federal Drug Administration are available for cancers with epidermal growth factor receptor (EGFR) mutations and anaplastic lymphoma kinase (ALK) gene rearrangement. Improvement of clinical outcomes with targeted therapies has been demonstrated in clinical trials, and new advances continue to be made [CFH14].

The major contribution of this work is the Contextualized Semantic Map, a formalism that ties relations to their contexts. Figure 1.2 provides a visual overview of the components of a contextualized semantic map, and the technical contribution of each component. First, to instantiate a contextualized semantic map in a particular domain, it is necessary to define the concepts and relations relevant to that domain. Then, these relations and concepts can be identified from a body of literature. Manual annotation of a subset of documents from a large corpus of biomedical literature such as PubMed provides a gold standard against which

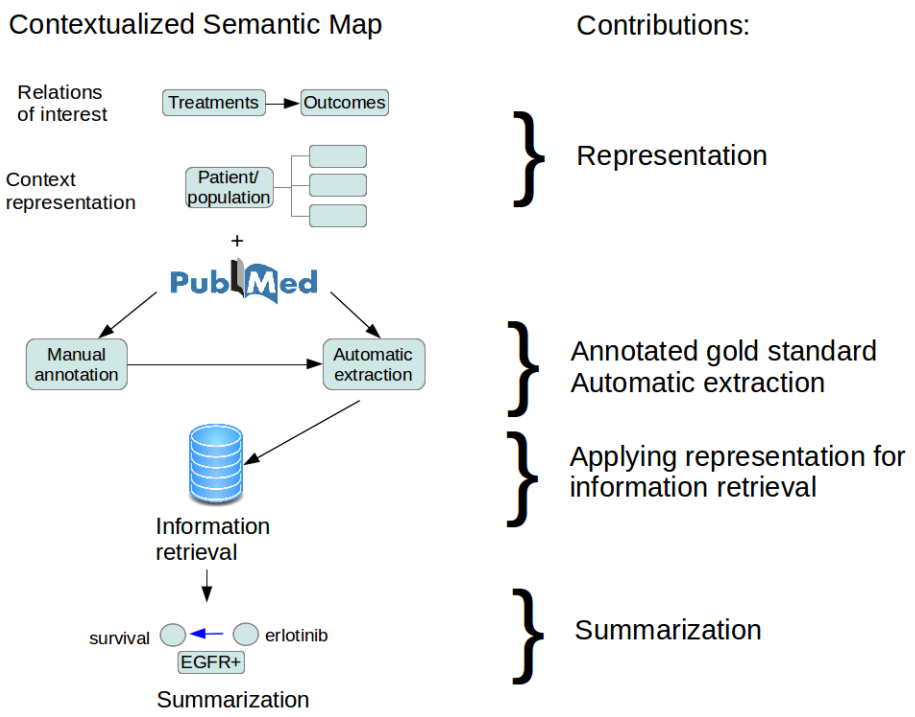


Figure 1.2: An overview of the Contextualized Semantic Map formalism and the contributions of each element.

current and future automatic extraction tasks can be evaluated. These extraction methods automatically tag the document set with concepts and relations in the representation. The document set can then be indexed and searched to retrieve documents most relevant to a user-defined query, i.e., the attributes of individual patient. Finally, a contextualized semantic map, in which the contexts of each relation are represented as edge attributes, provides a summary of the desired knowledge.

The collection of representations, gold standards, and methods for extraction, retrieval, and summarization shall herein be referred to as "Casama," a Tagalog word meaning "together."

Formally, this dissertation addresses the following specific aims:

1. Developing a new formalism, the "contextualized semantic map," that abstracts the findings, strength of evidence, and population attributes in a scientific study.

2. Demonstrating how this formalism can be used to facilitate evidence-based and patient-

tailored information retrieval and automatic summarization.

These aims are achieved via the following technical contributions:

1. A formal definition of the contextualized semantic map, and an instance of this representation for NSCLC.

2. An annotated gold standard conforming to this representation.

3. Methods for automatic extraction of concepts, relations, and their contexts.

4. Evaluation of information retrieval and automatic summarization methods that leverage the contextualized semantic map representation.

The dissertation is organized as follows. Chapter 2 provides an overview of related work in automatic summarization, relation extraction, study and patient/population representations, and information retrieval. Chapter 3 introduces the representation for a contextualized semantic map, including targeted relations, study context, and patient/population context. Chapter 4 describes the annotation of a corpus of literature on driver mutations in NSCLC. Chapters 5 and 6 describe methods for automatic extraction of relations and context. In Chapter 7, these extractors are used to improve information retrieval on a previously unseen data set. Chapter 8 examines the use of contextualized semantic map for summarization, and Chapter 9 reviews the findings and points to future work.

CHAPTER 2

Background

This literature review is organized as follows. First, relation extraction as a summarization paradigm is examined, investigating the various rule-based, statistical, and pattern-based approaches developed in the last twenty years. Then, various representations are presented for two types of context: strength of evidence and patient/population context, followed by a presentation of methods for automatic extraction of key entities from clinical trial reports and other scientific literature. Discussed next are annotation methods, metrics of evaluation, and existing biomedical corpora. Finally, two applications of contextualized relations are reviewed: information retrieval and automatic summarization.

2.1 Relation extraction

The representation of knowledge as concepts and relations was first explored in the 1970s by Novak, who applied this representation for education [Nov77], and Sowa, who developed a computable formalism that supports querying and inference [Sow76]. The basic element of knowledge proposed was the proposition — two or more concepts linked by a relationship to form a semantic unit. The collection of these propositions, referred to as "concept maps" or "conceptual graphs" have been shown to be an effective way to represent, visualize, and communicate knowledge [NG84]. As natural language processing techniques developed and matured, the automatic extraction of relational knowledge became a logical continuation of their work.

Relation extraction has been an active area of research for over twenty years. Like many artificial intelligence tasks, early approaches relied on rule-matching. Rule-based systems

tend to be brittle, because no finite set of rules can capture the infinite possibilities of language. As statistical natural language processing grew to fruition, machine learning methods were applied to the relation extraction task. While statistical methods are more robust, they require annotated training data, which can be time-intensive to produce. For these reasons, both rule-based and supervised statistical approaches typically focus on a small, pre-specified set of relations of interest. As such, much of the current work in relation extraction is driven by the existence of annotated training sets via shared tasks. Extraction methods that do not rely on a set of targeted relations have grown in popularity with the development of large, heterogeneous sources of natural language text such as the World Wide Web.

2.1.1 SemRep: a rule-based system

This dissertation was influenced most significantly by SemRep [RFL05], a rule-based relation extraction system developed by Rindflesch et al. SemRep automatically extracts the relationships between concepts, referred to as "semantic predications," from natural language text, usually in the form of journal abstracts. SemRep combines a syntactic parse and a set of grammar rules with domain knowledge from the Unified Medical Language System (UMLS): the text is mapped to concepts in the Metathesaurus, and the relations are from the Semantic Network [Bod04].

The development of SemRep and related work by the National Library of Medicine over the past decade has been thoroughly described in published literature. EDGAR [RTW00], a predecessor to SemRep, focuses on the extraction of drugs and genes for a document clustering task. EDGAR uses several strategies for cell and gene identification: a short list of characteristic signal words (such as "cell," "line," "gene," and "mutated"), rules for identifying cell and gene references, and MetaMap. SemGen [RLH03], a system based on SemRep, focuses on identifying gene-disease relations (**causes**, **predisposes**, **associated_with**) and gene-gene interactions (**inhibits**, **stimulates**, **interacts_with**). SemGen adds to SemRep with a labeled categorizer (for recognizing content relating to molecular genetics) and by in-

corporating domain knowledge for identifying genetic phenomena. In [AFD07], Ahlers et al. developed Enhanced SemRep, a system for identifying semantic predications in the domain of pharmacogenomics. The semantic predications targeted by Enhanced SemRep encompass genetic etiology (**predisposes**, **causes**), substance interactions (**inhibits**, **stimulates**), pharmacological effects (**disrupts**, **augments**), clinical actions (**administered_to**, **treats**), organism characteristics (**part_of**, **process_of**), and co-existence (**coexists_with**).

2.1.2 Machine learning approaches

SemEval (Semantic Evaluation), a general domain shared task, included a relation extraction task in 2007 and 2010 [GNN07, HKK09]. Targeted relations that are potentially applicable to the biomedical domain include: **cause-effect**, **product-producer**, **origin-entity**, and **part-whole**. The highest-performing system in 2007 by Beamer et al. [BBC07] trained a support vector machine (SVM) classifier on a set of linguistic features. The feature set consisted of core features (such as position of arguments) and context features (sentence-level grammatical and semantic information). Rink and Haribagiu developed the winning system in 2010 [RH10]; they also used an SVM classifier and a number of external knowledge sources including WordNet, PropBank, FrameNet, TextRunner, and Google N-grams.

The BioNLP Event Extraction shared task focuses on the extraction of bio-molecular events such as gene expression, transcription, and regulation [NBK13]. Best performance was achieved by Bjorne et al. in 2009 [BHG09], 2011 [BS11], and 2013 [BS13]. Their system, TEES (Turku Event Extraction System), divides the task into multiple stages: trigger detection (identifying predicates that signify an event), edge detection (discovering event arguments), and unmerging (constructing valid trigger and argument relations). An SVM is used at each stage to classify triggers, arguments, and relations as positive or negative.

SemEval included a drug-drug interaction task in 2011 and 2013. The overviews of all the participants' results for both years [SMS11, SMH13] highlighted the advantage of non-linear kernel methods over linear approaches such as SVMs. In 2011, the winning system developed by Thomas et al. used ensemble learning, combining multiple kernel-based classifiers in a

voting approach [TNS11]. In 2013, Chowdhury and Lavelli devised a hybrid kernel combining shallow linguistic and tree features [CL13].

Outside of shared tasks, similar trends can be seen in the use of machine learning methods such as SVMs ([PKT06, YRH11]), kernel methods ([MB05, GLR06]), and others (neural networks: [RH04, BWC09], conditional random fields: [BDS08], maximum entropy: [FLD11]).

Machine learning approaches provide more robust performance and better generalizability than rule-based systems. However, the supervised learning techniques described above require an annotated training corpus. While it is possible to leverage existing annotated corpora, for many applications the cost of manually annotating a large training corpus is significant, especially as the number of targeted relations increases. Semi-supervised and unsupervised approaches, many of which are pattern-based, attempt to mitigate this problem. These pattern-based approaches, which require little-to-no annotated data, can then be scaled up to tackle vast corpora such as the World Wide Web.

2.1.3 Pattern-based approaches and Web scale relation extraction

Hearst's seminal paper on extracting **is-a** relations [Hea92] describes an approach for automatically discovering lexico-syntactic patterns (LSPs or Hearst patterns) from Grolier's Academic Encyclopedia. Given a list of terms in which the relation holds (e.g., England **is-a** country), the input corpus (Grolier's Academic Encyclopedia) is searched for co-occurrences of those terms. Frequently occurring syntactic patterns surrounding these co-occurrences are hypothesized to indicate the relation of interest. New patterns are used to gather more instances of the relation, and the process is repeated.

In [GM02], Girju and Moldovan apply Hearst's LSP discovery algorithm to the task of causal relation extraction, extracting relations from news articles, with WordNet used as the source of known relations. In [PP08], Pantel and Pennacchiotti describe Espresso, a relation extraction system that modifies Hearst's approach by discovering broad coverage noisy patterns, then ranking them to create a set of reliable patterns. Performance

of Espresso was evaluated with respect to five relations (**is-a**, **part-of**, **succeeds**, **reacts with**, **produces**) on two corpora (TREC-9 and CHEM). Rong and Xu extract drug-disease treatment pairs by learning patterns bootstrapped from known drug-disease relations found in ClinicalTrials.gov [XW13].

In [IB11], [SKW08], and [MAG14], Wikipedia is used as the external knowledge source for relation extraction. Ittoo and Bouma [IB11] apply Hearst patterns for extracting explicit (**causes**) and implicit (**affects**, **inflicts**) causal relations. Sentences from Wikipedia are transformed into lexico-syntactic patterns, then a small number of seed cause-effect pairs are used to identify patterns that encode causality. The most reliable patterns are used to extract relations from a domain-specific, sparse corpus. Suchanek et al. leverage the structured knowledge in Wikipedia and WordNet to create YAGO, a high precision, broad coverage ontology [SKW08]. YAGO uses Wikipedia page titles, infoboxes, category hierarchies, and WordNet synsets to collect entities, classes (groups of similar entities), and facts (relations between two entities). Mousavi et al. automatically extract structured information from the text of Wikipedia articles to populate a knowledge base known as TextGraph [MAG14].

Information extraction at the Web scale presents a new set of challenges. Because the corpus is vast and heterogeneous, extraction algorithms must be efficient, generalizable, and minimally supervised. Thus, several systems use a Hearst-like approach to populate large knowledge bases with relations extracted from the World Wide Web. DeepDive uses "distant supervision" to automatically create a noisy training set, leveraging the existing ontology FreeBase to provide positive examples of relations. DeepDive then uses Markov logic to statistically infer which linguistic patterns are the most reliable [NZR12]. In [BCS07], Banko et al. develop TextRunner, a system that uses no human input to extract a large number of unspecified relations from the input corpus. First, the system uses a few manually-defined heuristics to automatically tag positive and negative examples of relations in a small corpus. Syntactic patterns extracted from these examples are used to train a naive Bayes classifier, which labels patterns as trustworthy or untrustworthy. These labeled patterns are then used to extract relations during a single-pass over the input corpus. At the time of this writing, this new paradigm of Open Information Extraction is in its fourth generation, adding a semantic

role labeling step that enables support for n-ary relations and conditionality [CSE10].

There has been some interest in adapting Open Information Extraction to the biomedical domain. BioNELL, a biomedical variant of the Never-Ending Language Learner, bootstraps its extraction patterns from several biomedical ontologies and a small number of seed examples [MC12]. Nebot and Berlanga utilize semantic annotation of entities, lexico-syntactic patterns, and clustering to identify biomedical relations [NB14]. PASMED uses syntactic patterns observed in the GENIA bio-molecular event corpus to extract relations from MEDLINE with an emphasis on recall over precision [NMT15]. Coulet et al. use rules on the dependency graph of a sentence to extract relationships between key entities in pharmacogenomics [CSG10].

In summary, both rule-based and statistical methods have been applied to the relation extraction task. SemRep, a popular biomedical relation extraction tool, is rule-based. Supervised machine learning methods are often used for extraction of limited sets of relations; however, an annotated corpus is required, and manually annotating a corpus can be very labor-intensive. Unsupervised methods, such as those used for web-scale relation extraction, mitigate the high cost of manual annotation and are not limited to a pre-specified set of relations.

The Casama representation for relations (described in detail in Chapter 3) includes a large set of relations, some of which are not covered by any existing ontology or relation extraction system. Thus, a relation extraction system was implemented that combines rule-based concept recognition, document classification using machine learning, and Web-scale extraction based on OpenIE 4.0. This system is presented in Chapters 5 and 6.

Crucially, few of the systems described above attempt to ascertain the context of the extracted relations. The following section explores the use of context in knowledge-based systems, both in general and biomedical applications. Then, existing representations for Casama's contextual domains (strength of evidence and patient/population) are investigated.

2.2 Context in artificial intelligence and biomedicine

Contextualization in artificial intelligence (AI) has been explored since the 1990s, notably by McCarthy who first motivated the need for a formalism of context. Arguing that existing AI systems lacked generality, McCarthy aimed to develop a mathematical definition of context and introduced the *ist*(c, p) relation (signifying that proposition p was true under context c) [McC93]. Since then, many in the AI community have recognized the importance of context in knowledge-based systems. Cyc [Len95], a large common sense knowledge base, explicitly includes a context mechanism — each assertion is placed in the appropriate location on a lattice of hundreds of contexts. Walther et al. develop a context ontology in [WEM92]; Turney addresses the issue of context for machine learning in [Tur96]; Giunchiglia [Giu93], Serafini and Bouqet [SB04], and Brezillon [Bre03] develop formalisms for representing and reasoning with context. While context in general has been explored in the domain of artificial intelligence, there has been relatively little development of context-sensitive systems to enhance biomedical relation extraction.

A few biomedically-oriented systems augment their extracted relations with some notion of context. Lussier et al. describe PhenoGO [SML09], a natural language processing system based on BioMedLEE, which identifies concepts and semantic types from multiple ontologies to assign phenotypic context such as anatomical structure, body substance, and body system to Gene Ontology annotations. In [GSB12a], Gerner et al. develop BioContext, a text mining system that contextualizes biomolecular events in terms of species involved, anatomical location, and speculation or negation. BIOSMILE augments relations with the surrounding words signifying the location, manner, and timing of an event [TCS07].

This dissertation aims to build upon current work in relation extraction by developing a framework in which the context of relations is represented and extracted, thus providing a more comprehensive summary that includes relevant knowledge such as experimental context and population attributes. The inclusion of additional knowledge in its summaries, and the tying of contextual knowledge to relations, can then be used to improve information retrieval (Chapter 7) and facilitate discovery of relevant facts by users (Chapter 8). Furthermore, by

using context to constrain truth values (as proposed by McCarthy), spurious relation chains (e.g., A **correlated with** B under one context, B **correlated with** C under a different context, which does not imply A **correlated with** C) can be expressed correctly by the system.

Casama targets two types of contextual knowledge: study context and patient/population context. To discover current research trends related to these types of context, a literature review was conducted on systems that represent and apply knowledge from clinical trials. Articles from the last five years related to clinical trials were retrieved from two major medical informatics journals: *Journal of Biomedical Informatics* and *Journal of the American Medical Informatics Association*. A few research trends emerged: 1) the development of formal representations for clinical trial protocols; 2) the evaluation of bias in clinical trials (and the larger goal of representing and extracting strength of evidence); 3) a variety of methods for representing and extracting eligibility criteria for matching patients to clinical trials. A comprehensive review, presented in the next section, provides an overview of efforts to represent these types of knowledge.

2.3 Study context

2.3.1 Standardizing clinical trials and other biomedical literature

Several research endeavors have aimed to standardize the reporting of randomized clinical trials (RCTs). An early effort to standardize the reporting of RCTs led to the publication of the CONSORT statement [SAM10], which consists of a flow diagram illustrating the progress through the phases of a randomized trial, and a checklist describing the details of the clinical trial in terms of background, methods, results, discussion, and more. Studies showed that use of CONSORT improves the quality of clinical trial reports. For example, in [MSA01], Moher et al. showed that omission of allocation concealment dropped from 61% to 39% after the enforcement of the CONSORT standard. Related work by Tong et al. describes a process model representation of clinical trials for structuring experimental

context such as therapy descriptions [TT12] and numerical data [THT13].

The Cochrane Collaboration [HG08] focuses on providing a framework for communicating systematic reviews of RCTs. Each review in the Cochrane Database of Systematic Reviews includes contact information about the reviewers, sources of support for preparing the review, a structured report of the review, citations of studies included, characteristics of the trials, and statistical analyses.

In [SN12], Sim and Niland go beyond standardization and advocate for a computable protocol model, or "e-protocol." E-protocols allow computational approaches to data organization, information management, and knowledge discovery. The authors argue that utilization of an e-protocol will improve study design, clinical study efficiencies, and application to care and research. The e-protocol developed includes background rationale, hypotheses, eligibility criteria, measurements and variables, and statistical analysis plans.

Sim et al. subsequently initiated the Human Studies Database (HSDB) Project [SCT10], which aims to define and implement an infrastructure for the sharing of human study designs. The authors define a study typology for classifying studies into various study types (human or non-human, qualitative or quantitative, interventional or observational). The Ontology of Clinical Research (OCRe) [TCR09] is used by HSDB as the semantic model for the federated sharing of studies. Important entities in the ontology include Study (scientific hypothesis, study plan, investigators, subjects, data sets, start date and end date, study sites, study purpose, study design features, study status), Study Protocol (characteristics of the subjects, activities to be performed, data to be collected, outcomes to be assessed, analysis methods), and Events (observations, interventions).

While these efforts aim to represent the details of a clinical trial for purposes of systemized reporting, others have examined the representation of scientific studies in terms of clinical questions. In [RWN95], Richardson et al. develop a framework for formulating patient-specific questions, in order to facilitate searching for precise answers in the clinical setting. The framework consists of four parts: Problem/Population, Intervention, Comparison, and Outcome (PICO). Huang et al. evaluate the suitability of the PICO framework for

representing clinical questions [HLD06]. They analyzed 59 primary-care clinical questions and found prevalent structural patterns for four types of clinical questions: therapy, diagnosis, prognosis, and etiology. Huang et al. conclude that PICO is best suited for therapy questions.

In [DPS07], Dawes et al. explore the feasibility of a modified PICO framework for indexing and retrieval of medical journal abstracts. "Intervention" is replaced with "Exposure" in order to enable the inclusion of case control and cohort studies, in addition to randomized clinical trials. "Duration" and "Results" were added to the framework. The final representation is: Patient/Population/Problem, Exposure, Comparison, Outcome, Duration, and Results (PECODR). Twenty synopses from the journal *Evidence-Based Medicine* and their corresponding PubMed abstracts were manually examined for PECODR elements. The six PECODR elements were found in nearly all abstracts.

2.3.2 Strength of evidence

Clinical trials and other biomedical studies sometimes fall short of the optimal level of evidence in terms of design, implementation, or the nature of the population. In order for a clinician to determine whether the results provide sufficient evidence to influence his or her clinical decision, the clinician must consider the strength of evidence.

The National Cancer Institute has published guidelines on how to assess quality of evidence in cancer treatment studies [Ins15]. Types of study designs, endpoints, and populations are organized hierarchically, in descending order of strength. Study designs include blinded and non-blinded randomized controlled clinical trials, nonrandomized controlled clinical trials, and case series. Study endpoints include total mortality, cause-specific mortality, carefully assessed quality of life, and indirect surrogates (event-free survival, disease-free survival, progression-free survival, tumor response rate).

The Oxford Centre for Evidence-Based Medicine provides a more detailed hierarchy for grading study designs [Evi11]. Experimental studies such as clinical trials provide the highest level of evidence, followed by several observational study types. In descending order

of strength of evidence, the study types are: prospective cohort studies, retrospective cohort studies, cross-sectional studies, case control studies, and case series.

In [JAE01], Juni et al. summarize another set of elements used to assess the quality of controlled clinical trials for the purposes of meta-analysis. Selection bias, performance bias, detection bias, and attrition bias determine the level of systematic error in a clinical trial. Patient characteristics, treatment regimens, care settings, and outcome measures are used to assess the generalizability of a clinical trial. Other potential sources of bias in clinical trials include publication bias [VH13] and under-representation in study populations [HRH16].

Previous work in classifying studies by strength of evidence relies on independently established standards of evidence, often reduced to two or three classes of evidence level. Aphinyanphongs et al. designated their input articles as ACP+ or ACP- depending on whether they were listed in the *American College of Physicians Journal Club* [ATS05]. Kilicoglu et al. used the Clinical Hedge Database, the manually-annotated input set used to produce PubMed's Clinical Queries filters; articles were tagged with regard to their "scientific rigor" (a binary yes/no assessment) [KDR09, FBS10, CRY12]. Mollá and Gyawali used strength of recommendation scores (A, B, or C) as a metric of evidence [MS12, GSB12b]. In the domain of neuroscience research, Landreth proposes a graphical summary of published literature in which study reproducibility and convergence are used to weight evidence [LS13].

2.3.3 Casama and study context

Current work in the representation of scientific studies ranges in granularity from the highly detailed (e.g., CONSORT, Cochrane, OCRe) to the highly distilled (e.g., works that automatically grade studies by strength of evidence). The representation developed for Casama is the backbone of the semantic map produced; thus, an efficient and well-structured representation of clinical studies is crucial. In this dissertation, a subset from existing resources was selected in order to create a representation that is well-suited to application in information retrieval and summarization in a clinical setting. Thus, the representation described in Chapter 3, rather than focusing on every detail relating to the conducting and reporting

of a clinical study, instead targets one important aspect of clinical studies: strength of evidence. Furthermore, these concepts are extracted in a more granular fashion than existing automatic classification systems to enable meaningful interpretation by human users.

2.4 Patient/population context

Much of the current research in representing patient populations involves the representation and extraction of eligibility criteria. The main purpose of examining eligibility criteria is to identify candidates for enrollment in clinical trials or retrospective studies. Secondarily, eligibility criteria help inform a clinician as to whether an intervention described in a clinical trial report is appropriate for a specific patient. ClinicalTrials.gov reports eligibility criteria in free text form – a list of patient characteristics that include or exclude a patient from a study. These characteristics can be straightforward ("confirmed advanced non-small cell lung cancer") or may include temporal information ("disease free from a previously treated malignancy for more than three years"), numeric comparisons ("white blood cell count $>$ 3,000 mm^3"), and conjunctions of these ("alkaline phosphatase, aspartate aminotransferase and alanine transaminase $<$ 2.5 x upper limits of normal").

In [TPC11], Tu et al. describe ERGO, an annotation system which formalizes eligibility criteria into a computable representation. Criteria are classified into one of three categories (simple statements, comparison statements, and complex statements) and are then rewritten according to a set of rules to conform to the ERGO representation. Automatic extraction relies on a set of heuristics to extract noun phrases, modifiers, comparisons, and quantities.

Luo et al. manually defined 27 eligibility criteria categories, organized into six topic groups: demographics, health status, treatment or health care, diagnostic or lab tests, ethical consideration, and lifestyle choice in [LJL11]. The authors also define an ontology for representing temporal constraints in eligibility criteria. Conditional random fields classify eligibility criteria into one of the classes in the ontology.

In contrast to the top-down, representational approaches described above, other systems use a bottom-up approach, mining eligibility criteria to discover common elements and enable

automatic clustering and filtering. Weng et al. developed EliXR, which uses categories from the UMLS Semantic Network to automatically induce commonly-found semantic patterns and semantic role labels found in eligibility criteria [WWL11]. They found that the 12 most common semantic role labels are: medical condition, therapy or surgery, medication, patient group, modifiers, temporal constraint, body location, manifestation, diagnosis or assessment, consequence, medical specialist, and device. He et al. describe VITTA (Visual Analysis Tool of Clinical Study Target Populations), a tool for identifying and visualizing common recruitment features [HCS15]. Luo et al. annotate clinical trials with UMLS terms to discover common data elements in eligibility criteria [LMW13]. Hao et al. [HRB14] locate similar clinical trials by clustering them based on their eligibility criteria. Miotto [MJW13] and Riccardo [MW13] use eligibility criteria to filter and index clinical trials.

While a detailed representation is necessary for capturing all the criteria involved in matching patients to clinical trials, this level of granularity is overly strict for the purposes of summarization and information retrieval and may lead to the exclusion of potentially useful information [Geo96]. Matching patients to published literature for the purposes of information retrieval has thus far relied on a general representation. For example, Demner-Fushman and Lin use the PICO representation to retrieve literature regarding a clinical question; P stands for "Population" or "Problem" [DL07]. Their extractors were designed for broad coverage, tagging mentions of semantic types Group and Disorder from the UMLS vocabulary.

2.4.1 Casama and patient/population context

Again, defining the intended level of granularity in Casama's representation is key. Current work in eligibility criteria is highly granular, as it is designed to capture the strict requirements needed to include or exclude patients from participation in clinical trials. Demner-Fushman and Lin use a much less granular representation for purposes of information retrieval; Casama aims for a more information-rich approach that includes many relevant details of patient population to enable patient-tailored information retrieval. Greater ex-

pressive capabilities are possible with a domain-specific representation, such as the representation developed for Casama in the domain of lung cancer. These capabilities are explored in Chapter 3.

2.5 Information extraction from clinical trials

Complementary to the efforts of representing clinical trial entities is the automatic extraction of these entities. Most of the current work in information extraction from clinical trial reports is based on the preliminary step of sentence classification. Xu et al. use text classification and hidden Markov models (HMMs) to label sentences for the automatic structuring of clinical trial abstracts [XSH06]. Naive Bayes, maximum entropy, and decision tree classifiers were trained to categorize sentences into one of five classes: Introduction, Objective, Method, Result, and Conclusion.

In [CC07], Chung and Coiera use conditional random fields (CRFs) and support vector machines (SVMs) to label sentences in an abstract with one of five rhetorical roles (Aim, Method, Participants, Results, Conclusion). Using similar methods, Chung classified sentences referring to Intervention, Participants, and Outcome Measures. Classification features included unigram bag-of-words, part-of-speech tags, sentence position, features from previous and following sentences, and rhetorical roles. Kim et al. use CRFs to classify sentences into 6 categories: Background, Population, Intervention, Outcome, Study Design, and Other [KMC11]. In addition to features used by Chung, Kim et al. also used bigrams, semantic information from the UMLS, and section headings. Blake and Lucic [BL15] identified comparison sentences from full-text articles then used SVMs to extract the endpoints.

In [DCK08], de Bruijn et al. describe an architecture for extracting information elements from the full-text of clinical trial reports. The elements extracted are based on CONSORT Plus, an extension of the CONSORT statement that includes eligibility criteria, experimental and control treatments, intervention parameters, sample size, start and end date, primary and secondary outcomes, funding information, and publication details. The information extraction process consists of two stages – an SVM classifier that determines which sentences

contain information elements, and a regular expression based stage that extracts the exact information value. In [KBC10], Kiritchenko et al. describe ExaCT, an end-to-end system based on [DCK08] that highlights key information from clinical trial reports, and displays them in a web-based interface.

Less work has been done in this area without a sentence classification stage. Summerscales et al. frame the problem as a named-entity recognition task [SAB11]. The authors trained a CRF classifier for identifying treatments, treatment groups, and outcomes from *BMJ* abstracts. Features included part-of-speech, MeSH concept ID, semantic tags, position in the abstract, and features of surrounding words.

Chung examines the use of linguistic features (specifically, coordinating constructions) for the identification of intervention arms in RCTs in [Chu09]. Coordinating constructions consist of constituent phrases that are linked by conjunctions ("and," "or," "but"). A maximum entropy classifier was trained using syntactic and bag-of-word features.

In [DL07], Demner-Fushman and Lin describe extractors for elements in the PICO representation. The population, problem, and intervention extractors are rule-based; the outcome extractor consists of a set of supervised classifiers. Boudin et al. exploit document structure to extract PICO elements for an information retrieval task [BSN10]. The detection of PICO elements is framed as a sentence classification task. When available, document headings such as "Patients" or "Outcomes" are used to locate sentences containing PICO elements. Syntactic and semantic features are used to build a hybrid classifier consisting of multiple classifiers (decision tree, support vector machine, multi-layer perceptron, and naive Bayes). Dawes et al. identified commonly-occurring textual patterns which could be used for automatic extraction of Comparisons ("placebo," "compared," "than"), Outcomes ("mortality," "outcome," "incidence"), and Results ("differ," "increase," "significant") [DPS07].

There have been many approaches to the automatic extraction of study elements from scientific literature. Often these approaches are based on section and sentence classification, followed by rule- or pattern-based extraction of individual elements. Chapter 6 will present Casama's use of these techniques for extraction of contextual elements.

2.6 Creation of annotated gold standards

An annotated gold standard is a collection of documents that have been marked up by human readers with respect to a standardized representation of the knowledge contained within the documents. Annotated gold standards are valuable to the research community because they bridge the gap between human-level understanding of free text and computable representations. These gold standards can then be used as "ground truth" for computational tasks such as automatic concept extraction, relation extraction, and information retrieval. Typically, the development of annotated gold standards involves: a training phase, in which annotators work through a small number of sentences or documents together; an iterative annotation phase, in which annotators independently annotate the document collection; and an adjudication phase, in which disagreement between annotators is resolved. To ensure high quality of annotations and facilitate reproducibility, concepts and relations in the ontology are clearly defined and annotation guidelines provide instructions and examples for the annotators. Inter-rater agreement measures the "trustworthiness" of the gold standard. For document classification tasks (i.e., placing documents into pre-specified categories), the Kappa statistic is typically used [LK77]. For tasks involving the tagging of mentions in free text, F-score is a commonly used metric [HR05].

2.6.1 Annotated gold standards for biomedical concepts and relations

Several annotated gold standards exist in the domain of biomedicine, spanning a variety of subdomains. Gold standards annotated for biomedical concepts include PhenoCHF, the Mantra Gold Standard Corpus, the Colorado Richly Annotated Full-Text (CRAFT) corpus, and GENETAG. PhenoCHF [ATA14] is a corpus of biomedical articles and clinical notes annotated with phenotypic information on congestive heart failure. The Mantra Gold Standard Corpus [KCA15] is a multi-lingual corpus of 5,530 concept annotations from MEDLINE abstract titles, drug labels, and biomedical patents based on UMLS. The CRAFT corpus includes over 100,000 annotated entities in 97 full-text articles for concepts from nine biomedical ontologies [BEE12]. GENETAG is a corpus of 20,000 MEDLINE sentences tagged for

mentions of gene/protein names for the BioCreative shared task [TXT05].

In the domain of biomedical relations, Mihaila et al. developed BioCause, 851 mentions of causal relations from 19 full-text biomedical articles [MOP13]. The BioText corpora include sentences tagged for seven types of disease-treatment relations [RH04]. A gold standard corpus for chemical-induced disease extraction was developed for BioCreative V [LSJ15]. Kilicoglu et al. describe the SemRep gold standard [KRF11], 1,371 semantic predications based on the UMLS Semantic Network from 500 MEDLINE sentences. The Clinical E-Science Framework (CLEF) corpus focuses on conditions, investigations, interventions, and results found in clinical reports [RGH07].

Many gold standard corpora exist for bio-molecular events, particularly interactions between proteins and genes [RH05, BGK05, CMB09, PGH07]. The largest effort to date in annotating bio-molecular events is the GENIA corpus. Based on the GENIA ontology, the GENIA corpus includes nearly 100,000 annotated concepts from 1,000 MEDLINE abstracts. These entities participate in relations or biological "events" such as binding, localization, and regulation [KOT08]. GENIA also includes meta-knowledge annotations that indicate origin of knowledge, certainty, and negation [TNM11]. Other corpora for bio-molecular events include GREC (Gene Regulation Event Corpus) [TIM09] and PASBio [WSC04].

2.6.2 Justification for a Casama gold standard

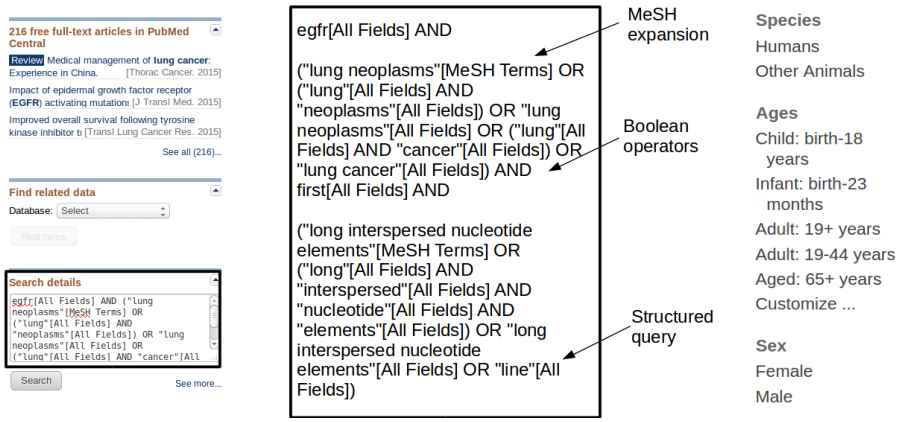
While several ontologies and annotated gold standards exist for biological concepts and events, each corpus focuses on its own set of targeted relations (e.g., treatment-disease, protein-protein, causal relations, bio-molecular events). The SemRep ontology includes the widest variety of clinically-oriented relations; however, at the time of this writing it does not contain a gold standard for all the relations targeted by Casama, such as relations between patient features and outcomes and the improvement of survival with therapy. Furthermore, few gold standards include an element of contextualization (the parameters under which relations held true). Thus, an annotated gold standard of relations and their contexts was developed for Casama; a description of this process is found in Chapter 4.

2.7 Patient-tailored information retrieval

The most familiar form of literature retrieval, from Google to PubMed, is via an ad hoc query, in which the user issues an unstructured, free text query to the search engine. Under this paradigm, the corpus of documents is often represented as an inverted index: a list of all the terms appearing in the corpus and links to the documents containing each term. For each term of each document, a term frequency $\times$ inverse document frequency (TFIDF) score is computed. This metric favors terms that appear frequently in the document and infrequently in the corpus, thus giving greater weight to rare terms. Documents and queries can be represented as vectors of TFIDF scores. Similarity between documents and queries is computed by taking the cosine of these vectors.

The vector space model is useful for general information retrieval because no domain knowledge is required. However, biomedical search engines such as PubMed include features that leverage the known document structure of scientific literature and the availability of standardized terminologies. For example, PubMed performs query expansion using Medical Subject Headings (MeSH). As illustrated in Figure 2.1a, a query for "lung cancer" would be automatically expanded to include the MeSH synonym "lung neoplasms." PubMed also supports several types of metadata that can be used to filter the result set in a structured manner. Figure 2.1b depicts PubMed's filters on coarse patient attributes, such as species, age, and sex.

Many have investigated the use of patient/population information to automatically match relevant research articles to individual patients. Early research focused on identifying the thought processes used by physicians to judge relevance of papers to individual patients. Florance interviewed three clinicians on their information seeking behavior and discovered that clinicians prioritized information such as patient age, sex, condition, history, and diagnosis, in addition to study-related information such as study size, study location, treatment applied, and outcomes [Flo92]. Rennels et al. developed Roundsman, a system that automatically critiques proposed therapy regimens for a given patient [RSS87]. Critiques were generated by means of a structured representation of the clinical literature that includes



(a) (b)

Figure 2.1: Screenshots from PubMed illustrating (a) query expansion, and (b) population filters.

sample size, interventions compared, and outcomes. Relevance of a study to an individual patient is assessed by means of a distance metric. In the domain of patient/treatment matching for alcoholism interventions, Finney and Moos identified three conceptual issues: selecting the patient/treatment matching variables, identifying the desired outcome, and determining when matching decisions should be made [FM86].

Other work has focused on integrating information from the electronic health record for information retrieval or identification of clinical trial candidates. ERGO, an eligibility criteria formalization system, matches patients to eligibility criteria by translating ERGO annotations to SQL and storing them in a relational database [TPC11]. Clinical trials for which a patient is eligible are retrieved by issuing a query corresponding to the patient's attributes. In [EKK05], Elhadad et al. describe PERSIVAL, a patient-tailored summarization system that extracts information from the patient record and matches it to the findings of the paper. Cimino et al. explore methods for mapping clinical data such as narrative text to standardized terminologies; these terms are used to explore information resources such as PubMed [CEZ97].

There has been some work in integrating patient/population information with standard vector space model of information retrieval. Boudin et al. developed automatic annotators for PICO elements ("P" standing for a coarse representation of Problem/Population) and use these to enhance retrieval [BSN10]. The authors discovered that retrieval could be improved

significantly despite only moderate accuracy of the annotators. Recently, the Text REtrieval Conference (TREC) Clinical Decision Support shared task sought to retrieve full-text articles from PubMed Central given a short, free text description of a patient case [RSD14]. During the inaugural run in 2014, most participants implemented traditional information retrieval techniques such as query expansion and vocabulary standardization. Five teams included a simple representation for patient attributes, consisting of age, sex, and race. Two of these teams (Soldaini et al. [SCY14], Garcia-Gathright et al. [GMH14]) saw an improvement in performance when including patient attributes in the retrieval process.

Notably, the systems described above use varying levels of granularity in representing patients/populations, depending on the overarching goal of the system. Representations of eligibility criteria use a fine-grained representation in order to discover strict matches to clinical trials. PERSIVAL tags relevant findings in papers and matches them to patient attributes; however, this process is based on syntactic patterns rather than a representation of salient patient features. Retrieval systems such as PubMed, Boudin's PICO-based system, and systems developed for TREC 2014 all use coarse representations of patients/populations (i.e., basic demographic features). In contrast, Casama incorporates a more fully developed patient representation (tailored specifically to lung cancer) to improve upon standard information retrieval methods. In doing so, more detailed queries may be posed and more relevant results retrieved. This is demonstrated in Chapter 7.

2.8 Automatic summarization

Researchers have been interested in automatic summarization as early as the 1950s [Luh58]. Previous work in automatic summarization has spanned multiple domains, including summarization of news articles [MMM97, RJB00, Nen05] and search engine results [RF00, KLR04, ZE99]. Automatic summarization of biomedical journal articles is an active research area, reviewed thoroughly by Mishra in [MBF14]. Mishra enumerates several trends among state-of-the-art summarization systems: 1) interest in multi-document summarization is increasing with the move toward evidence-based medicine; 2) most existing summarization

systems are extractive, but a significant minority of systems are abstractive; 3) knowledge-rich techniques are growing in popularity; 4) most systems used natural language processing methods, sometimes in combination with statistical or machine learning techniques; and 5) most evaluations were intrinsic (judged against a gold standard for accuracy, relevancy, etc.) rather than extrinsic (i.e., a task-oriented evaluation).

Casama, the system described in this dissertation, is abstractive and visual; thus, this portion of the literature review focuses on abstractive, visual summarization systems, followed by a brief discussion of evaluation.

2.8.1 Abstractive summarization systems

Automatic summarization can be divided into two general approaches: *extractive*, in which summaries are composed of fragments of the source material identified as containing salient information, and *abstractive*, in which a new natural language or visual summary is generated by the system. The majority of the summarization systems reviewed by Mishra were extractive. Among the abstractive summarization systems, most generated a textual summary.

The most significant work in abstractive, visual summarization is the National Library of Medicine's Semantic MEDLINE [RKF11]. Semantic MEDLINE uses a relational framework based on SemRep to summarize claims made in scientific literature. In [FRK04], Fiszman et al. describe an approach for transforming relations (or "semantic predications") into a graphical summary. Semantic MEDLINE utilizes four principles to select which predications should be included in the summary: *relevance* to the topic, *connectivity* of related predications, *novelty* of extracted knowledge, and *saliency* or high frequency of predications within the source text. These are determined by examining the graph-based or statistical features of the semantic network: relevance is the number of edges between a node and the central topic node; connectivity is the number of edges between a node and nodes adjacent to the central node; novelty is the distance of node from a more specific concept, and saliency is the frequency of the node (concept) within the document set.

Further work with Semantic MEDLINE has developed novel methods for focusing its graphical summaries. Zhang explored concepts from graph theory (degree centrality and clustering of cliques) for producing more focused summaries [ZFS11, ZFS13]. Workman introduces the Combo algorithm, a technique for refining the summary based on desired point-of-view (treatment, substance interaction, diagnosis, pharmacogenomics, and etiology) [WFH12].

A few other biomedical summarization systems are relation-oriented or perform visual abstraction to present their summaries. Telemakus [FRB04] exploits document structure, concept annotations produced by MetaMap, and relations extracted from tables and figures to represent claims in biomedical documents. AliBaba uses pattern matching and co-occurrence filtering to extract protein-protein, gene-gene, and drug-disease relations, among others. These relations are then visualized as a graph for real-time browsing of PubMed query results [PSP06]. BIOSQUASH, an extractive summarizer, produces a semantic graph to determine which propositions span multiple documents to aid the sentence selection process [SMW07]. Similarly, Morales et al. represent documents as a graph and cluster the sentences within the graph to determine which sentences are most significant [MEG08].

2.8.2 Evaluation of summarization systems

Mishra categorizes the evaluation of summarization systems into two groups: intrinsic and extrinsic. Intrinsic methods assess the quality of summaries in terms of comprehensiveness, accuracy, and relevance with respect to a gold standard. As no reference standards exist for summarization in biomedicine, usually the gold standards used in evaluation are produced manually in a proprietary fashion. Alternatively, some systems use knowledge sources (such as the abstracts of papers) as their gold standard. Common evaluation metrics include precision and recall; in the case of text-based summaries, ROUGE metrics (Recall-Oriented Understudy for Gisting Evaluation) are often used [Lin04]. Most of the systems reviewed by Mishra perform intrinsic evaluations. Extrinsic evaluations measure the task-oriented success of a system (e.g., time to completion, decision making accuracy, usability).

In [FDK09], Fiszman et al. performed semi-automated and manual evaluations of Semantic MEDLINE on the topic of pharmacological treatments for a variety of disease classes. A reference standard consisting of two resources for evidence-based medicine was used for the semi-automated evaluation. Performance was evaluated based on two metrics: mean average precision and "clinical usefulness score," a metric devised to positively weight beneficial treatments extracted by the system, and negatively weight harmful treatments. The performance was compared with that of a baseline system based on co-occurrence. The authors acknowledge the weaknesses of their summarization system: namely, identification of overly general concepts, incorrect mappings to Metathesaurus, and lack of information about the quality of evidence.

2.8.3 Summarization and Casama

Central to all summarization systems is the method for determining which information is the most relevant. Semantic MEDLINE relies on graph-based or statistical features of the semantic network to determine relevance, connectivity, novelty, and saliency of presented knowledge. Degree centrality measures focus the graph further. In contrast, Casama uses a semantic approach to visual summarization, in which study context or patient/population context can be leveraged by the user to focus the summary. Thus, Casama's summaries are both semantically-grounded and transparent: the user has ultimate control over which data he or she wishes to view. Furthermore, Casama addresses some of the weaknesses of Semantic MEDLINE identified by Fiszman et al. Inclusion of overly general concepts is mitigated by a more granular representation than that of Semantic MEDLINE, and strength of evidence is included explicitly in the Casama representation.

Casama follows many of the research trends identified by Mishra. Casama aggregates multiple documents to reveal current research directions, abstracts the relevant knowledge to produce a visual summary, uses knowledge of study design and population to enrich the summary semantically, combines lexical approaches with machine learning to extract relations and context, and is evaluated both extrinsically (with respect to an information

retrieval task) and intrinsically (using UpToDate as a manually-curated reference standard).

2.9 Conclusion

This work develops a novel paradigm for summarization by bringing together current research areas in clinical study representation, patient/population representation, relation extraction, and information retrieval. While each individual element offers a small contribution to the current body of knowledge, the sum of these parts makes a significant step forward in evidence-based and patient-tailored summarization through contextualized relations.

CHAPTER 3

Representation

3.1 Introduction

The formal definition for a contextualized semantic map — a set of concepts and relations tied to their contexts — is presented in this chapter. The concepts, relations, and contexts needed to instantiate a contextualized semantic map in the domain of driver mutations in non-small cell lung cancer (NSCLC) are described in detail.

3.2 Formal definition of contextualized semantic maps

A contextualized semantic map consists of the following:

Primitives

- relation_concept_types, a set of concept types that may participate in relations
- relation_names, a set of relation types
- context_concept_types, a set of concept types that provide context to relations

Frames

- context_frame, the collection of context_concept_types:

  $context\_concept\_type1 \in context\_concept\_types$

  $context\_concept\_type2 \in context\_concept\_types$

  ...

- contextualized_relation_frames, a set of frames comprising the following for each relation_name:

  subject_type $\in$ *relation_concept_types*

  object_type $\in$ *relation_concept_types*

  relation_name $\in$ *relation_names*

  context frame is_ a context_frame

Instances

An instance of a context frame contains a list of values for each context_concept_type:

context_instance:

is_ a context_frame

context_ concept_ type1 (inherited from parent)

context_ concept1 is_ a context_ concept_ type1

context_ concept_ type2 (inherited from parent)

context_ concept2 is_ a context_ concept_ type2

...

A contextualized semantic map is a set of instances of contextualized relation frames:

relation_instance:

is_ a contextualized_ relation_frame

subject_ type (inherited from parent)

object_ type (inherited from parent)

relation_ name (inherited from parent)

subject is_ a subject_ type

object is_ a object_ type

context instance is_ a context_ instance

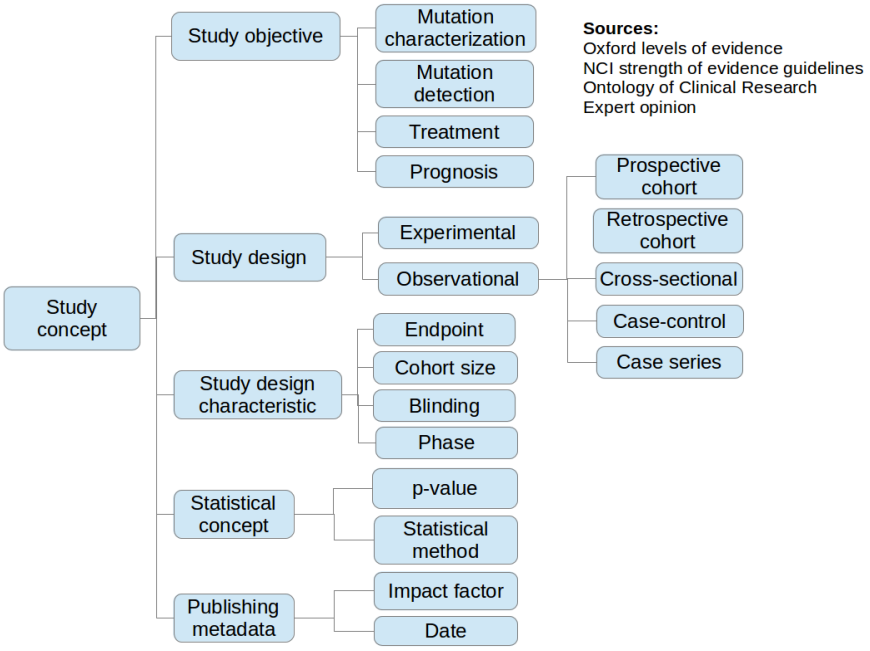


Figure 3.1: An ontological representation for study context. The leaves of this graph are the study context types.

The following sections define the context concepts, relation concepts, and relations needed to instantiate a contextualized semantic map in the domain of driver mutations in NSCLC. The Casama representation includes concepts for two types of context: study context and patient/population context. Relations of interest and concepts that may participate in these relations are then defined. Finally, two examples of contextualized semantic maps drawn from recent articles in PubMed are described.

3.3 Study context

Figure 3.1 illustrates Casama's representation for study context. Concepts were chosen based on guidelines for judging strength of evidence from the Oxford Centre for Evidence-Based Medicine [Evi11], the National Cancer Institute [Ins15], and expert opinion. The ontological organization of the representation was informed by the Ontology of Clinical Research [TCR09]. Study concepts are organized into five main classes: study objective, study design, study design characteristics, statistical concepts, and publishing metadata.

3.3.1 Study objective

Top-down and bottom-up strategies were used to identify key classes and elements that inform clinical decisions. The top-down aspect identifies clinical information needs by means of expert opinion. For NSCLC, a thoracic oncologist and thoracic radiologist specializing in lung cancer clinical trials were both asked to identify a set of patient-oriented questions perceived as being important in a clinical study. The questions were: 1) how likely is it that my patient has this mutation; 2) is there a treatment available for this mutation; and 3) is my patient likely to respond?

The bottom-up approach subsequently employs information gathered manually from the literature to suggest ways to stratify the document collection to enable retrieval of studies and to guide the representation of relations that address these questions. Four *study objective* classes were consequently identified: mutation characterization (relevant to question 1), mutation detection (question 1), treatment (question 2), and prognosis (question 3).

In *mutation characterization studies*, clinical-pathologic features (e.g., age, race, smoking history) are correlated with biomarker status (e.g., EGFR gene mutation). Other types of characterization papers report the prevalence of the mutation, either specifying the numerical prevalence within a population, the prevalence relative to other populations, or the prevalence with respect to other mutations.

Mutation detection studies describe analytical platofrms for detecting mutation status, such as polymerase chain reaction (PCR) or fluorescence in-situ hybridization (FISH). Mutation detection studies sometimes specify the type of biological specimen used by the analytical platform.

Treatment studies examine the association between treatments and outcomes. Treatments can improve outcomes (e.g., longer survival), worsen outcomes (e.g., side effects), or have no effect on outcomes. Treatments may also be recommended for a specific subpopulation.

Finally, *prognosis studies* associate clinical-pathologic features, biomarkers, and detection methods with outcome.

3.3.2 Study design

Representation of *study designs* was informed by a hierarchy of epidemiological study designs identified by the Oxford Centre for Evidence-Based Medicine [Evi11]. *Experimental* studies such as clinical trials provide the highest level of evidence, followed by several *observational* study types. *Prospective cohort* studies provide the highest level of evidence among the observational study types, followed by *retrospective cohort* studies, *cross-sectional* studies, *case control* studies, and *case series*.

3.3.3 Study design characteristics

Additional concepts that contribute to a clinician's judgment of strength of evidence include *endpoints* and *cohort size*. The National Cancer Institute has published guidelines for assessing quality of evidence in cancer treatment studies [Ins15]. Study endpoints, in descending order of strength, include: total mortality, cause-specific mortality, carefully assessed quality of life, and indirect surrogates (event-free survival, disease-free survival, progression-free survival, tumor response rate).

Cohort size is an attribute that may be helpful as a point of comparison across studies. In general, a larger cohort size suggests greater strength of evidence (although this comparison should be done carefully to account for differing study designs and effect sizes).

Phase and *blinding* are terms specific to clinical trials. Findings from advanced trials (phase III, IV) are more salient than those from early trials (phase I, II); double-blind studies are stronger than open label studies.

3.3.4 Statistical concepts

The *statistical test* concept informs the user of which methods were used to analyze the results; *p-value* is a metric used for judging the statistical significance of a test. A p-value of 0.05 or less is often used as a threshold of statistical significance; however, it should be noted that other factors such as study design should also be considered. Due to the complexity of

statistical knowledge, interpretation of statistical concepts is left to the user.

3.3.5 Publishing metadata

Publication date and *impact factor* of journal are two types of publication metadata that contribute to strength of evidence. Recent publications may be viewed more favorably, as these represent the latest discoveries in the field. Studies published in prestigious journals with high impact factors also play a role in judging the credibility of a study.

Formal definitions of study context terms are given in Table 3.1. When possible, definitions were borrowed from existing sources, such as the National Cancer Institute Thesaurus [SCH07] or Medical Subject Headings [NJH01].

3.4 Patient/population context

The attributes that compose the patient/population context are based on the National Lung Cancer Audit (LUCADA) [RTS11], an effort in the United Kingdom to create a registry of lung cancer patients and their treatments and outcomes. The LUCADA representation includes information on patient demographics, risk factors, treatment history, and tumor features. The representation was augmented based on expert opinion to include information on driver mutations, targeted therapy, imaging features, and clinical response. The organization of the representation was inspired by the National Cancer Institute Thesaurus [SCH07]. Patient/population information consists of three main classes: personal attributes, treatment-related concepts, and cancer features. An overview of the representation is given in Figure 3.2. Formal definitions are given in Table 3.2.

3.4.1 Personal attributes

Included in the representation are the most common demographic features found in the lung cancer literature — *age*, *sex*, *race*, and *smoking history*. Note that "age" may refer to a specific age group ("> 65 years of age") as well as general categories ("elderly", "younger"). Similarly,

Concept	Definition
Study designs	
Experimental	Studies in which individuals are assigned by an investigator based on a protocol to receive specific interventions.[1]
Prospective cohort	A research study that follows over time groups of individuals who are alike in many ways but differ by a certain characteristic and compares them for a particular outcome.[1]
Retrospective cohort	A research study in which the medical records of groups of individuals who are alike in many ways but differ by a certain characteristic are compared for a particular outcome.[1]
Cross-sectional	A study in which participants are examined at only a single time for characteristics of a disease.[1]
Case-control	A study that compares two groups of people: those with the disease or condition under study (cases) and a very similar group of people who do not have the disease or condition (controls).[1]
Case series	A group or series of case reports involving patients who were given similar treatment.[1]
Study design characteristics	
Cohort size	The number of units (persons, animals, patients, specified circumstances, etc.) in a population to be studied.[2]
Endpoint	Health events that lead to completion or termination of follow-up of an individual in a trial or cohort study.[3]
Phase I	Studies performed to evaluate the safety of diagnostic, therapeutic, or prophylactic drugs, devices, or techniques in healthy subjects and to determine the safe dosage range (if appropriate).[2]
Phase II	Studies that are usually controlled to assess the effectiveness and dosage (if appropriate) of diagnostic, therapeutic, or prophylactic drugs, devices, or techniques.[2]
Phase III	Comparative studies to verify the effectiveness of diagnostic, therapeutic, or prophylactic drugs, devices, or techniques determined in phase II studies.[2]
Phase IV	Planned post-marketing studies of diagnostic, therapeutic, or prophylactic drugs, devices, or techniques that have been approved for general sale.[2]
Open label study	A type of study that stipulates both the health provider and the subject be aware of the drug or treatment assignment.[1]
Blinded clinical study	A type of study in which the patients (single-blinded) or the patients and their doctors (double-blinded) do not know which drug or treatment is being given.[1]
Statistical concepts	
Statistical test	A test used to determine the statistical significance of an observation.[1]
P-value	A measure of the probability that a result happened by chance. The lower the p-value, the more likely it is that the result was caused by phenomenon of interest.[1]
Publication metadata	
Journal impact factor	A quantitative measure of the frequency on average with which articles in a journal have been cited in a given period of time.[2]

Table 3.1: Definitions of study design concepts.

Sources:

[1]National Cancer Institute

[2]Medical Subject Headings

[3]McMaster University Epidemiology Terms

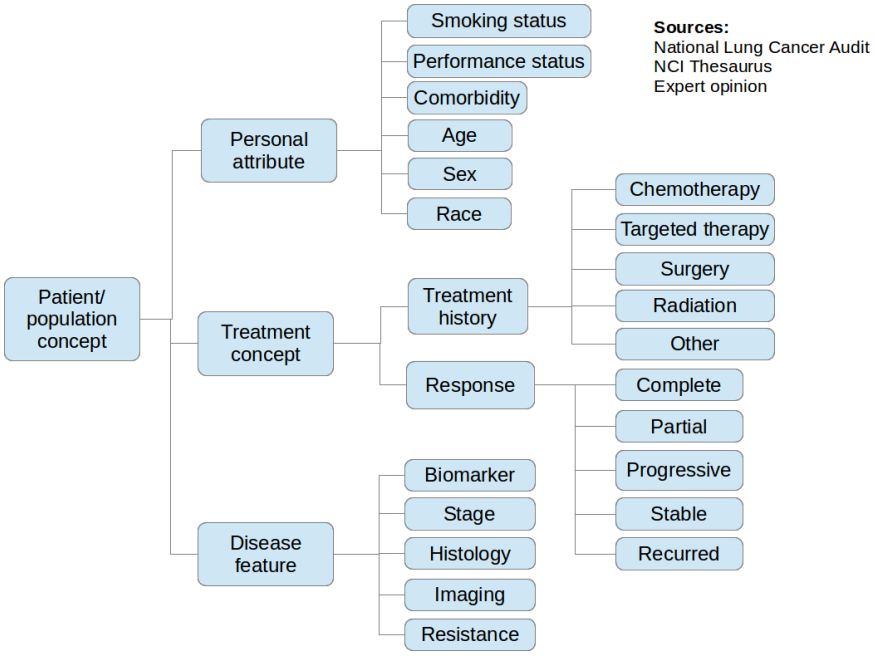


Figure 3.2: An ontological representation for patient/population context. The leaves of this graph are the patient/population context types.

race refers to any information indicating the racial background of the patient/population, either specifically (e.g., "Chinese") or generally (e.g., "Western population"). Smoking history categories (never smoking, former smoking, current smoking) are borrowed from the National Health Interview Survey [DP09]. Also included are qualitative terms describing smoking behavior (e.g., light smoking, heavy smoking).

Performance status may be qualititative ("fit") or quantitative by any metric (e.g., Karnofsky [SHG84], Eastern Cooperative Oncology Group [SKP93]). *Comorbidities*, as suggested by the LUCADA data manual, include additional conditions of the patient/population such as other malignancies, renal failure, weight loss, etc.

3.4.2 Cancer features

Features of disease are an important factor in lung cancer, consisting of several subfeatures: *biomarkers*, *clinical stage*, *histology* (as defined by World Health Organization/International Association for the Study of Lung Cancer classification of lung tumors [TBN11]), *resistance*, and *imaging features* as defined in the RadLex lexicon [Lan06]).

3.4.3 Treatment concepts

Treatment concepts consist of two subclasses: treatment history and treatment response. Treatment history is a highly relevant feature, as the history of successful or failed treatments can help suggest a strategy for future treatments. LUCADA identifies three major treatment types: *surgery*, *radiotherapy*, and *chemotherapy*. Because the domain of interest involves driver mutations in lung cancer, the Casama representation also includes *targeted therapy* (treatments targeting specific driver mutations).

Treatment response concepts correspond to clinical response categories defined by RECIST (Response Evaluation Criteria in Solid Tumours) [ETB09] and are always associated with the treatment history during which the response occurred. *Progression* is worsening or spreading of disease, whereas *recurrence* is a return of cancer after a period of no cancer (such as, after surgical resection). *Complete response* refers to complete disappearance of targeted tumors; *partial response* means shrinkage of tumor was observed; *stable disease* means there was little to no change to the number and size of tumors.

3.5 Comparison context

Casama also includes a special type of context for representing comparisons. In a controlled experiment, findings are often communicated in terms of a comparison to placebo or other treatment. Comparison context permits Casama to include information about a relation with respect to this comparator. Comparisons are represented as the context of the subject or object of the relation, and are always of the same type as the subject or object. Comparators can be especially useful when communicating negative findings, such as "gefitinib did not improve survival compared with erlotinib." To say that "gefitinib did not improve survival" would be inaccurate, as it may improve survival compared with chemotherapy or no treatment. Rather, this sentence expresses less improvement compared with another treatment, erlotinib.

Concept	Definition
Personal attributes	
Race	Shared heredity, physical attributes and behavior, and in the case of humans, by common history, nationality, or geographic distribution.[1]
Never smoker	An adult who has never smoked, or who has smoked less than 100 cigarettes in his or her lifetime.[3]
Former smoker	An adult who has smoked at least 100 cigarettes in his or her lifetime but who had quit smoking at the time of interview.[3]
Current smoker	An adult who has smoked 100 cigarettes in his or her lifetime and who currently smokes cigarettes.[3]
Performance status	A measure of how well a patient is able to perform ordinary tasks and carry out daily activities.[1]
Comorbidity	A concomitant but unrelated disease or pathologic process.[4]
Cancer features	
Biomarker	A characteristic that can be objectively measured and serves as an indicator for normal biologic processes, pathogenic processes, state of health or disease, the risk for disease development and/or prognosis, or responsiveness to a particular therapeutic intervention.[1]
Stage	The extent of a cancer in the body. Staging is usually based on the size of the tumor, whether lymph nodes contain cancer, and whether the cancer has spread from the original site to other parts of the body.[1]
Histology	The combined microscopic physical features of cells and their surrounding extracellular environment in tissues.[1]
Resistance	The failure of cancer cells, viruses, or bacteria to respond to a drug used to kill or weaken them.[1]
Treatment history	
Surgery	Operative procedures on organs, regions, or tissues in the treatment of diseases.[2]
Radiotherapy	The therapeutic use of ionizing and nonionizing radiation.[2]
Chemotherapy	The use of chemical-based agents to treat cancer. Antineoplastic chemotherapy works by arresting or killing the growth and spread of cancer cells.[1]
Targeted therapy	A type of treatment that uses drugs or other substances, such as monoclonal antibodies, to identify and attack specific cancer cells.[1]
Treatment response	
Progression	The worsening of a disease over time.[2]
Recurrence	The return of a sign, symptom, or disease after a remission.[5]
Complete response	The disappearance of all signs of cancer in response to treatment.[1]
Partial response	A decrease in the size of a tumor, or in the extent of cancer in the body, in response to treatment.[1]
Stable disease	Cancer that is neither decreasing nor increasing in extent or severity.[1]

Table 3.2: Definitions of patient/population concepts.

Sources:

[1]National Cancer Institute

[2]Medical Subject Headings

[3]National Health Interview Survey

[4]Computer Retrieval of Information on Scientific Projects

[5]Consumer Health Vocabulary

3.6 Relations

3.6.1 Definitions of relations

Four study objective types (mutation characterization, mutation detection, treatment, and prognosis) guide the selection of relations of interest. Relations appropriate for each study type are given below. With the exception of the **detects** relation (which is similar to the "diagnoses" relation of UMLS), relations defined for Casama had no analogs in the UMLS Semantic Network. In all definitions given below, relation names appear in **bold** and arguments of relations appear in *italics*. Example sentences expressing the relation are also given.

3.6.1.1 Characterization studies

The term "correlation" indicates a statistically significant finding between biomarkers and clinical features.

- *biomarker* **positive correlation** *clinical feature*: The biomarker and clinical feature tend to occur together. Example: "*EGFR mutation* rate was higher in *female* patients." [GCZ12]

- *biomarker* **negative correlation** *clinical feature*: The biomarker and clinical feature tend to not occur together. Example: "Number of *cigarette pack years* were significantly lower in patients with *EGFR mutations*." [BMZ13]

- *biomarker* **correlation** *clinical feature*: The biomarker and clinical feature are related, but the direction is not stated. Example: "*IGF1R/EGFR FISH+* correlates with *IGF1R/EGFR IHC+*." [LFB13]

- *biomarker* **no correlation** *clinical feature*: The biomarker and clinical feature appear to have no relationship. Example: "No significant correlation between *LKB1 alterations* and *mutations in EGFR pathway genes* was found." [LCN13]

Similarly, the "has rate in" relations describe relations between biomarkers and clinical features that trend toward positive or negative correlation, but without statistical significance. These relations include:

- *biomarker* **has higher rate in** *clinical feature.* Example: "Most of the tumors with *EGFR mutations* were *acinar with lepidic or papillary subtypes.*" [ZMO13]
- *biomarker* **has lower rate in** *clinical feature.* Example: "The frequency of *EGFR mutations* was lower in *African-American* patients compared to Caucasian patients but did not reach statistical significance." [BMZ13]
- *biomarker* **has similar rate in** *clinical feature.* Example: "The frequency of *EGFR mutations* is similar in *US Hispanics* compared with non-Hispanic whites." [ZMO13]

Casama includes several relations for describing the prevalence of a biomarker.

- *biomarker* **has rate** *rate*: This relation states the numerical prevalence of a biomarker. Example: "*EGFR mutation* was detected in *36.7%* of patients with NSCLC." [GCZ12]

The following relations compare prevalences between biomarkers. Example: "*EGFR exon 20 insertion* represents the third most common type of EGFR mutation after *exon 19 deletions* and *L858R.*" [ANC13]

- *biomarker* **has higher rate than** *biomarker*
- *biomarker* **has lower rate than** *biomarker*
- *biomarker* **has similar rate to** *biomarker*

3.6.1.2 Mutation detection studies

Mutation detection studies identify relations between analytical platforms for mutation detection, biomarkers, and biological materials. Relations associated with detection are:

- *detection method* **detects** *biomarker*. Example: "The *ARMS-TaqMan real-time PCR* method for the detection of *L858R mutations* was applicable in the clinical setting." [ZZZ13]

- *biomarker* **detected in** *material*. Example: "We have demonstrated the feasibility of using *cytological specimens* for *EGFR mutation* analysis." [CWC13]

- *detection method* **detects in** *material*. Example: "*Pyrosequencing* on *cytological blocks* is feasible." [SKU13]

3.6.1.3 Treatment studies

Relations associated with treatment studies are defined below.

- *treatment* **improves** *outcome*: The treatment led to a desired outcome. Example: "Patients receiving *erlotinib* experienced improvements in *quality of life*." [CFZ13]

- *treatment* **worsens** *outcomes* The treatment led to an unfavorable outcome (e.g., side effects). Example: "*Diarrhea, dysphagia*, and *sore mouth* were worse with *afatinib*." [YHS13]

- *treatment* **associated with** *outcome*: The treatment led to an outcome where "improves" or "worsens" is not appropriate. Example: "There was evidence of *anticancer activity* in relation to *matuzumab*." [HKC13]

- *treatment* **recommended for** *clinical feature*: A treatment is appropriate for a certain population. Example: "*EGFR tyrosine kinase inhibition* may be the treatment of choice for NSCLC patients with *miliary intrapulmonary carcinomatosis* at initial diagnosis." [WHC13]

Each of the treatment relations has a negated analog: **does not improve**, **does not worsen**, **not associated with**, **not recommended for**.

3.6.1.4 Prognosis studies

Finally, prognosis studies explore relations between biomarkers or clinical features and outcomes.

- *biomarker or clinical feature* **predicts better** *outcome*: The biomarker or clinical feature is associated with a more desirable outcome. Example: "*EGFR mutations* were associated with longer *overall survival*." [JSC13]
- *biomarker or clinical feature* **predicts worse** *outcome*: The biomarker or clinical feature is associated with an unfavorable outcome. Example: "*KRAS mutations* were associated with shorter *survival*." [JSC13]
- *biomarker or clinical feature* **predicts** *outcome*: An outcome is predicted, but the direction is not stated. Example: "*Metastatic status* was significantly associated with *survival* time." [DLW13]

The Casama representation includes a mechanism for distinguishing between prognostic factors (characteristics that predict outcome independent of treatment) and predictive factors (characteristics that predict response to treatment). A "predicts" relation that is tied to a *treatment* context indicates a predictive factor; otherwise, the feature is prognostic rather than predictive.

Each "predicts" relation also has a negated analog (**does not predict better**, **does not predict worse**, **does not predict**).

Formal definitions of concepts that participate in relations are provided in Table 3.3.

3.6.2 Hierarchical organization of relations

Note that Casama's granular relation types can be organized into broad families of relations: correlation, prediction, and treatment relations. This hierarchical organization, depicted in Figure 3.3, permits granular relations to be subsumed by their parents, regardless of

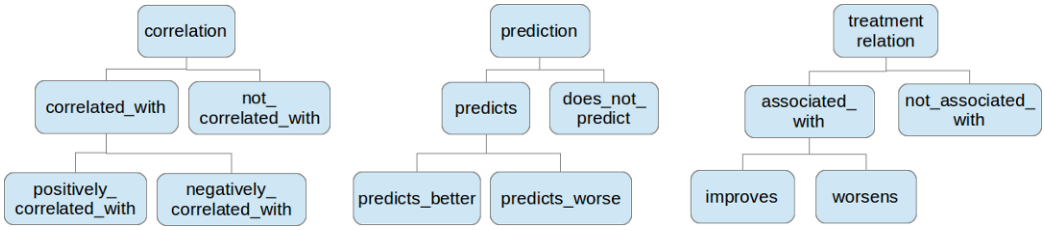


Figure 3.3: Hierarchical organization of relations associated with correlation, prediction, and treatment effects.

Concept	Definition
Biomarker	A characteristic that can be objectively measured and serves as an indicator for normal biologic processes, pathogenic processes, state of health or disease, the risk for disease development and/or prognosis, or responsiveness to a particular therapeutic intervention.[1]
Clinical feature	Clinical-pathologic features of a patient/population, same features as patient/population context representation.[2]
Detection method	A specific test or series of steps done to help diagnose a disease or condition.[1]
Material	A group or layer of cells that work together to perform a specific function.[1]
Outcome	A place of termination or completion. This may be a primary or secondary outcome variable used to judge the effectiveness of a treatment.[1]
Treatment	Any type of intervention intended to treat a condition in a patient.[1]
Rate	The ratio (for a given time period) of the number of occurrences of a disease or event to the number of units at risk in the population.[1]

Table 3.3: Definitions of concepts that may participate in relations.
Sources:
[1]National Cancer Institute
[2]National Lung Cancer Audit

polarity. This property is utilized in evaluating the manual annotation task (Chapter 4, Section 4.4.2.3) and automatic extraction task (Chapter 6, Section 6.2.1).

3.7 Contextualized relations

Any relation may be augmented with study or patient/population context, indicating the parameters under which the relation was found. Context may be found at the whole-study level (e.g., eligibility criteria, study design), indicating that the context applies to all relations for that study. Alternatively, context may be relation-specific, meaning that the context applies to a single relation only (e.g., p-values, population subgroups). A contextualized semantic map is the set of instances of these contextualized relations.

3.8 Discussion

The representation described here was designed to capture the main findings of a study and their associated contexts. The following example illustrates how these relations and contexts can be used to produce tailored graphical summaries of a collection of biomedical articles on lung cancer.

Consider the following relations discovered in three recent publications examining EGFR mutation in lung adenocarcinoma:

1. *Our results show that docetaxel improves progression-free survival for patients with NSCLC who have wild-type EGFR tumours.* [GMB13]

2. *The OPTIMAL study found that erlotinib improved progression-free survival in patients with EGFR mutation-positive non-small-cell lung cancer (NSCLC).* [CFZ13]

3. *The results proved a significant improvement in progression-free survival for patients harboring wild-type EGFR treated with the erlotinib-pemetrexed sequence.* [FPF13]

The following context-free relations may be extracted from these sentences:

1. docetaxel **improves** progression-free survival

2. erlotinib **improves** progression-free survival

3. erlotinib-pemetrexed sequence **improves** progression-free survival

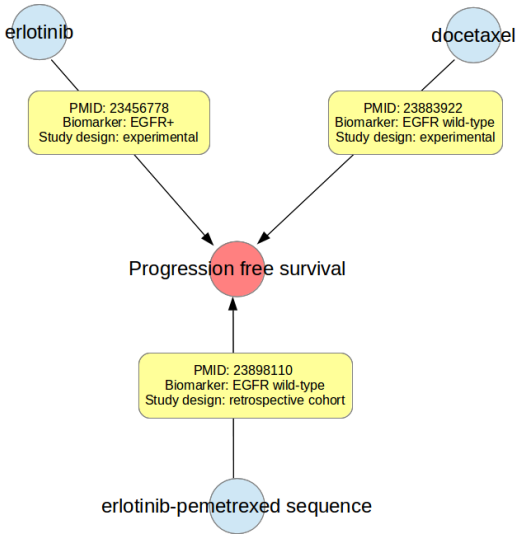
These relations can be aggregated and visualized as a contextualized semantic map, shown in Figure 3.4a. Each of these sentences occurred in different contexts — the first, in a clinical trial of EGFR wild-type patients; the second, in a clinical trial of EGFR mutation positive patients; the third, in an observational study of mixed EGFR+ and EGFR wild-type patients where the relation was found in the wild-type subgroup. By including this information as attributes of the edges of the semantic map, users can now pose more detailed queries than those accommodated by traditional, context-free graphs. They may choose to only include

results found in clinical trials, creating a threshold for a minimum level of evidence (Figure 3.4b). They can filter studies based on biomarker status (Figures 3.4c and 3.4d) in order to find studies relevant to a particular patient.

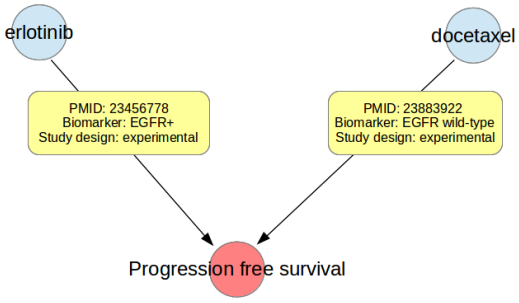
A more complex example (Figure 3.5) shows how contextualized semantic maps may be used to present an overview of current research from various perspectives. Two new relations are added to the semantic map (*tyrosine kinase inhibitors* **improve** *progression-free survival* and *afatanib* **improves** *progression-free survival*). In addition to contexts for biomarker and study design, the relations are further augmented with comparators, treatment history, and stage.

This example depicts five total relations; realistically, a semantic map may include dozens or hundreds of relations. Context may be used to filter the knowledge space, such that the remaining relations are specific to the user's information need. For example, only erlotinib and afatinib were demonstrated in clinical trials to be effective in EGFR+ patients. For EGFR wild-type patients, docetaxel was shown to be more effective than erlotinib. However, two retrospective studies found that tyrosine kinase inhibitors (such as erlotinib) improved progression-free survival in EGFR wild-type patients or patients of unknown EGFR status. Two of the five relations were found in study populations with a prior history of chemotherapy; these relations may be of interest to a clinician seeing a patient matching this history. All of the study populations included in this semantic map were of advanced stage.

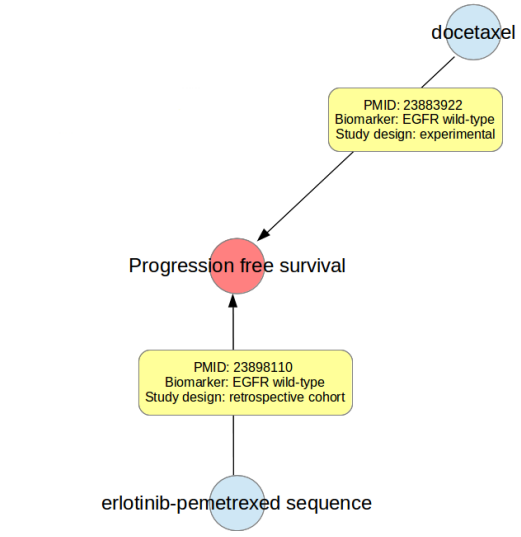
Context can also be used to resolve conflicting information (Figure 3.6). For instance, two studies examined the relationship between EGFR mutation and survival. However, one study found that EGFR mutation predicted better survival; another study claims that EGFR mutation does not predict survival. This apparently conflicting information can be resolved by examining the context – EGFR mutation predicted better survival in a cohort of patients with metastatic disease and a history of radiotherapy. In contrast, EGFR mutation did not predict survival in a cohort of early stage patients who received surgical resection.



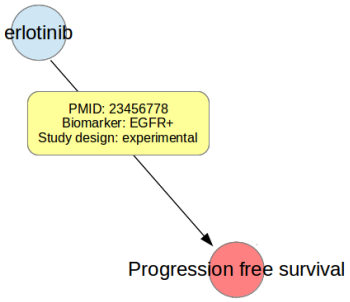
(a) All relations.



(b) Clinical trials only.



(c) biomarker = EGFR-wild type.



(d) biomarker = EGFR mutation positive.

Figure 3.4: A few examples of how semantic maps can be filtered through contextualization.

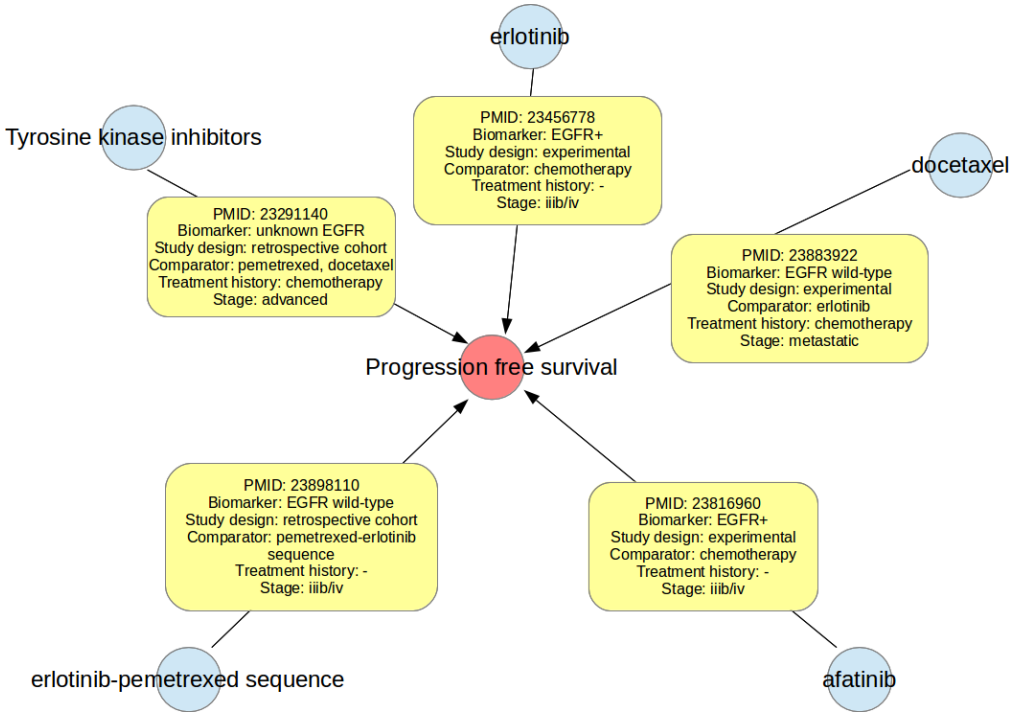


Figure 3.5: A more complex contextualized semantic map.

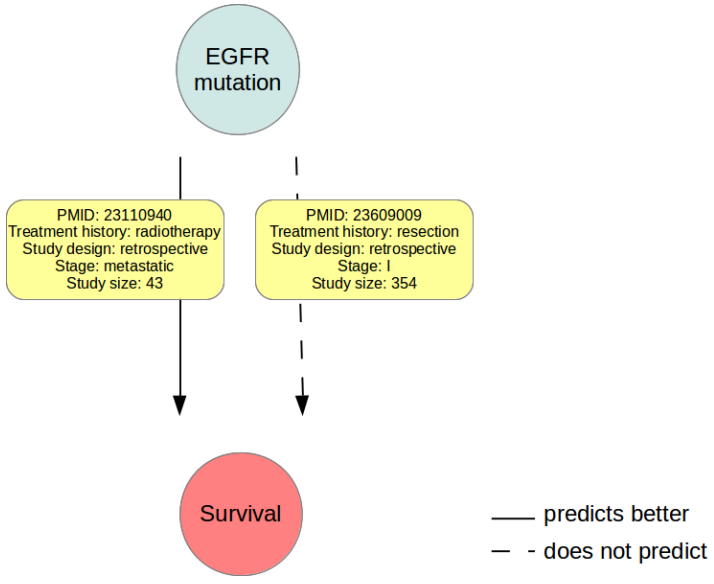


Figure 3.6: A contextualized semantic map that resolves seemingly discrepant findings.

3.9 Conclusion

This chapter described the representation of a contextualized semantic map in both its general and domain-specific forms. The representation presented here is unique in that it ties relations to their study and patient/population contexts. The following chapters will show how this representation enables the creation of powerful, contextualized summaries.

CHAPTER 4

Creating and evaluating an annotated gold standard

4.1 Introduction

This chapter describes the development of an annotated gold standard conforming to the representation described in Chapter 3. Recent abstracts on epidermal growth factor receptor (EGFR) mutation and anaplastic lymphoma kinase (ALK) rearrangement in non-small cell lung cancer were acquired. Two annotation studies were carried out: the first study classified papers by study objective and study design; the second annotation study tagged mentions of study population, relations, and remaining study concepts on a subset of the abstracts. Inter-annotator agreement was evaluated to validate the suitability of this annotated document set as a gold standard for automatic classification and extraction tasks.

4.2 Data collection

Both annotations tasks were performed against a snapshot of PubMed from September 2013. PubMed was searched for "EGFR" and "lung" in the titles of articles published between January 2012 and August 2013. Restricting the search to titles ensured that the retrieved abstracts belonged to the domain of lung cancer (as opposed to a study in another cancer domain that cites previous work on lung cancer in the abstract). Excluded from the search were empty abstracts, case reports, reviews, and pre-clinical studies. 211 studies on EGFR mutation in lung cancer were retrieved via PubMed. A similar query replacing "EGFR" with "ALK" resulted in 61 articles. The full PubMed query is given in Table 4.1.

Also included in the data set were abstracts from the "Non-Small Cell Lung Cancer -

Original query	egfr [Title] AND lung [Title] AND ("2012/01/01" [PDAT]:"2013/09/01" [PDAT])
Exclusion filter	NOT review [ptyp] AND hasabstract [text] NOT "cells" [title/abstract] NOT "cell lines" [title/abstract] NOT systematic [sb] NOT case reports [ptyp]

Table 4.1: Baseline PubMed queries for retrieving abstracts on EGFR mutation in lung cancer.

Metastatic" category of the American Society of Clinical Oncologists (ASCO) annual meetings from 2011-2013. This data source was chosen because of its high value as a source of information on current, clinically-oriented cancer research. A longer time frame was chosen compared to PubMed in order to retrieve a sufficient number of articles for automatic classification and extraction tasks. Similar to the PubMed query, the ASCO archive was searched for abstracts not containing "cell lines" whose titles contained "EGFR" or "ALK." 124 studies on EGFR and 34 studies on ALK were retrieved.

4.3 First annotation study: study objective and study design

4.3.1 Methods

A set of annotation guidelines was developed to enable annotation by multiple readers. One physician and four non-physicians with 0.5 - 2 years of clinical lung cancer research experience annotated the document collection. The document collection was divided into five sets of 86 abstracts each. Each annotator reviewed two sets; thus, each abstract was read by two annotators. The annotators placed each abstract into one or more study objective categories (mutation characterization, mutation detection, treatment, prognosis), and identified the methodological design of the study (experimental, prospective cohort, retrospective cohort, cross-sectional, case-control, case series).

In order to utilize the most information available for classification, annotators were permitted to consult the full-text of the article if available (21% of the document set). Typically, this information would be used to determine the finer details of a study design, such as whether a cohort study was prospective or retrospective. As full-text was not available for most studies, a catch-all class for cohort studies was used when the direction of inquiry was not stated.

Annotation was performed iteratively. After each round of annotation, agreement was calculated by Kappa analysis. Classes with low Kappa scores were targeted for discussion. The annotators met to identify differing interpretations of the guidelines, developing strategies for unifying their interpretations by talking through difficult cases.

The annotation guidelines were updated to remove ambiguities identified during the discussion. For instance, one point of disagreement involved whether naming the percentage of patients in a study who were EGFR-positive constituted a mutation characterization study. After a period of discussion, the annotators agreed that a study should only be considered a mutation characterization study if one of its aims was to identify the rate of mutation within a population, selecting its study population carefully for this purpose. Thus, the annotation guidelines were modified to specify this distinction.

Readers then re-annotated their sets of abstracts according to the revised annotation guidelines, and the process was repeated until sufficient agreement across the collection was reached. The Kappa scores presented here were obtained after three rounds of annotation. In order to produce a gold standard, two annotators were selected to resolve discrepancy. They viewed the annotations provided by the first pair of readers, and provided a tie-breaking vote. The two annotators were selected such that no annotator performed tie-breaking on a study for which he or she was one of the original annotators.

The counts in the gold standard for each category are summarized in Table 4.2.

4.3.2 Results

Table 4.3 details the inter-annotator agreement after three iterations of annotation. Kappa agreement for study objectives over all document subsets ranged from 0.518 to 0.846, indicating moderate to substantial agreement [VG05]. Standard deviations over each category ranged from 0.061 to 0.109. Mutation detection studies had the highest Kappa agreement at 0.792, while prognostic studies had a Kappa of 0.604. Over the entire document space and all study objectives, Kappa agreement was 0.684.

For the major classes of study design (experimental, cohort, cross-sectional), Kappa

Category	Counts: EGFR PubMed	Counts: ALK PubMed	Counts: EGFR ASCO	Counts: ALK ASCO
Mutation Characterization	74	26	40	8
Mutation Detection	35	20	14	7
Treatment	38	5	40	15
Prognosis	81	12	68	8
Experimental	20	3	27	10
Cohort (all)	89	14	63	12
Prospective cohort	7	1	1	0
Retrospective cohort	47	1	35	8
Unknown	35	12	27	4
Cross-sectional	60	27	20	10
Case-control	3	0	0	0
Case series	5	5	4	0

Table 4.2: Number of documents in the training and test sets for each study objective and study design type.

agreement ranged from 0.518 to 0.860, with intraclass standard deviations ranging from 0.031 to 0.128. Experimental studies had the highest overall Kappa score (0.728) while cohort studies had the lowest (0.608). Overall, the Kappa agreement for this subset of study design classes was 0.688.

Kappa agreement for the smaller study design types (subtypes of cohort studies, case control, case series) was significantly lower, with greater deviations from the mean. Of these, retrospective studies had the best agreement, ranging from 0.352 to 0.634, indicating fair to substantial agreement. For the study design classes that had less than 0.5 Kappa agreement, I reviewed the gold standard and confirmed that the value in the gold standard was in agreement with the annotation guidelines.

4.3.3 Discussion

Inter-rater agreement (per the Kappa score) for study objectives was moderate to substantial. One source of disagreement between annotators stemmed from the fact that studies could have more than one objective. Indeed, 86% of studies had at least one study objective that was agreed upon by both annotators; thus, primary objectives were "easy" to annotate whereas it was more difficult to determine secondary aims of a study. Also, some study objective classes differed from each other in subtle ways, such as mutation characterization

Category	Set A	Set B	Set C	Set D	Set E	Mean	Std Dev
Mutation Characterization	0.725	0.563	0.65	0.718	0.65	0.661	0.066
Mutation Detection	0.846	0.821	0.689	0.813	0.793	0.792	0.061
Treatment	0.634	0.552	0.705	0.649	0.845	0.677	0.109
Prognosis	0.606	0.518	0.643	0.725	0.527	0.604	0.086
Experimental	0.781	0.860	0.649	0.621	0.731	0.728	0.097
Cohort (all)	0.622	0.636	0.573	0.577	0.633	0.608	0.031
Retrospective Cohort	0.519	0.635	0.560	0.438	0.352	0.501	0.109
Prospective Cohort	0.378	0.488	0.312	0	0	0.236	0.224
Unknown Cohort	0.254	0	0.270	0.497	0.239	0.252	0.176
Cross-sectional	0.835	0.673	0.518	0.74	0.569	0.667	0.128
Case control	n/a	0	0	0	0	0	0
Case series	0	0.222	0.271	0.467	0.271	0.246	0.167

Table 4.3: Inter-rater agreement (Kappa) for the entire document collection. The collection was divided into five sets; each set was reviewed by two annotators.

and prognostic studies, which both aim to characterize various aspects of a mutation. This subtle difference is reflected in the lower Kappa score for prognostic studies.

Kappa scores for study design were moderate to substantial for the main study design classes. Experimental studies and cross-sectional studies had better Kappa agreement within this set of classes, as these are clearly associated with certain study types (clinical trials and mutation detection studies, respectively) and therefore were easier to agree upon. More granular study design types were more difficult to annotate. In particular, the difference between retrospective and prospective study designs was not often communicated clearly in abstracts. Annotators had varying levels of confidence in identifying cohort studies as prospective, whereas retrospective studies often stated their study design explicitly.

4.4 Second annotation study: Study context, population context, and relations

4.4.1 Methods

The remainder of the annotations were performed by myself and a graduate student specializing in medical informatics. The brat rapid annotation tool [SPT12] was used to highlight the spans of text in the abstracts that referred to a contextual element or relation. Com-

pared to the first annotation study, this second study was wider in scope (more elements to annotate) and more labor-intensive (highlighting spans rather than simple classification). Thus, a subset of the data was selected for annotation. For this task, only the EGFR data set was used. Furthermore, only abstracts published in 2013 were included. Excluding the "out of scope" abstracts, as determined in the first annotation study, there were 99 PubMed abstracts and 36 ASCO abstracts.

Study context, population context, and relations were converted to brat format (i.e., a list of concepts and relations, where relations are constrained by permitted concept type). Annotation guidelines were composed to formalize the annotation process. The annotation guidelines provided definitions of each concept type and relation, as well as specific directions regarding the scope of the annotations. For example, population context is limited to eligibility criteria and populations under which relations were observed, rather than descriptive statistics of the cohort. Also out of scope are relations for which interpretation of numeric data is required. For example, the sentence "Progression-free survival was 6 months for erlotinib vs. 3 months for docetaxel" would not be annotated, whereas "Progression free survival was longer with erlotinib vs. docetaxel" constitutes a valid relation. The underlying assumption is that key findings are often restated qualitatively in the Conclusion section of an abstract, thus simplifying automatic extraction by obviating the need to interpret quantitative data.

Because information is often repeated in different sections of an abstract, the guidelines also provided suggestions for where to expect certain types of information. For instance, mentions of population context are likely to be found in Background and Methods sections, whereas study context may be found in the Methods or Results sections. While exceptions to these rules did exist in the data, these suggestions aimed to enable better agreement between annotators in the case of repeated information. Sentences expressing the findings of the study were assumed to be found in the Results and Conclusion sentences of the abstracts. Thus, only these sections were annotated for relations.

During the training phase, the annotators reviewed ten papers together, noting and correcting ambiguities in the annotation guidelines prior to independently annotating the

entire corpus. Agreement was calculated on this set, and entities with low agreement were selected for discussion. For instance, one area of disagreement was appropriate usage of the *treatment* **recommended for** *population* relation. Consider the sentence, "These results show clinical benefit with afatinib in EGFR+ patients." One reader annotated, "afatinib **associated with** clinical benefit in EGFR+ patients," while the other annotated, "afatinib **recommended for** EGFR+ patients." Discussion of this example led to the conclusion that because the sentence is not an explicit recommendation of treatment, the first annotation type should be used. Consequently, the agreed-upon definition of "outcome" should include general terms such as "benefit." After discussing several points of clarification, annotation guidelines were again updated. The readers corrected their annotations until consensus was achieved.

Annotator agreement was calculated in terms of F1-score, holding my annotations as the ground truth for the purposes of evaluation. The Kappa statistic was not used because the nature of the task ensured that the chance of random agreement was very low, and Hripcsak showed that F1-score approaches Kappa as the number of negative cases grows large [HR05]. For study context and population context, agreement was calculated using document-level matching (i.e., annotations may appear anywhere in the abstract). For relations, agreement was calculated for each relation as well as semantically related groups of relations (e.g., **improves** and **associated with**, **predicts better** and **predicts**, **positive correlation** and **correlation**, see: Chapter 3, Figure 3.3). To create a gold standard, discrepant annotations were reviewed and resolved by discussion between the annotators.

4.4.2 Results

4.4.2.1 Study context

Annotation agreement for study context was substantial, with an overall F1-score agreement of 0.88. Full results are given in Table 4.4.

Concept	Total (both annotators)	Precision	Recall	F1-score
p-value	136	0.95	0.86	0.90
cohort size	119	0.89	0.88	0.88
endpoint	98	0.83	0.90	0.86
statistical test	46	0.87	0.84	0.85
phase	12	0.86	1.0	0.92
blinding	3	1.0	1.0	1.0
All concepts	414	0.89	0.88	0.88

Table 4.4: Annotator agreement for study context types.

4.4.2.2 Population context

Of the 20 population context types, 10 types contributed to over 90% of the total annotations. Among these, F1-score agreement ranged from 0.56 for *progression* to 0.92 for *surgery history*. F1-score agreement over all concept types was 0.78. Full results are given in Table 4.5.

4.4.2.3 Relations

Of the 27 relation types, 13 contributed to roughly 90% of the total annotations. Among these, F1-score agreement ranged from 0.40 for **associated with** to 0.80 for **does not predict** and **predicts worse**. Overall agreement for relations was 0.68. Combining semantically similar relations proved beneficial to F1-score agreement. Notably, the combined relation **improves or associated with** had an F1-score of 0.79. Full results are given in Table 4.6.

4.4.3 Discussion

During the consensus phase of the annotation task, major causes of annotator disagreement were noted. First, human error contributed to annotations being missed completely. Second, annotators had differing interpretations of the annotation guidelines, indicating that the guidelines were vague in some cases. These cases were identified and resolved for the final guidelines and gold standard. Finally, the inherent ambiguity of language made some sentences impossible to resolve, even by human readers (Figure 4.1).

One finding from the annotation study was that the representation was overspecified — in some cases, there was more than one way to express a concept or relation. Indeed, the

Concept	Total (both annotators)	Precision	Recall	F1-score
biomarker	151	0.68	0.88	0.77
stage	146	0.80	0.85	0.82
histology	93	0.75	0.90	0.82
targeted therapy history	91	0.78	0.80	0.79
race	67	1.0	0.63	0.78
chemotherapy history	40	0.94	0.77	0.85
progression	32	0.53	0.60	0.56
other treatment history	29	0.56	0.91	0.69
resistance	27	0.71	0.77	0.74
surgery history	26	0.86	1.0	0.92
performance status	11	0.83	1.0	0.91
radiotherapy history	9	0.80	1.0	0.89
smoking history	9	0.60	0.75	0.67
other clinical feature	6	0.50	0.25	0.33
recurrence	6	0.67	0.67	0.67
sex	6	0.25	0.50	0.33
age	4	1.0	1.0	1.0
comorbidity	2	0.0	N/a	N/a
partial response	2	1.0	1.0	1.0
stable disease	2	1.0	1.0	1.0
All concepts	759	0.75	0.81	0.78

Table 4.5: Annotator agreement for population concepts.

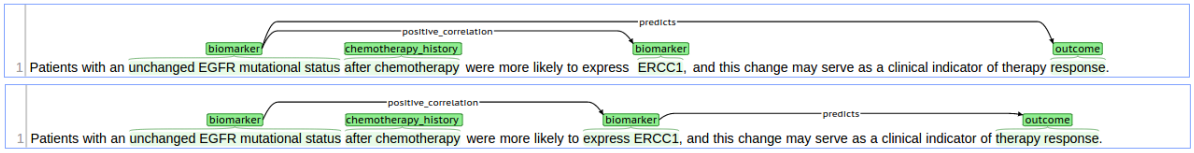


Figure 4.1: An example of annotator disagreement caused by the inherent ambiguity of natural language. Annotator 1 (top) and Annotator 2 (bottom) disagreed on whether response was predicted by unchanged EGFR mutation status or expression of ERCC1.

Relation	Total (both annotators)	Precision	Recall	F1-score
has rate	207	0.60	0.90	0.72
improves	131	0.88	0.70	0.78
does not predict	113	0.87	0.74	0.80
predicts better	112	0.82	0.66	0.73
predicts worse	112	0.92	0.71	0.80
predicts	107	0.61	0.67	0.64
detects	94	0.66	0.58	0.62
positive correlation	93	0.80	0.63	0.71
detected in	56	0.67	0.77	0.71
correlation	47	0.48	0.78	0.60
associated with	45	0.33	0.50	0.40
has higher rate in	41	0.87	0.50	0.63
recommended for	25	0.75	0.69	0.72
not associated with	21	0.14	0.29	0.19
detects in	18	0.50	0.25	0.33
has higher rate than	16	1.0	1.0	1.0
no correlation	16	1.0	0.60	0.75
worsens	16	0.07	1.0	0.13
negative correlation	14	0.60	0.33	0.49
has lower rate in	12	0.50	0.25	0.33
does not improve	8	0.33	0.20	0.25
has lower rate than	5	0.67	1.0	0.80
has similar rate in	5	0.33	0.50	0.40
does not predict better	2	1.0	1.0	1.0
not recommended for	2	1.0	1.0	1.0
does not predict worse	1	0.0	N/a	N/a
has similar rate to	1	0.0	N/a	N/a
All relations	1320	0.68	0.68	0.68
Combined relations				
predicts better or predicts	215	0.82	0.80	0.81
improves or associated with	173	0.80	0.77	0.79
positive correlation or correlation	140	0.84	0.84	0.84

Table 4.6: Annotator agreement for relations.

representation was designed for maximal expressiveness, in which relations are organized hierarchically by level of granularity (Chapter 3, Figure 3.3). For example, "erlotinib **improves** progression-free survival" could equivalently be annotated as "erlotinib **associated with** longer progression-free survival." Thus, higher precision and lower recall for **improves** were observed; conversely, **associated with** had lower precision and higher recall. While the guidelines stated that the most specific relation should be used, variation between annotators still came into play when making these types of judgments.

Lower agreement was also observed for less common relation types, such as **not associated with**, **detects in**, and **worsens**. As in any task, performance improves with experience, and annotators were more likely to disagree when faced with a rare case. However, the rarity of these cases ensured that they did not contribute significantly to the overall agreement score.

4.5 Conclusion

This chapter presented the annotation and evaluation of a gold standard corpus for the Casama representation. While annotation tasks are challenging by nature, the results show that Casama's representation is well-defined and specific enough to produce moderate to substantial inter-rating agreement for study objective (κ=0.68), study design (κ=0.69), relations (F1=0.68), study context (F1=0.88), and population context (F1=0.78). Certain types of annotations showed lower agreement, such as secondary study objectives, sub-types of cohort studies, and more granular relation types such as **associated with** vs. **improves**. However, these disagreements were resolved manually for the final gold standard, establishing a ground truth for automatic classification and extraction tasks described in Chapters 5 and 6.

CHAPTER 5

Automatic Classification

5.1 Introduction

This chapter describes the automatic classification of abstracts by the Casama concepts of study objective and study design. A document classification algorithm as developed by Joachims using support vector machines (SVMs) was implemented [Joa02]; this algorithm has been used by many for automatic text classification in the biomedcal domain [DMD03, PNR05, YP05, CMS06, YCD08, WTL10, YXT10, KC14]. Classification performance was compared to that of PubMed's Clinical Queries, a set of Boolean filters derived by empirically discovering combinations of search terms that yield optimal sensitivity and specificity [HMW05].

Classification performance was evaluated against the gold standard described in Chapter 4, Section 4.3. The classification results and the top features of the classifiers were examined to determine whether this scheme could generalize to other mutations in lung cancer and studies on driver mutations in other cancer domains.

5.2 Methods

5.2.1 Document classification algorithm

Each article in the document set is associated with one or more study objectives (mutation characterization, mutation detection, treatment, prognosis), whereas articles are assigned a single study design label (experimental, prospective cohort, retrospective cohort, cross-sectional, case-control, case-series). Described below are the steps involved in pre-processing

the document collection, performing the classification using SVMs, and evaluating performance on the training set (EGFR PubMed) and test sets (ALK PubMed, EGFR ASCO, ALK ASCO).

Joachims' document classification algorithm was implemented using Python's natural language toolkit (NLTK) and machine learning package scikit-learn [BKL09, PVG11]. If full-text was available for an article, the patient-selection portion of the Methods section (determined by matching regular expressions to the section headings) was concatenated with the abstract in order to improve detection of study design.

NLTK pre-processed the text by stemming and removing stop words. Unigram and bigram frequency distributions over the document collection were calculated; a binary feature vector indicating whether each unigram or bigram appeared in the text was created for each abstract.

ASCO abstracts were further processed to expand common abbreviations (e.g., "pts" for "patients", "PFS" for "progression-free survival"). Regular expressions were also used to detect abbreviation definitions (e.g., "patients previously treated with E (erlotinib)"; abbreviations were then replaced with their full names.

To classify study objective, scikit-learn trained a set of two-class linear-kernel SVMs; each SVM in the set corresponded to one of the study objective classes. The hyperplane constructed by each SVM was used to decide whether the document belonged in the corresponding study objective class or not.

A multi-class, one-versus-rest SVM was trained to classify documents by study design. The multiple study design classes were reduced to a set of binary SVMs; each abstract was classified according to the SVM that produced the highest output score. For study design classes with very few training examples (case-control studies, case series, sub-types of cohort studies), documents were classified by a set of hand-crafted rules, as described in Table 5.1.

5-fold cross validation was performed on the training set (EGFR PubMed); precision and recall across the folds were calculated. To test the performance of the classifier to previously unseen data, the SVMs were then trained on the entire training set and tested on the ALK

Study design	Extraction rules
Retrospective	title/abstract contains "retrospective" OR "review" OR "data" OR "charts" OR "records" OR "analyze"
Prospective	title/abstract contains "prospective"
Unknown cohort	any cohort study not matching rules for retrospective or prospective study
Case-control	title/abstract contains "case" AND "control"
Case series	title/abstract contains "series"

Table 5.1: Casama uses hand-crafted rules to extract sparsely-represented study designs.

PubMed, EGFR ASCO, and ALK ASCO sets.

The generalizability of these classifiers was further assessed by examining the most discriminative features of the linear-kernel SVM. Features with the highest-magnitude coefficients were considered highly discriminative. Features that are not domain-specific suggest the potential for generalizability. Study design classes that were classified by rules were not included in the analysis of top features.

5.2.2 Creating a baseline set for comparison

A baseline for classification performance was calculated by evaluating PubMed's filters against the manually-annotated input set of EGFR PubMed abstracts. PubMed Clinical Queries or Medical Genetics filters analogous to Casama's categories were applied to the original PubMed query (see: Chapter 4, Table 4.1), resulting in a subset of retrieved documents. For each filter, the retrieved documents were matched by PMID (PubMed identifier) to the annotated set; the number of results in each Casama category was then tabulated to calculate precision and recall. Newly added studies that were not found in the original set (i.e., studies that were added between the time of retrieval in September 2013 and the time of evaluation) were excluded. The PubMed queries examined are summarized in Tables 5.2-5.3.

PubMed Filter	**Query**
Clinical Description	Natural History OR Mortality OR Phenotype OR Prevalence OR Penetrance AND Genetics
Genetic Testing	DNA Mutational Analysis OR Laboratory techniques and procedures OR Genetic Markers OR diagnosis OR testing OR test OR screening OR mutagenicity tests OR genetic techniques OR molecular diagnostic techniques AND genetics
Diagnosis (broad)	sensitiv*[Title/Abstract] OR sensitivity and specificity[MeSH Terms] OR diagnose[Title/Abstract] OR diagnosed[Title/Abstract] OR diagnoses[Title/Abstract] OR diagnosing[Title/Abstract] OR diagnosis[Title/Abstract] OR diagnostic[Title/Abstract] OR diagnosis[MeSH:noexp] OR diagnostic * [MeSH:noexp] OR diagnosis,differential[MeSH:noexp] OR diagnosis[Subheading:noexp]
Diagnosis (narrow)	specificity[Title/Abstract]
Therapy (narrow)	randomized controlled trial [Publication Type] OR (randomized [Title/Abstract] AND controlled [Title/Abstract] AND trial [Title/Abstract])
Therapy (broad)	(clinical [Title/Abstract] AND trial [Title/Abstract]) OR clinical trials [MeSH Terms] OR clinical trial [Publication Type] OR random* [Title/Abstract] OR random allocation [MeSH Terms] OR therapeutic use [MeSH Subheading]
Management	Therapy [Subheading] OR treatment [Text Word] OR treatment outcome OR investigational therapies AND Genetics
Etiology (broad)	risk*[Title/Abstract] OR risk*[MeSH:noexp] OR risk *[MeSH:noexp] OR cohort studies[MeSH Terms] OR group[Text Word] OR groups[Text Word] OR grouped [Text Word]
Etiology (narrow)	relative[Title/Abstract] AND risk*[Title/Abstract]) OR (relative risk[Text Word]) OR risks[Text Word] OR cohort studies[MeSH:noexp] OR (cohort[Title/Abstract] AND study[Title/Abstract]) OR (cohort[Title/Abstract] AND studies[Title/Abstract]

Table 5.2: PubMed Clinical Queries and Medical Genetics filters, discovered by Haynes in [HMW05].

Casama Category	Analogous PubMed Query
Mutation characterization	Original query + Exclusion filter + Clinical Description [filter]
Mutation detection	Original query + Exclusion filter + Genetic Testing [filter] Original query + Exclusion filter + Diagnosis/Broad [filter] Original query + Exclusion filter + Diagnosis/Narrow [filter]
Treatment	Original query + Exclusion filter + Therapy/Broad [filter] Original query + Exclusion filter + Therapy/Narrow [filter] Original query + Exclusion filter + Management [filter]
Prognosis	Original query + Exclusion filter + Prognosis/Broad [filter] Original query + Exclusion filter + Prognosis/Narrow [filter]
Experimental studies	Original query + Exclusion filter + Clinical Trial [ptyp]
Cohort studies	Original query + Exclusion filter + Etiology/Broad [filter] Original query + Exclusion filter + Etiology/Narrow [filter] Original query + Exclusion filter + "cohort studies" [MeSH]
Prospective cohort studies	Original query + Exclusion filter + "cohort studies" [MeSH] AND "prospective studies" [MeSH]
Retrospective cohort studies	Original query + Exclusion filter + "cohort studies" [MeSH] AND "retrospective studies" [MeSH]
Cross-sectional studies	Original query + Exclusion filter + "cross-sectional studies" [MeSH]
Case-control studies	Original query + Exclusion filter + "case-control studies" [MeSH]

Table 5.3: Map of Casama categories to PubMed queries. Original query and exclusion filters can be found in Chapter 4, Table 4.1.

5.3 Results

5.3.1 Study objective classification

Table 5.4 presents the results of Casama's automatic classification of its four study objective categories (mutation characterization, mutation detection, treatment, prognosis), and compares them to PubMed's results with analogous filters. Casama outperformed PubMed in all categories based on 5-fold cross validation. Classification of study objectives had better F1-scores (balanced precision and recall) than PubMed's narrow filters (high precision, low recall) and its broad filters (high recall, low precision). As shown in Table 5.5, there was a decrease in performance on the test sets compared to the training set.

Receiver operating characteristic (ROC) curves for study objective classification by Casama are presented in Figure 5.1. (ROC curves for PubMed are not available as only a single set of results is returned per query.)

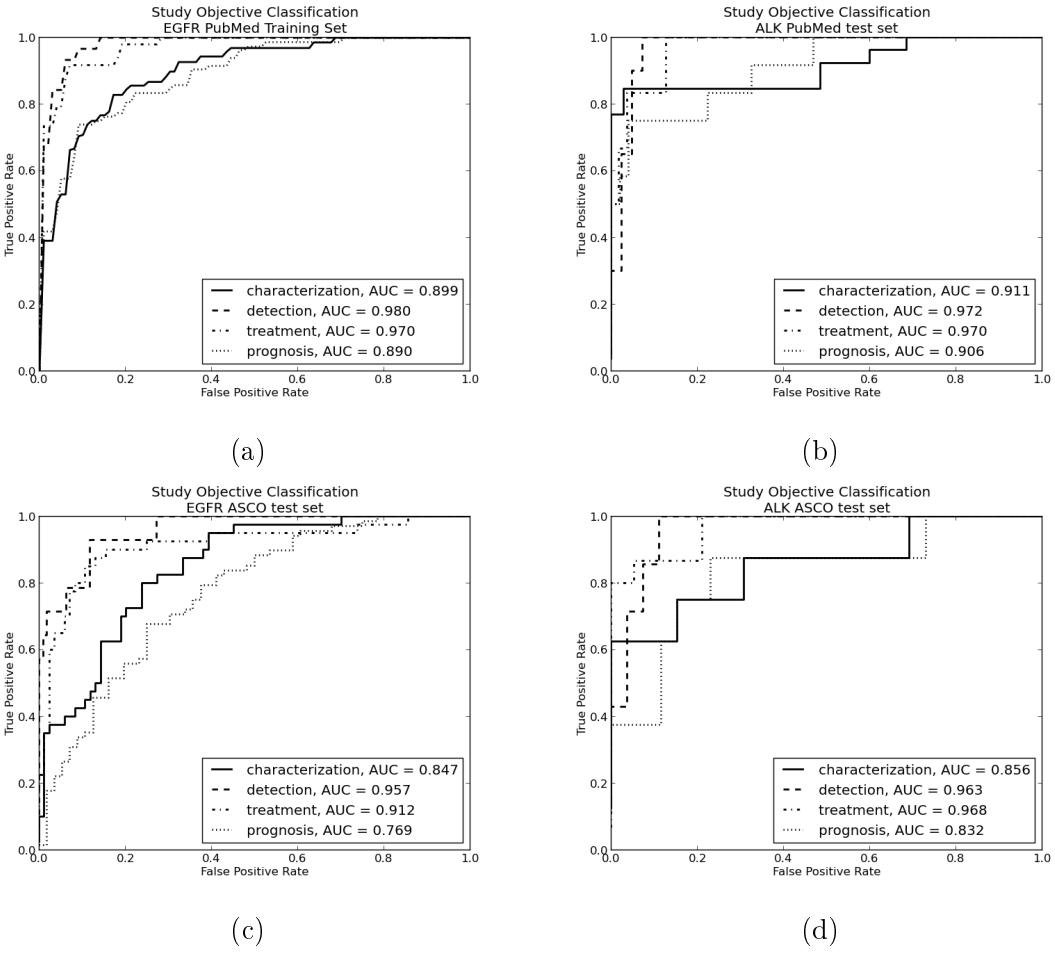


Figure 5.1: Casama's receiver operating characteristic and area under the curve for study objective classification on (a) EGFR PubMed, (b) ALK PubMed, (c) EGFR ASCO, (d) ALK ASCO.

Casama Cross Validation				PubMed Filter			
	P	R	F1		P	R	F1
Mutation characterization	0.82	0.68	0.74	Clinical description	0.48	0.39	0.43
Mutation detection	0.96	0.69	0.80	Genetic testing	0.22	0.94	0.35
				Diagnosis (broad)	0.36	0.89	0.52
				Diagnosis (narrow)	0.92	0.31	0.47
Treatment	0.84	0.71	0.77	Therapy (broad)	0.35	0.95	0.51
				Therapy (narrow)	0.77	0.26	0.39
				Management	0.25	0.71	0.37
Prognosis	0.76	0.77	0.76	Prognosis (broad)	0.58	0.78	0.67
				Prognosis (narrow)	0.71	0.51	0.59

Table 5.4: Comparison of precision (P), recall (R), and F1-scores (F1) between Casama and PubMed filters for study objective classification.

	ALK PubMed			EGFR ASCO			ALK ASCO		
	P	R	F1	P	R	F1	P	R	F1
Mutation characterization	0.80	0.62	0.69	0.66	0.48	0.55	0.56	0.63	0.59
Mutation detection	0.93	0.65	0.76	0.82	0.64	0.72	0.75	0.43	0.55
Treatment	0.67	0.67	0.67	0.82	0.78	0.79	1.0	0.80	0.89
Prognosis	0.77	0.58	0.67	0.76	0.65	0.70	1.0	0.38	0.55

Table 5.5: Casama's precision (P), recall (R), and F1-scores (F1) on test sets for study objective classification.

5.3.2 Study Design Classification

Tables 5.6 and 5.7 summarize the results of Casama's study design classifier. In Table 5.6, classification performance is compared to that of PubMed's filters (if available).

Casama outperformed PubMed in classification of cross-sectional studies, cohort studies, and prospective cohort studies. Casama's performance was similar to PubMed in retrieval of experimental and retrospective cohort studies. PubMed slightly outperformed Casama in classification of case-control studies. Rule-based classification worked best for retrospective studies; for the remaining classes, F1-scores were less than 0.50. There was no degradation in performance between the training and test sets.

5.3.3 Representational Class Features

Tables 5.8 and 5.9 specify the top features used to discriminate between each pair of study objective classes. Characterization studies aim to find *correlations* with mutation *status*;

Casama Cross Validation				PubMed Filter			
	P	R	F1		P	R	F1
Experimental	0.46	0.65	0.54	Clinical trials	0.48	0.60	0.53
Cross-sectional	0.77	0.80	0.79	Cross-sectional (MeSH)	0	0	0
Cohort (all)	0.81	0.72	0.76	Cohort (MeSH)	0.65	0.48	0.55
				Etiology (broad)	0.68	0.54	0.61
				Etiology (narrow)	0.60	0.10	0.17
Prospective Cohort	0.67	0.29	0.40	Prospective cohort (MeSH)	0.14	0.29	0.19
Retrospective Cohort	0.53	0.66	0.59	Retrospective cohort (MeSH)	0.63	0.57	0.60
Unknown Cohort	0.44	0.23	0.30				
Case-control	n/a	0	n/a	Case-control (MeSH)	0.05	0.67	0.08
Case-series	0.29	0.40	0.33				

Table 5.6: Comparison of precision (P), recall (R), and F1-scores (F1) between Casama and PubMed for study design classification.

	ALK PubMed			EGFR ASCO			ALK ASCO		
	P	R	F1	P	R	F1	P	R	F1
Experimental	0.60	1.0	0.75	1.0	0.63	0.77	0.75	0.60	0.67
Cross-sectional	0.79	0.85	0.82	0.73	0.80	0.76	0.80	0.80	0.80
Cohort (all)	0.81	0.75	0.78	0.81	0.86	0.83	0.63	0.58	0.60
Prospective Cohort	n/a	0	n/a	n/a	0	n/a	n/a	0	n/a
Retrospective Cohort	0.17	1.0	0.26	0.56	0.80	0.65	0.50	0.38	0.43
Unknown Cohort	0.75	0.25	0.375	0.64	0.33	0.44	0.20	0.25	0.22
Case-control	n/a	0	n/a	n/a	0	n/a	n/a	0	n/a
Case-series	0.33	0.20	0.25	0.50	0.50	0.50	n/a	0	n/a

Table 5.7: Casama's precision (P), recall (R), and F1-scores (F1) on test sets for study design classification.

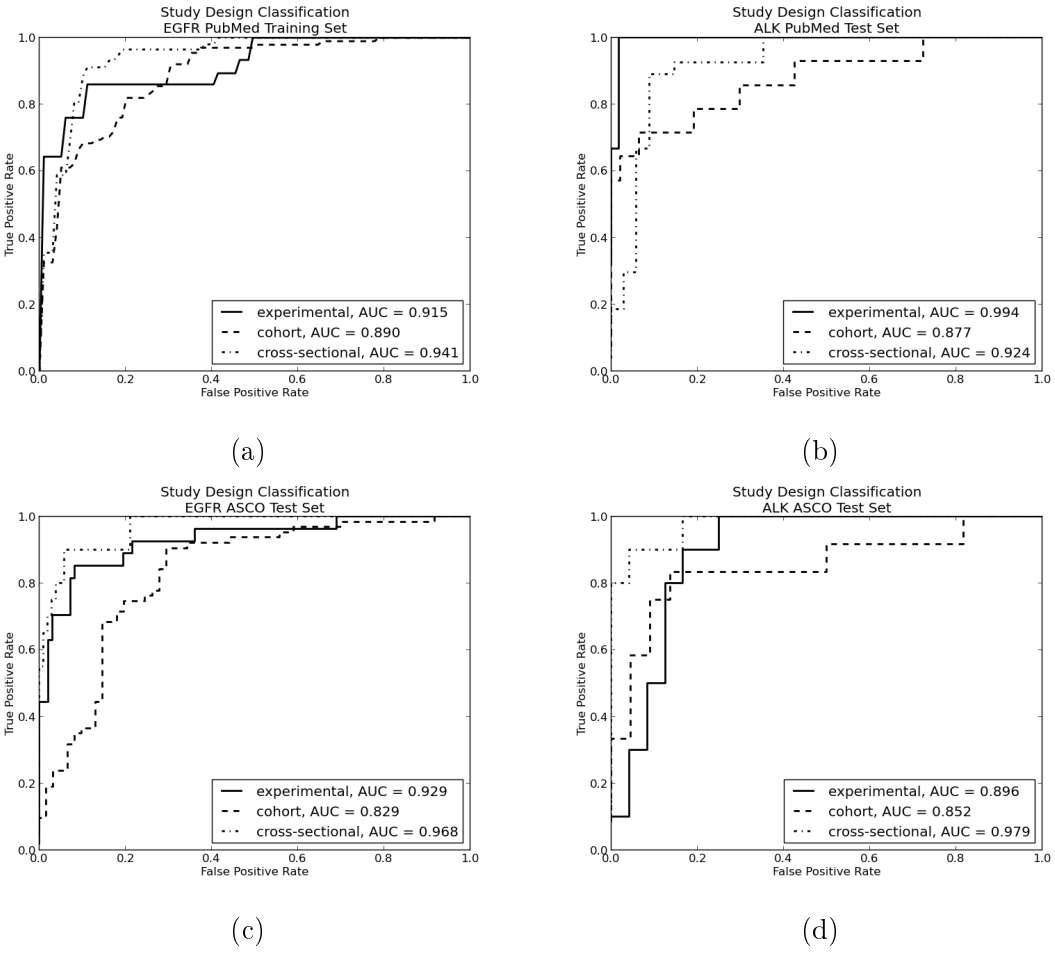


Figure 5.2: Casama's receiver operating characteristic and area under the curve for study design classification on (a) EGFR PubMed, (b) ALK PubMed, (c) EGFR ASCO, (d) ALK ASCO.

Mutation Characterization	Mutation Detection	Treatment	Prognosis
status	sample	progression	survival
kras	method	advanced	overall survival
higher	serum	mg	prognosis
correlated	detect	median epidermal	overall
conclusive	evaluate	control	analyze
patient	tumour	month	overall prognostic
smoker	dna	symptom	patient egfr
hospital	rearrangement	receive	month
egfr kras	copy	chemotherapy	differ
result	sensitivity	follow	significantly

Table 5.8: Top features for study objective classification.

Experimental	Cohort	Cross-sectional
patient epidermal	cancer patient	exon
toxicity	prognostic	detect
mg	retrospective	result
receive	worse	evaluate
clarify	observe	egfr kras
day	worse	examine
progression	month	prevalence
grade	prognosis	specimen
progression free	differ	pcr
six	significant difference	exon egfr

Table 5.9: Top features for study design classification.

mutation detection studies *evaluate sensitivity* of *detection methods* in *DNA samples*. Top features for treatment studies include explicit references to treatment (*chemotherapy*, *mg* (dosage)). Prognostic studies usually explicitly mention *prognosis* and examples of outcomes such as *overall survival*.

Discriminative features for the study design classifier indicate that experimental studies describe the details of the intervention (*mg*, *toxicity*). Top features for the other study design classes reveal that there is a relationship between study objective and study design – cohort studies tend to overlap with prognostic studies; detection or prevalence studies tend to be cross-sectional. In both cases, this relationship is unsurprising. Cohort studies by definition include follow-up and enable assessment of outcomes, as in a prognostic study. No follow-up is required to demonstrate a mutation detection technique, so these studies are often cross-sectional.

Category	Precision	Recall	F-score
Characterization	0.67	1.0	0.8
Detection	1.0	0.29	0.44
Treatment	1.0	0.87	0.93
Prognosis	0.53	1.0	0.70
Experimental	0.75	0.90	0.82
Cohort (all)	0.86	0.58	0.69
Prospective cohort	n/a	n/a	n/a
Retrospective cohort	0.63	0.63	0.63
Unknown	n/a	0	n/a
Case-control	n/a	0	n/a
Cross-sectional	1.0	0.80	0.89
Case series	n/a	0	n/a

Table 5.10: Classification performance on ALK ASCO when trained on EGFR ASCO.

5.4 Discussion

Casama's automatic classification performance was comparable to or better than PubMed's retrieval in every category. Notably, Casama automatically classified experimental studies with similar F1-score compared to PubMed's manual tagging of clinical trials.

For study objective classification, a decrease in performance was observed between the training set and the test sets (Tables 5.4 and 5.5). The ALK PubMed test set had the smallest decrease in performance; the decrease was greatest in the "treatment" category. A manual review of the incorrectly classified abstracts revealed that many errors could be attributed to differing stages of research between EGFR and ALK (e.g., ALK treatment studies were missed because they were descriptive rather than analytical).

In contrast, the ASCO test sets had a more dramatic drop in performance compared to the training set. In this case, a major source of error was the difference in vocabulary, writing style, and types of knowledge reported between PubMed and ASCO. As shown in Table 5.10, performance improved when the classifier was trained on the EGFR ASCO set and tested on the ALK ASCO set. This finding indicates that performance is indeed sensitive to vocabulary differences between PubMed and ASCO. Thus, creating a large training set of ASCO abstracts would be a useful direction for future work.

For study design classification, performance was preserved between training and test sets

(Tables 5.6 and 5.7). This is a very promising finding, as it suggests that the automatic extraction of study designs is a viable and generalizable strategy. However, rule-based performance was generally poor. Part of this stems from the effect of few examples of prospective cohort studies, case-control studies, and case series in the data set – small n results in a large penalty for missed abstracts. The other contributing factor is the fact that most studies do not explicitly name their study design in the abstract. Semantic modeling of study design, including identification of exposures, outcomes, and direction of inquiry for improved study design classification is a possible avenue for future work.

5.4.1 Top features

An examination of the top features reveals some interesting characteristics of the vocabulary used across studies. Many of these features would be expected (e.g., *chemotherapy* for treatment studies), and some are even included in PubMed's filters (e.g., *DNA* for mutation detection studies). The top features also reveal less obvious terms that can be used to discriminate between studies (e.g., *receive* for experimental studies vs. *observe* for cohort studies). However, simply entering a few top features into a PubMed search query is unlikely to produce good retrieval results as the vocabulary is modeled in a high-dimensional feature space via an SVM, going beyond the basic Boolean querying available in PubMed. Indeed, issuing the baseline query to PubMed with the top term for treatment studies (*progression*) results in an F1-score of 0.54. AND-ing the two most discriminative terms (*progression, advanced*) results in decreased recall; OR-ing them results in decreased precision.

Given the domain-specific nature of this representation, it is important to assess if the classifiers developed here can be applied outside the target domain (i.e., EGFR mutations in lung cancer). Markedly, many of the top features for the study objective classifier are not specific to EGFR mutation. As such, this classifier may be applicable to other driver mutations in NSCLC, especially those with similar treatment strategies. Furthermore, the top features of the study design classifier are not domain dependent and may generalize well to other disease and cancer domains.

5.5 Conclusion

In this chapter, the automatic classification of study objective and study design in abstracts on EGFR and ALK mutation in lung cancer was explored. Improved classification performance was achieved on the training and test sets compared to PubMed. Study objective classification was sensitive to differences in vocabulary between corpora; however, study design classification was robust to these differences. Based on an examination of top features, both classifiers could generalize outside the lung cancer domain.

CHAPTER 6

Automatic Extraction

6.1 Introduction

This chapter describes the automatic extraction of relations, population context, and study contexts (except study objective and study design, discussed in Chapter 5). A combination of lexical matching and OpenIE 4.0 was used to extract the most frequent concepts and relations in the Casama representation. Performance was evaluated using the gold standard described in Chapter 4, Section 4.4. Extraction performance varied by type of concept or relation; suggestions for improvement of the extraction system are discussed.

6.2 Methods

6.2.1 Selecting concepts and relations to extract

Given the large number of concepts and relations in the representation, a subset of concepts and relations was selected for automatic extraction to demonstrate proof of concept and establish a starting point for future development. The following relations were selected, corresponding to the most frequent relation(s) for each study objective type. This subset of relations accounts for 65% of the total relations found in the gold standard.

- *biomarker* **correlation** *clinical feature*
- *detection method* **detects** *biomarker*
- *biomarker* **predicts** *outcome*

- *biomarker* **predicts worse** *outcome*
- *biomarker* **does not predict** *outcome*
- *treatment* **treatment relation** *outcome*

Recall that relations associated with correlation, prediction, and treatment effects are organized hierarchically (see: Chapter 3, Figure 3.3). Based on the observed distribution of relations (see: Chapter 4, Table 4.6), 80% of relations are stated in the positive, except for **does not predict** and **predicts worse**, which make up a significant portion of the total relations. For this automatic extraction task, the remaining negated relations (e.g., **not correlated with**, **does not improve**) are subsumed by their parent relations. Thus, the extraction system tags any associations found, regardless of polarity.

A subset of clinical features / population context was selected based on an analysis of most frequently-occurring types in the manually-annotated gold standard. This subset comprises: *stage*, *histology*, *biomarker*, and *treatment history* (including *targeted therapy history*, *chemotherapy history*, and *surgery history*). *Sex, race* and *smoking history* were also targeted as these were readily extracted by lexicon or regular expression. These contexts represent 75% of the observed population context space. In this first attempt at extracting contextualized relations, extraction was limited to population contexts mentioned in the same sentence as a relation.

Similarly, the most frequent study contexts were selected for automatic extraction. The extracted subset, accounting for 85% of study contexts, comprises: *cohort size*, *p-value*, and *endpoint*.

6.2.2 Extraction algorithm

Automatic extraction of relations and population context was achieved by a four-step process. An overview of this process is given in Figure 6.1.

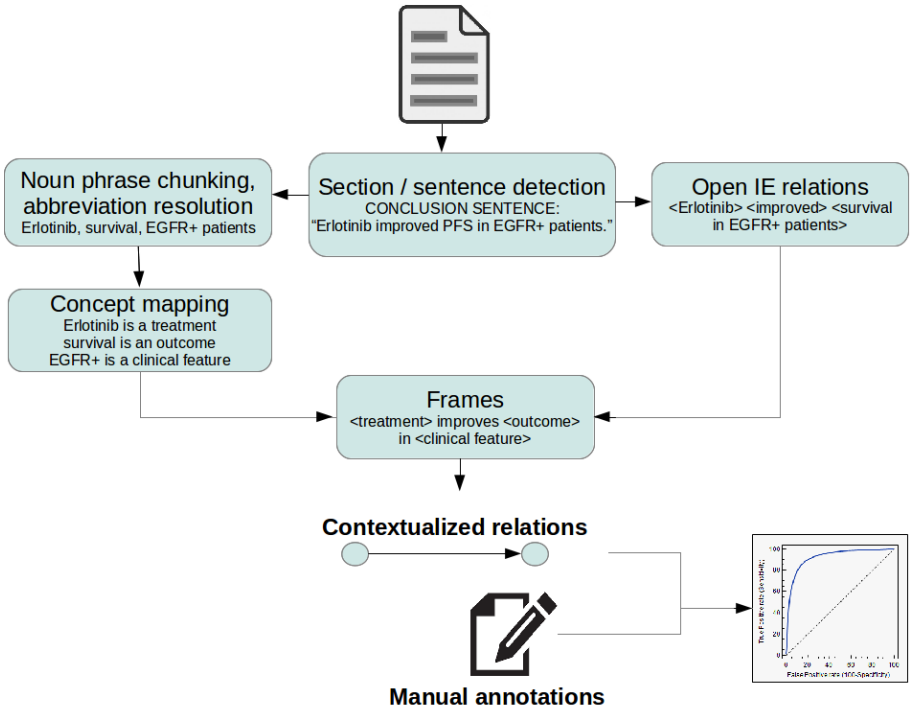


Figure 6.1: Overview of the automatic extraction method. First, abstracts are pre-processed to extract sentences, sections, and noun phrases. Lexicons are used to identify concepts from the noun phrase chunks. In parallel, OpenIE is used to extract relations from raw sentences in a domain-independent manner. Tagged concepts and relations are mapped to frames to produce the final relations. These relations are compared to the manual annotations to calculate precision and recall.

Semantic type	Source
Biomarker	Hensing et al.
Detection method	Ellison et al.
Endpoint/Outcome	National Cancer Institute
Histology	WHO/IASLC
History of chemotherapy	SNOMED-CT (anti-neoplastic agents)
History of targeted therapy	SNOMED-CT (protein-tyrosine kinase inhibitors)
History of surgery	LUCADA
Race	SNOMED-CT (ethnic group)
Treatment	RxNorm

Table 6.1: Sources used for the development of concept lexicons.

6.2.2.1 Pre-processing

Pre-processing was performed by the Python library NLTK (natural language toolkit). Abstracts were tokenized into sentences and noun phrase chunks. To leverage the structured nature of most abstracts, XML section headings were parsed (for the PubMed corpus) and regular expressions used (for the PubMed and ASCO corpora) to detect section boundaries (i.e., Background, Methods, Results, Conclusion). ASCO abstracts were further processed to expand common abbreviations (in the same manner described in Chapter 5, Section 5.2.1).

6.2.2.2 Concept matching using lexicons and regular expressions

The collection of pre-processed sentences was semantically tagged using manually-curated lexicons based on existing resources, such as SNOMED-CT [Don06], LUCADA [RTS11], RxNorm [NZK11], the WHO/IASLC Histologic Classification of NSCLC [TBN11], National Cancer Institute guidelines [Ins15], National Cancer Institute Common Terminology Criteria for Adverse Events (CTCAE) [TCS03], and papers on driver mutations in lung cancer [HS13, EZM13]. The resources on which these lexicons are based are given in Table 6.1. Each noun phrase in the sentence was searched for in the lexicon and tagged if a match was found. *Stage, sex, smoking history, cohort size* and *p-value* were extracted by regular expression.

6.2.2.3 Extraction of OpenIE relations

OpenIE 4.0 utilizes a self-supervised classifier, automatically labeling parsed syntactic patterns as trustworthy or untrustworthy based on a set of heuristic rules [BCS07]. A semantic role labeling step improves recall by including information about a noun phrase's relation to the verb in the sentence [CSE10]. The system is trained on a general English language corpus, designed to work at the Web scale on highly heterogeneous data. OpenIE was chosen as the relation extraction tool for Casama because its free availability and domain-independent nature were attractive features upon which to build a relation extraction system for Casama's unique representation.

Relations were extracted from the set of Results and Conclusion sentences. OpenIE relations are of the form <noun phrase> <verb phrase> <noun phrase>, where each argument in the relation may be a complex phrase such as, "<EGFR+ patients receiving erlotinib> <experienced> <clinically relevant improvements in quality of life>." The concatenation of these arguments form the "relation sentence."

6.2.2.4 Negation detection

An approach similar to that of NegEx [CBH01] is used for detection of **does not predict** and **predicts worse** relations. The EGFR PubMed data set was searched manually for linguistic triggers indicating no evidence of a relation or evidence of worse outcome, respectively. If one of these triggers was found in a relation sentence between a *biomarker* and an *outcome*, the relation was tagged accordingly. Otherwise, the relation was tagged as a **predicts** relation (implicitly, **predicts_better**). The full list of triggers is given in Table 6.2.

6.2.2.5 Frame matching

Having semantically tagged the noun phrases and identified relations between them, these tagged relations were then compared with the set of relations in the representation. Consider the relation, "<EGFR+ patients receiving erlotinib> <experienced> <clinically relevant

No relation	Worse outcome
not	low
similar	short
no relationship	poor
no significant difference	unfavorable
not an independent predictor	worse
no significant association	negative prognostic
not associated with	adverse prognostic
comparable	
no difference	
significant differences were not observed	
no correlation	

Table 6.2: For the "predicts" family of relations, linguistic triggers indicate no relation or worse outcome.

improvements in quality of life>" and its associated tags, EGFR+ (*biomarker / clinical feature*), erlotinib (*treatment*), and quality of life (*outcome*). An expected relation in treatment studies is <treatment> improves <outcome> in <clinical feature>. Thus, each element of the relation frame can be filled as "<erlotinib> improves <quality of life> in <EGFR+ patients>".

The system was designed to favor precision over recall – it is important for a summary to state facts that are true (high precision); furthermore, a low-recall system can still achieve broad coverage by using a large data set. To maximize precision, the subject of the relation (in this example, "erlotinib") must be identified by OpenIE as such; the object of the argument ("quality of life") must appear somewhere in the entire relation sentence; and the context ("EGFR+") may appear anywhere else in the sentence.

6.2.2.6 Training and test sets

The extraction algorithm was trained on the EGFR PubMed corpus, incrementally adjusting the lexicons and frame matching parameters to optimize performance. EGFR ASCO data formed the blind test set.

	training			test		
Relation	**Total in gold standard**	**Precision**	**Recall**	**Total in gold standard**	**Precision**	**Recall**
correlation	52	0.64	0.31	14	0.63	0.36
detects	24	0.71	0.21	20	0.38	0.15
treatment_relation	61	0.60	0.30	39	0.44	0.21
predicts	53	0.58	0.42	8	0.0	0.0
does not predict	20	0.67	0.20	11	0.75	0.55
predicts worse	33	0.81	0.52	6	1.0	0.17

Table 6.3: Precision and recall for each relation type extracted by Casama.

6.3 Results

6.3.1 Relation extraction

On the training set, performance ranged from 0.58-0.81 for precision and 0.20-0.52 for recall. On the test set (for relations having 10 or more examples), precision ranged from 0.38-0.75 and recall ranged from 0.15-0.55.

Negation detection favored precision over recall, achieving 0.67-0.81 precision on the training set and 0.75-1.0 precision on the test set. Unfortunately, no **predicts** relations were correctly identified on the test set after negation detection was performed.

6.3.2 Context extraction

Extraction of study design context was excellent, ranging from near perfect precision and recall for *p-value* to 0.61 precision and 0.62 recall for *endpoint*.

Extraction of study population context varied by semantic type, ranging from 0.79 precision and 0.52 recall for *histology*, to precision and recall of 0.27 for *biomarker*. Some semantic types had few or no instances in the training and test sets, making it difficult to estimate precision and recall.

	training			test		
Concept	**Total in gold standard**	**Precision**	**Recall**	**Total in gold standard**	**Precision**	**Recall**
p-value	78	1.0	1.0	23	0.96	0.96
cohort size	58	0.88	0.88	13	0.85	0.85
endpoint	63	0.61	0.62	59	0.85	0.80

Table 6.4: Precision and recall for the study contexts extracted by Casama.

	training			test		
Concept	**Total in gold standard**	**Precision**	**Recall**	**Total in gold standard**	**Precision**	**Recall**
biomarker	22	0.27	0.27	16	0.22	0.31
histology	21	0.79	0.52	5	0.0	0.0
stage	21	0.80	0.38	7	0.83	0.63
history of targeted therapy	14	0.50	0.29	9	0.50	0.22
history of chemotherapy	10	0.60	0.30	0	n/a	n/a
history of surgery	5	1.0	0.60	0	n/a	n/a
race	4	1.0	0.25	3	1.0	0.67
sex	1	n/a	0.0	0	0.0	n/a

Table 6.5: Precision and recall for the extraction of study population context by Casama.

6.4 Discussion

These initial extraction results are promising: they show that a relation extraction system based on OpenIE 4.0 can be devised that conforms to the representation developed; no other system exists currently. With the exception of *biomarker*, the system performed well in precision, as designed. Study contexts in particular were extracted with high precision and recall.

6.4.1 Error analysis

For the relation extraction task, there were two main sources of false negatives: concepts missing from the lexicons, and incorrect or incomplete parses by OpenIE. These sources of error contributed approximately equally to the total error and together accounted for the majority of false negatives.

Several concepts missing from lexicons contributed to false negatives. In particular, several biomarkers and outcomes were not found in any of the source lexicons and could

not be anticipated. While *biomarker* often refers to genetic alterations found in tumor tissue, other biomarkers studied included levels of substances found in serum (e.g., CYFRA 21-1, RANTES, adiponectin). Uncommon measured outcomes included "time from bone metastasis to first skeletal-related event" and "incidence of leptomeningeal metastasis."

OpenIE was another source of false negatives due to incorrect or incomplete parses. Although OpenIE's domain-independence was a key reason for using it, its lack of biomedical knowledge was ultimately a weakness. Abbreviations, mutation names, and drug names were not always included in the relation sentence, despite their semantic significance. For example, "Global health status/QoL was also improved over time with afatinib compared with chemotherapy" was parsed by OpenIE as <Global health status/QoL> <was also improved> <over time>. (The ideal parse would be: <Global health status/QoL> <was also improved over time> <with afatinib>.) Furthermore, complex sentences were likely to be parsed incorrectly. "Analyses revealed significant predictors for having EGFRMUT to be: female gender, non-smoking status, and adenocarcinoma subtypes" was parsed as <EGFRMUT> <to be> <female gender> (missing the other predictors of EGFR mutation).

OpenIE also produced false positives by including contextual elements such as comparisons in the relation sentence. The sentence, "Progression-free survival was significantly better with docetaxel than erlotinib" was parsed as <Progression-free survival> <was significantly better> <with docetaxel than erlotinib>. A more useful parse would be: <Progression-free survival> <was significantly better> <with docetaxel>, thus identifying the object of the relation without its context. The context could then be extracted by examining the full sentence.

Sentences reporting numeric data were another source of false positives (e.g., "Progression-free survival was 6.2 months on erlotinib vs 3.4 months on docetaxel"). Because the interpretation of numeric data is outside the scope of this study, these sentences were not annotated. However, the automatic extraction system recognized the co-occurences and tagged them as valid relations.

There was a decrease in performance between the training and test sets, particularly for

treatment studies, mutation detection studies, and prognosis studies after negation detection. As seen in Chapter 5, Section 5.4, vocabulary differences between the PubMed training set and ASCO test set affected concept mapping performance. An error analysis also showed that EGFR ASCO sentences were longer and denser, contributing to errors caused by the combined effect of bad OpenIE parses and co-occurences not expressing a relation. Deeper linguistic analysis (for example, using dependency trees rather than syntactic pattern matching) may result in more robust relation extraction of complex sentences such as those found in ASCO abstracts.

In the case of prognosis studies after negation detection, one salient observation is that negated relations have less than perfect performance, even in the training set from which the linguistic triggers were mined. Thus, the main error contribution is from the relation extraction algorithm itself, rather than effects intrinsic to negation detection. Indeed, an error analysis on the test set showed that most of these errors were due to missed concept mapping.

For context extraction, low precision scores were seen for *biomarker*. In this case, low recall of relations contributed to low precision of context (because if a mention is incorrectly missed as a biomarker argument in a relation, it may be incorrectly tagged as a biomarker context).

6.4.2 Comparison to existing resources

A proprietary extraction method was developed rather than leveraging existing tools such as MetaMap (for concept extraction) or SemRep (for relation extraction). Although MetaMap is useful for providing access to a broad, comprehensive vocabulary, more granular entities required by Casama were missing from MetaMap's lexicons ("overall survival," a frequently used endpoint that is distinct from "progression-free survival," was mapped to the more general term "survival"; specific alterations such as "EGFR mutation" and "exon 19 deletion" were similarly mapped to general terms "mutation" and "deletion" respectively). The representation on which SemRep is based is also limited, particularly in its omission of clinical

outcomes (see Chapter 8 for a more detailed comparison of Casama to SemRep).

6.4.3 Future work

This chapter presented a first attempt at automatic extraction of contextualized relations. Improvements could be made in three major areas: increased granularity of targeted relation types, more comprehensive concept recognition, and more accurate relation extraction via a biomedical OpenIE framework.

Casama's relation extraction system in its current state targets only the broadest relations in its representation. As a result of this simplification, relations can be discovered by only examining the argument types (disregarding the signifiers between the arguments, e.g., "survival *was longer* with erlotinib"). The next iteration of Casama could include this information as a feature in a supervised learning setting to differentiate between the various relation types.

As discussed in Section 6.4.1, Casama's usage of a highly specialized lexicon led to false negatives. A hybrid method that combines Casama's specialized representation with MetaMap's broader lexicon might prove beneficial. For example, a system could first search the MetaMap lexicon for broad concepts, then narrow down to a more specific concept from Casama's lexicon.

Finally, performance could be improved by training an OpenIE/semantic role labeling system on a biomedical rather than a general vocabulary; this could improve recognition of mutation names and other biomedical terms, promoting their inclusion in relations. A few biomedical OpenIE systems have been developed recently [MC12, NB14]; investigating these methods would be a fruitful avenue for future work.

6.5 Conclusion

This chapter explored the automatic extraction of study context, population context, and relations. Lexicon-based extraction of study context and population context varied by con-

cept type, performing particularly well for study contexts. A method for automatic relation extraction based on OpenIE 4.0 was also investigated, showing modest but promising results. Chapters 7 and 8 will show that the performance of these automated tagging methods are sufficient to enhance information retrieval and summarization.

CHAPTER 7

Patient-tailored information retrieval

7.1 Introduction

This chapter examines the application of Casama's structured representation for improving retrieval with respect to a given patient case. The design of this experiment is based on the Text REtrieval Conference (TREC) 2014 Clinical Decision Support shared task, which challenged participants to retrieve full-text articles from PubMed Central given a short narrative of a patient case [SVH14]. The following sections cover in detail the creation and automatic annotation of a blind document set, the composition of ad hoc and structured queries, the relevance judgment process, the comparison of results between PubMed and Casama, and a sensitivity analysis.

7.2 Methods

7.2.1 Annotating and indexing a blind document set

A previously unseen document set was automatically annotated for study and population concepts using the annotators described in Chapters 5 and 6. On December 15, 2015, PubMed was queried for recent articles pertaining to EGFR mutation in lung cancer. The query parameters were similar to those from the initial retrieval step described in Chapter 4. Reviews, case reports, and pre-clinical studies were excluded. The initial retrieval query was restricted to articles containing "EGFR" and "lung" in the title to ensure highly relevant results. For this experiment, this portion of the query was expanded to include titles and abstracts. The annotated gold standard consisted of articles from January 1, 2012 - September

Original query	egfr [Title/Abstract] AND lung [Title/Abstract] AND ("2013/09/01" [PDAT]:"2015/12/15" [PDAT])
Exclusion filter	NOT review [ptyp] AND hasabstract [text] NOT "cells" [title/abstract] NOT "cell lines" [title/abstract] NOT systematic [sb] NOT case reports [ptyp]

Table 7.1: PubMed queries for retrieving previously unseen abstracts on EGFR mutation in lung cancer.

1, 2013; the blind set included articles published between September 1, 2013 and December 15, 2015. 1,340 articles were returned. The full query is given in Table 7.1.

In Chapters 5 and 6, annotation of frequently-occurring concepts was formally evaluated. These concepts included *study objective, study design, cohort size, p-values, endpoints, stage, histology, biomarkers,* and *treatment history.* For this information retrieval task, additional regular expression extractors were developed for *study phase, smoking history, treatment line, progression,* and *resistance.*

A Lucene index was created containing the raw data associated with each article (title, date, authors, abstract) and Casama's structured fields. Each concept type in the Casama representation corresponded to a Lucene field. The fields were populated with the concatenation of terms annotated for each concept type. Documents were indexed using standard information retrieval methods: the vector space model with term frequency $\times$ inverse document frequency (TFIDF) weightings. An overview of the information retrieval process is given in Figure 7.1.

7.2.2 Creating queries from patient cases

A lung cancer oncologist composed five narratives describing clinically relevant patient cases, and a clinical question associated with each case (Table 7.2). This domain expert also provided queries to be run on PubMed to produce a baseline performance set. For Cases 3, 4, and 5, the initial queries provided by the expert were non-optimal, returning fewer than ten results. These queries were broadened by the expert when possible. For Case 5, the expert left the query as-is, acknowledging that the query was intentionally narrow.

In parallel, I composed structured queries for the cases. Queries were informed by the

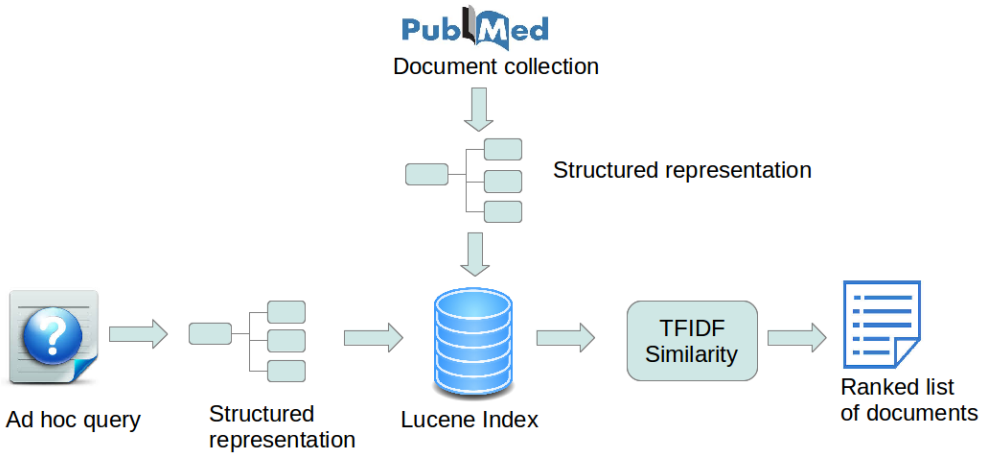


Figure 7.1: Casama's information retrieval approach is based on the standard vector space model. Casama adds a layer of structure to the documents and queries.

Casama representation and included as much information as possible — if there existed a concept and annotator for a term in the patient description, it was included in the query. If there was a relevant term in the patient description not covered by Casama (e.g., "germline mutation", "small cell transformation"), it was included in the query as free text. Study objectives, as determined by the clinical question, were also included in the query.

Construction of Casama queries also included manual and automatic query expansion. If the patient description included a biomarker (such as EGFR mutation), a negation of the opposite term ("NOT wild-type") was manually added to the query. If the patient description included a prior history of therapy, articles about "first-line" therapy were ruled out. Automatic query expansion was performed if names of drugs were detected in the query. A list of targeted therapies ("tyrosine kinase inhibitors") and chemotherapies ("anti-neoplastic agents") was extracted from SNOMED-CT. If a specific targeted therapy was named in the query, the term "tki" was automatically added. Similarly, if a chemotherapy was named, "chemotherapy" was added.

A simple query builder was developed to facilitate the construction of structured queries (Figure 7.2). The query builder includes a drop-down menu of all the searchable concept types in Casama, an entry widget for entering the search term, and checkboxes for selecting study objectives and negation if desired. As search terms are added to the query iteratively,

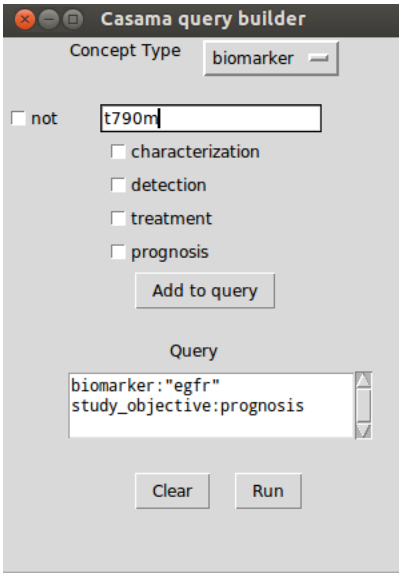


Figure 7.2: Screenshot of the Casama query builder.

a textbox displays the query in progress. The user can then run the query on the Lucene index, which produces a spreadsheet of results.

Patient descriptions, clinical questions, PubMed queries, and Casama queries are given in Table 7.2.

7.2.3 Evaluation

7.2.3.1 Relevance judgments

Three individuals with 6-8 months of experience in a lung cancer research setting (including using PubMed for literature reviews) were recruited to provide relevance judgments. Each judge was given 1-2 spreadsheets, each providing one patient description and the PMIDs, titles, and abstracts of at most twenty articles – the top ten results from PubMed and Casama. Ten was chosen as the cutoff due to its correlation with user satisfaction in web search tasks and limited availability of judges [MRS08]. The judges were blinded as to which set of results were produced by which system. The articles were presented as ranked by each system; the order in which each system was presented varied.

The judges were instructed to note in the spreadsheet whether each article was "definitely

Case	Case summary	Clinical question	PubMed query	Casama query
1	42 year old woman with newly diagnosed stage IV EGFR mutant disease and no prior therapy.	What is the best initial therapy for her?	EGFR mutation and stage IV and therapy	biomarker:egfr -biomarker:wild treatment_line:first study_objective:treatment stage:iv
2	63 year old woman with an EGFR mutation. Received erlotinib, followed by carboplatin and pemetrexed at progression followed by afatinib at progression. Now again progressing. Unknown T790M status.	Should she undergo a biopsy to evaluate whether she has a T790M mutation?	EGFR mutation and T790m and therapy	biomarker:"egfr t790m" -biomarker:wild targeted_therapy_history: "erlotinib afatinib tki" chemotherapy_ history:"carboplatin pemetrexed chemotherapy" study_objective: characterization -treatment_line:first progression:progressed
3	67 year old woman with an EGFR mutation. Received erlotinib. Now underwent repeat biopsy. T790M negative, but small cell transformation noted on repeat biopsy.	What is the optimal treatment approach for her?	EGFR and small cell transformation	small cell transformation biomarker:"egfr" -biomarker:"wild t790m" targeted_therapy_history: "erlotinib tki" study_objective:treatment -treatment_line:first
4	62 year old EGFR mutant man status post frontline carboplatin, paclitaxel and bevacizumab with maintenance bevacizumab and erlotinib, not progressing and rebiopsied. Noted to have a T790M mutation.	Would this patient benefit from a change in therapy from his current erlotinib and bevacizumab?	T790M and therapy	biomarker:"egfr t790m" -biomarker:wild chemotherapy_ history:"carboplatin paclitaxel bevacizumab" targeted_therapy_history: "erlotinib egfr tki" study_objective:treatment -treatment_line:first -progression:progressed
5	27 year old woman with newly diagnosed EGFR mutant NSCLC with a T790M mutation and L858R mutation in the EGFR gene.	Should she be tested for a germline T790M mutation?	EGFR and T790M and germline	germline biomarker:"egfr t790m l858r" -biomarker:wild treatment_line:first study_objective: characterization

Table 7.2: Patient cases and their corresponding PubMed and Casama queries. For Casama queries, "-" indicates negation. All PubMed queries were limited to papers published between September 1, 2013 and December 15, 2015.

relevant," "potentially relevant," or "non-relevant." Judges were given guidelines for determining the relevance of an article. A definitely relevant article is applicable to the patient and addresses the clinical question. If the judge were searching for information regarding this type of patient, he or she would definitely click through to learn more. A potentially relevant article may apply to the patient but not the clinical question, may apply to the patient for some features but not others, or may simply provide an overview of a topic related to the clinical question. A non-relevant article would be ignored during an information seeking activity.

7.2.3.2 Evaluation metrics

Evaluation metrics were normalized discounted cumulative gain ($NDCG$), *precision*, and binary preference ($bpref$), calculated for the top ten results for each system [MRS08].

Discounted cumulative gain (DCG) is a measure of the ranking of the documents retrieved. Definitely relevant documents, potentially relevant documents, and not relevant documents have rank scores (rel) of 2, 1, and 0 respectively. DCG, given by the formula below, penalizes relevant documents that appear lower in the result list by a factor logarithmically proportional to the position of the document.

$$DCG_k = rel_1 + \sum_{i=2}^{k} \frac{rel_i}{\log_2(i)}$$

k in this evaluation is 10, the number of results judged for each system; rel_i is the relevance of the i-th document retrieved.

Normalizing DCG ensures that value of this metric is between 0.0 and 1.0. $ideal\_DCG$ is the result of perfect ranking, i.e., sorted in descending order of relevance. $NDCG$ is obtained by dividing DCG by the $ideal\_DCG$.

$$NDCG = \frac{DCG}{ideal\_DCG}$$

$precision_k$ is the proportion of relevant documents retrieved within the top k=10 results.

$$precision_k = \frac{|relevant\_documents\_retrieved|}{k}$$

The final metric selected for this task, binary preference or *bpref*, is a measure of the ranking effectiveness of the system with respect to the known relevant documents pooled from both systems. This metric was designed to be robust to incomplete judgment sets [BV04]. *bpref* is defined as:

$$bpref = \frac{1}{R} \sum_r 1 - \frac{n\_ranked\_higher\_than\_r}{R}$$

R is the total number of known relevant documents, and r is a single relevant document. n is a member of the top R judged non-relevant documents; $|n\_ranked\_higher\_than\_r|$ is the number of documents in this set ranked higher in the result list than r. Thus, *bpref* is calculated by counting the number of non-relevant, higher ranked documents for each relevant document, and scaling by the total number of known relevant documents. For cases in which very few relevant documents were returned (i.e., all cases under strict evaluation, and case 5 under both strict and relaxed evaluation) an alternative to *bpref*, referred to as $bref\_10$, broadens $n\_ranked\_higher\_than\_r$ to span the top R+10 judged non-relevant results:

$$bpref\_10 = \frac{1}{R} \sum_r 1 - \frac{n\_ranked\_higher\_than\_r}{R+10}$$

For each metric, strict and relaxed versions were calculated. Strict evaluation only considers definitely relevant documents as true positives, whereas relaxed evaluation combines definitely relevant and potentially relevant documents into a single relevance class.

7.2.3.3 Sensitivity analysis

A sensitivity analysis was carried out to investigate Casama's retrieval behavior with variations on query structure and length. The aim of this experiment was to identify the causes of non-optimal behavior and suggest strategies for improving retrieval results. An overview of the sensitivity analysis pipeline is given in Figure 7.3.

Two types of query variations were examined. Structural variations assessed the contribution of structured fields to the retrieval process. For each term in the original query, the

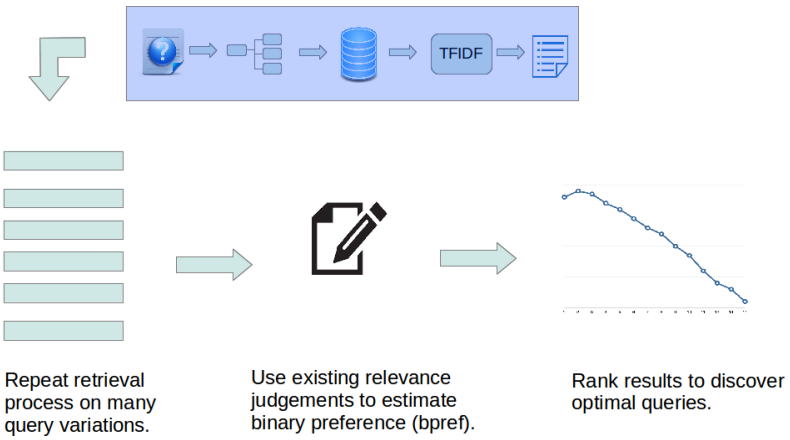


Figure 7.3: Overview of the sensitivity analysis pipeline.

Original query	Structural variations	Term varations
biomarker:egfr biomarker:t790m	biomarker:egfr biomarker:t790m biomarker:egfr abstract:t790m abstract:egfr biomarker:t790m abstract:egfr abstract:t790m	biomarker:egfr biomarker:t790m biomarker:egfr biomarker:t790m

Table 7.3: Structural variations and term variations for a simple two-term query.

structural variation query included either the term within a structured field or as a free-text search within the abstract. Every combination of structured or free-text query terms was generated automatically.

Term variations assessed the impact of query length and selection of query terms. Each term in the query was either included or not included in the term variation query. An example of structural variations and term variations for a simple two-term query is given in Table 7.3.

No new relevance judgments were carried out for this task. Thus, many queries returned results that were not in the original judgment set. *bpref-10* was chosen as the evaluation metric, as it was designed to remain stable despite incomplete judgment sets.

7.3 Results

7.3.1 Relevance judgments

Results for each case are presented in Tables 7.4-7.8.

Case 1: Casama outperformed PubMed on all metrics.

	Strict		Relaxed	
Case 1	PubMed	Casama	PubMed	Casama
Precision	0.2	0.3	0.5	0.9
Normalized discounted cumulative gain	0.57	0.87	0.68	0.99
Binary preference	0.30^{+}	0.63^{+}	0.28	0.68

Table 7.4: Case 1: Retrieval results comparing PubMed and Casama.
$^{+}$bpref-10 variation was used.

Case 2: PubMed outperformed Casama in *NDCG* in both strict and relaxed evaluation. For the other metrics, Casama outperformed PubMed in strict evaluation; the result was reversed when relaxed metrics were used. An investigation of the relevance judgments showed that the PubMed set consisted of eight consecutive potentially relevant documents, resulting in high relaxed *precision* and *NDCG*, but a strict *precision* of zero.

	Strict		Relaxed	
Case 2	PubMed	Casama	PubMed	Casama
Precision	0.0	0.2	0.8	0.6
Normalized discounted cumulative gain	0.95	0.43	0.95	0.74
Binary preference	0.0^{+}	0.67^{+}	0.57	0.38

Table 7.5: Case 2: Retrieval results comparing PubMed and Casama.
$^{+}$bpref-10 variation was used.

Case 3: PubMed outperformed Casama in strict evaluation; performance was similar in relaxed evaluation.

	Strict		Relaxed	
Case 3	PubMed	Casama	PubMed	Casama
Precision	0.5	0.3	0.8	0.8
Normalized discounted cumulative gain	0.88	0.73	0.89	0.85
Binary preference	0.56^{+}	0.35^{+}	0.45	0.46

Table 7.6: Case 3: Retrieval results comparing PubMed and Casama.
$^{+}$bpref-10 variation was used.

Case 4: Casama outperformed PubMed on all metrics, except strict *NDCG*, for which their performance was similar.

	Strict		Relaxed	
Case 4	PubMed	Casama	PubMed	Casama
Precision	0.3	0.6	0.5	0.9
Normalized discounted cumulative gain	0.73	0.75	0.73	1.0
Binary preference	0.36^{+}	0.75^{+}	0.31	0.74

Table 7.7: Case 4: Retrieval results comparing PubMed and Casama.
$^{+}$bpref-10 variation was used.

Case 5: PubMed retrieved only one document, which was judged definitely relevant. Thus, *precision* and *NDCG* at k=10 could not be determined. Casama also retrieved this document as the top result; additionally, the judge identified another definitely relevant document and five potentially relevant documents that were retrieved by Casama but not PubMed.

	Strict		Relaxed	
Case 5	PubMed	Casama	PubMed	Casama
Precision	n/a*	0.2	n/a*	0.7
Normalized discounted cumulative gain	n/a*	0.68	n/a*	0.80
Binary preference	0.50$^{+,*}$	0.79^{+}	0.14$^{+,*}$	0.88^{+}

Table 7.8: Case 5: Retrieval results comparing PubMed and Casama.
$^{+}$bpref-10 variation was used.
*Only one document was retrieved.

7.3.2 Sensitivity analysis

Figures 7.4-7.8 illustrate *bpref-10* for the structural and term variations, sorted in descending order for each case. Performance for PubMed and the original Casama query are also noted.

Case 1: When all terms in the query are structured terms (Figures 7.4c and 7.4d), no gains in improvement can be achieved by streamlining the query. However, Figure 7.4a shows that there exists a better query than the original Casama query. This optimal query was: biomarker:EGFR, biomarker:T790m, treatment_line:first, and free-text search for all other terms.

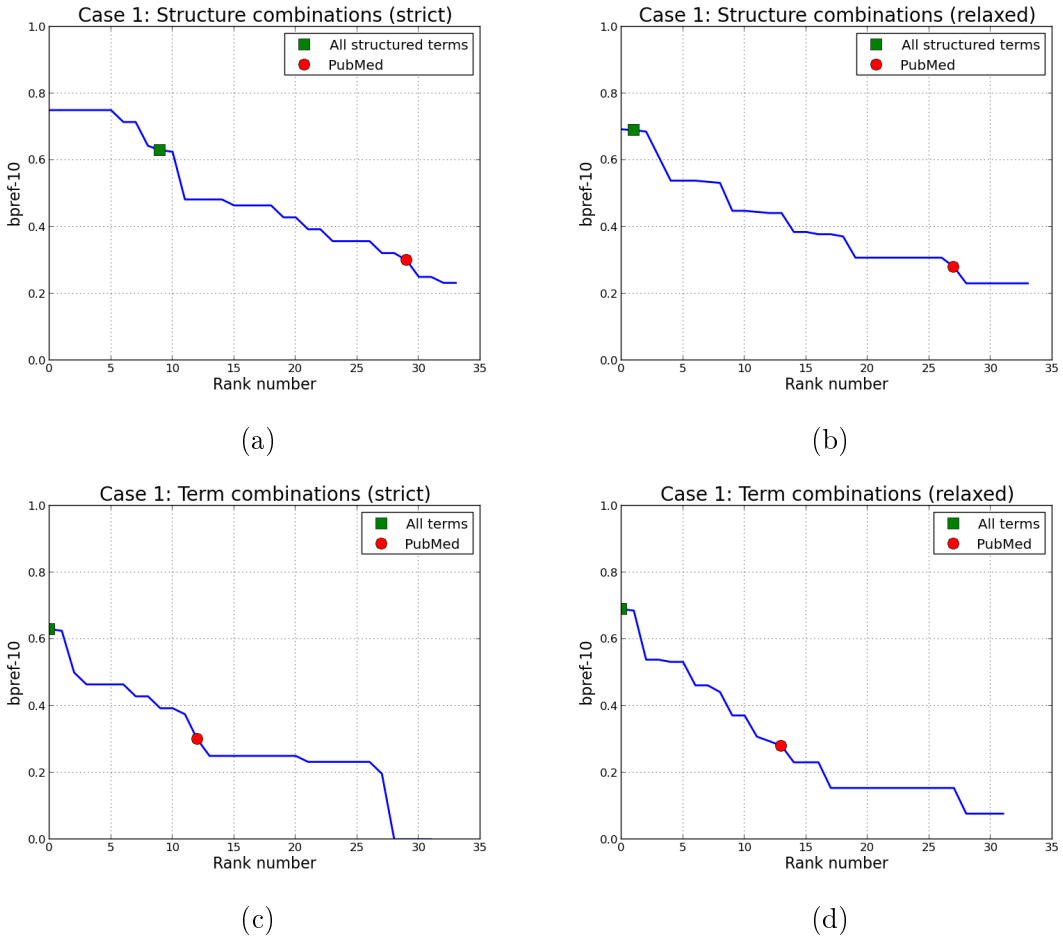


Figure 7.4: Case 1: The original Casama query was optimal compared to other term combinations (Figures 7.4c and 7.4d) and in relaxed evaluation (Figures 7.4b and 7.4d).

Case 2: Casama outperformed PubMed in strict evaluation, yet performance was further improved by using only the structured terms biomarker:EGFR and biomarker:T790m (Figures 7.5a and 7.5c). PubMed outperformed Casama in relaxed evaluation, despite optimal performance by Casama (Figures 7.5b and 7.5d).

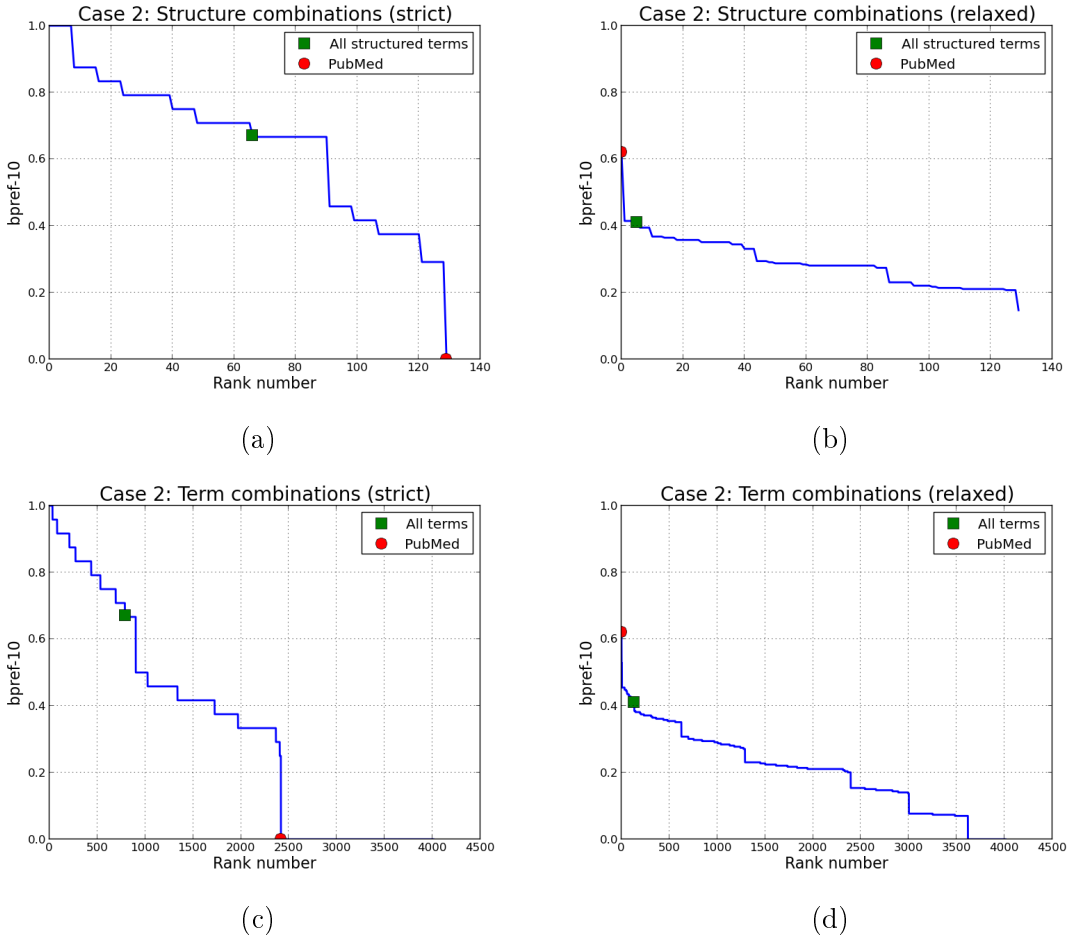


Figure 7.5: Case 2: Casama outperformed PubMed in strict evaluation, yet its performance could be further improved. The Casama query was optimal in relaxed evaluation, yet performance never reached that of PubMed.

Case 3: The original Casama query was near-optimal among both the structure variations and the term variations. Despite this, Casama's performance never reached that of PubMed in strict evaluation. Their performance was equal in relaxed evaluation.

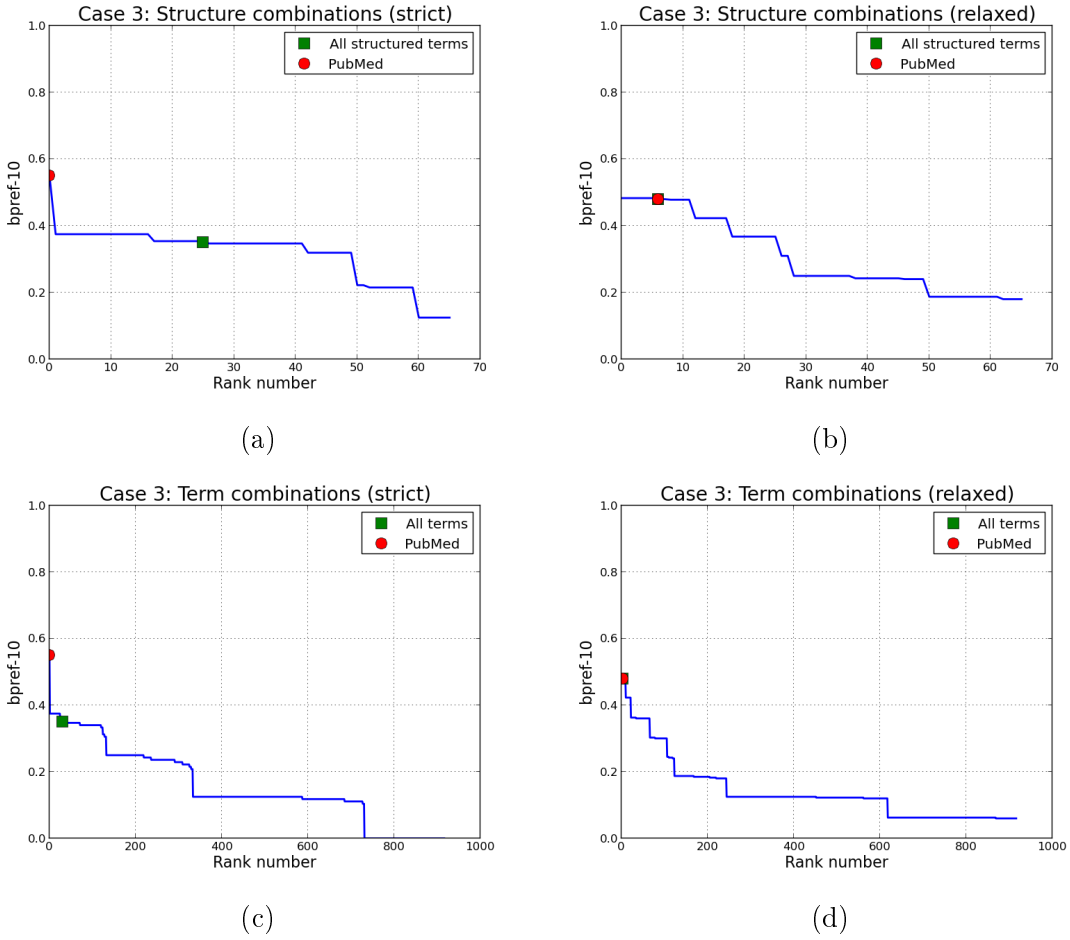


Figure 7.6: Case 3: The original Casama query was near optimal for both types of variations. Nonetheless, Casama never reached PubMed's level of performance in strict evaluation.

Case 4: The original Casama query was optimal over all query variations.

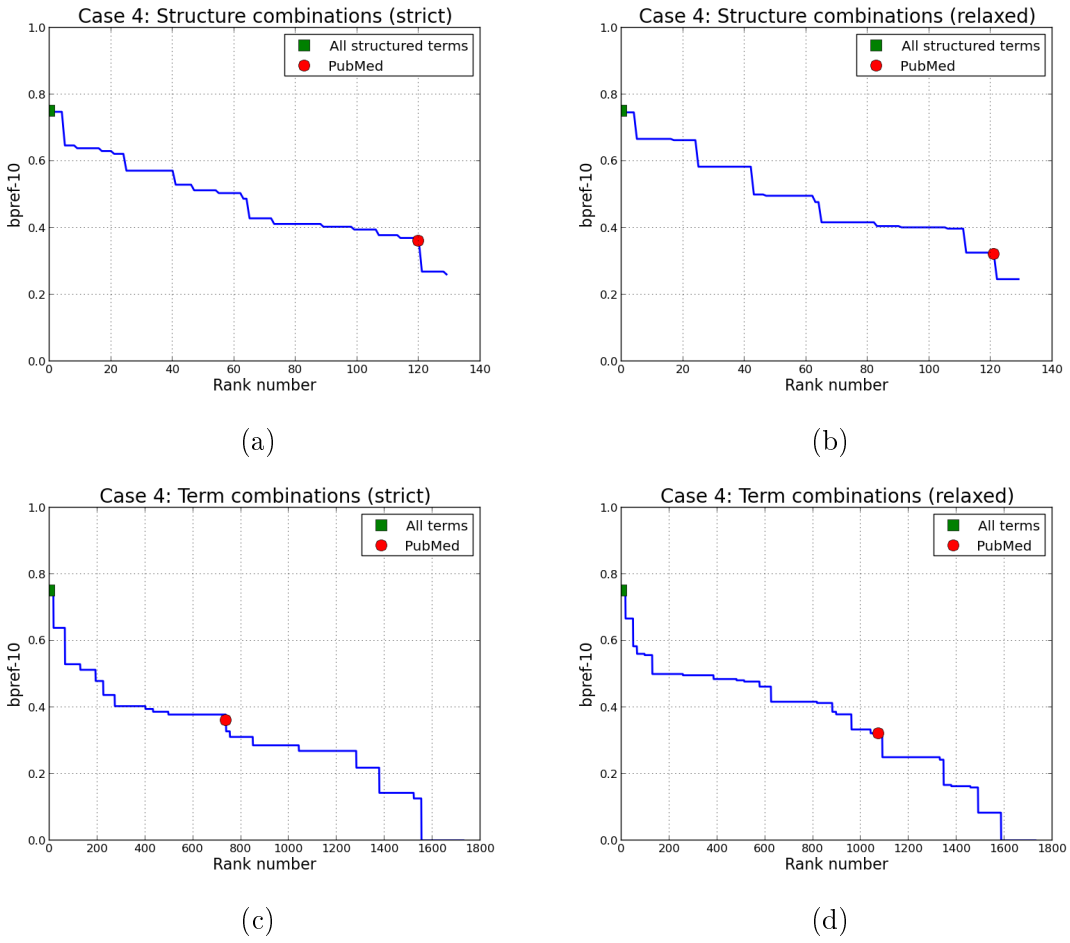


Figure 7.7: Case 4: The original Casama query was optimal over all variations.

Case 5: Casama performed optimally in relaxed evaluation; however, performance could be improved by broadening the search query.

The top structural variation was study_objective:characterization, with free-text search for the remaining terms.

The top term variation included the terms biomarker:egfr, biomarker:T790m, NOT biomarker:wild, and study_objective:characterization.

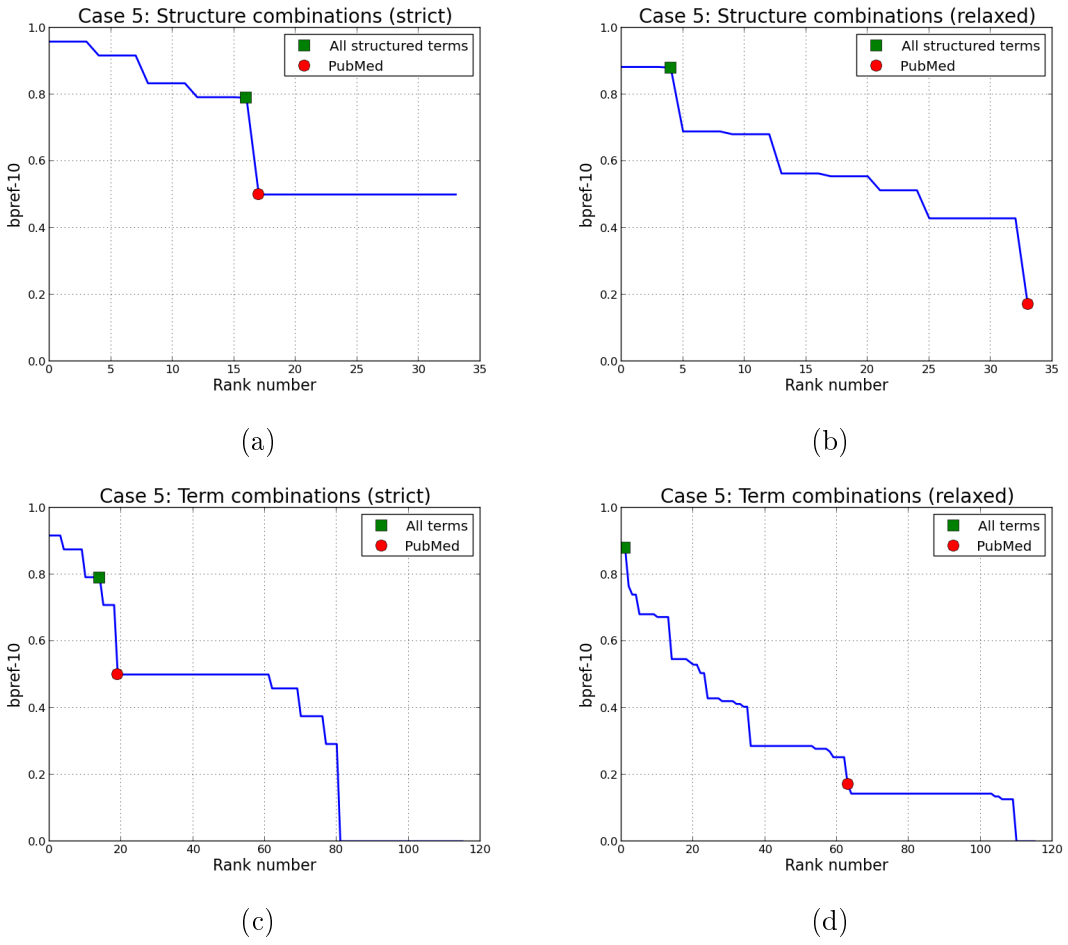


Figure 7.8: Case 5: Casama performed optimally in relaxed evaluation. However, in strict evaluation performance benefited from broadening of the search space.

7.4 Discussion

Casama performed optimally in many cases, such as in Cases 1 and 4 in which Casama outperformed PubMed and no query variation resulted in improved performance. Crucially, every term in the query was covered by Casama's structured representation. In this situation, a robust representation of patient characteristics proved beneficial to retrieval in terms of number of relevant documents retrieved and their ranking.

In contrast, in Case 3, the key concept "small cell transformation" was not included as a structured query term because it is not covered by the Casama representation. (The representation has a concept for the presence of resistance, but not the mechanism of resistance. The *biomarker* concept focuses primarily on genetic rather than histological alterations.) "Small cell transformation" was included in the query as free text, reducing Casama's search to a more standard approach. Thus, PubMed outperformed Casama in this case, most pronouncedly in strict evaluation, despite near optimal performance by Casama with respect to the query variations. Because Casama's approach is sensitive to terms missing from its representation, one key finding from this analysis is the importance of keeping Casama's representation and lexicons up to date as new discoveries are made.

Cases 2 and 5 were challenging for both systems: very few definitely relevant documents were discovered. For clinical questions that are known to be narrow in scope, one strategy for improving performance is through broadening of the search query. In Cases 2 and 5, broad queries (biomarker:egfr biomarker:T790m and study_objective:characterization respectively) gave the best results. (Interestingly, the queries composed by the lung cancer expert are also very broad, despite being based on a description consisting of multiple patient attributes. The lung cancer expert possesses a mental model of the patient and knowledge domain; these are distilled to a query containing very few terms.) One potential improvement to Casama would be a method for intelligently broadening search terms when unsatisfactory results are returned. A knowledge-based solution could leverage the hierarchical structure of the Casama representation to replace specific terms in the query with more general ones. This could also be performed interactively by presenting the user with a set of candidate

terms to on which to generalize.

In Chapter 6, it was shown that Casama's concept annotation algorithm achieved only moderate results (0.28-0.91 in precision, 0.23-0.48 for recall). However, even simple, lexicon-based matching was sufficient to improve retrieval significantly. These results are consistent with the findings of Boudin, who developed a retrieval system based on the PICO representation (which includes a less granular patient representation compared to Casama). Boudin demonstrated significant improvements in retrieval even with relatively low accuracy in the detection of PICO elements [BSN10].

7.4.1 Limitations and Future Work

An important limitation of this study was the relatively few number of relevance judgments performed. Each metric was assessed at k=10. Evaluating at the higher k would give more robust results; however, limited resources were available for performing relevance judgments.

It is also important to note that inter-rater agreement was not investigated. Rather, the results for each patient case were evaluated by a single judge. Ultimately, the results of each system were dependent on each judge's opinion. Nevertheless, improvements by Casama were seen across judges. Case 1 and Case 4, which showed substantial improvement by Casama over PubMed, were judged by different individuals. In the TREC Clinical Decision Support shared task, inter-rater agreement was lower than expected, indicating that standardized judgment criteria remain an area for future investigation [SVH14].

The results of the sensitivity analysis showed that out-of-date lexicons led to worse retrieval performance. Future versions of Casama should include a mechanism for keeping vocabularies up-to-date as new literature is published. This should be done automatically to ensure scalability. Existing resources that are frequently updated (e.g., RxNorm for drug names [NZK11], UpToDate for clinical reviews [Pro16]) could be leveraged for this task.

This study focused on an information retrieval task that was temporally static (i.e., retrieval was evaluated at a single point in time). However, real-world information retrieval systems experience many user interactions per day. Relevance feedback, the process of

continually training the system based on user interactions, was a common feature of the top-performing systems in the 2014 TREC Clinical Decision Support track [RSD14]. Incorporating these methods would enable Casama to improve over time, even as new knowledge is discovered.

7.5 Conclusion

This chapter evaluated the use of the Casama representation on an information retrieval task. Results on an automatically annotated, previously unseen document set showed that a structured retrieval approach based on the Casama representation outperformed PubMed in many cases and across several metrics. Areas of improvement include frequent updates to the concept lexicons to ensure maximum coverage, intelligent broadening of the search space in the case of unsatisfactory results, and improved evaluation with additional relevance judgments and assessment of inter-rater agreement. Next, Chapter 8 will present a second task-oriented evaluation of Casama: summarization.

CHAPTER 8

Summarization

8.1 Introduction

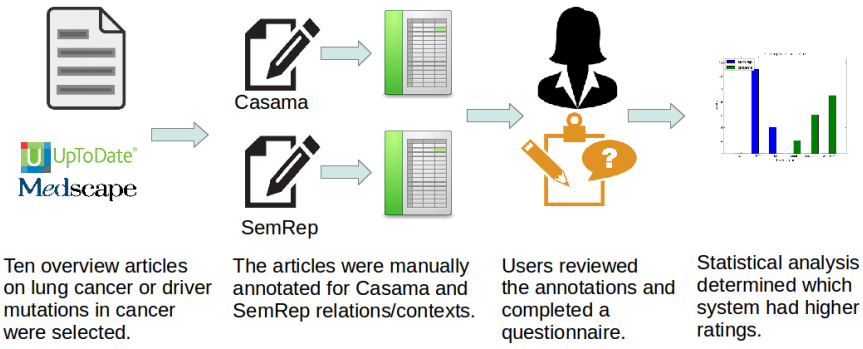
This chapter describes an evaluation study that compared the summarization capabilities of Casama with a baseline system SemRep, a representation based on the Unified Medical Language System [RFL05]. Manual and automatic annotations of several articles on driver mutations in cancer were reviewed and rated by multiple users. The results of the final analysis demonstrated significant advantages of Casama's contextualized relations over SemRep, particularly in the representation of strength of evidence.

8.2 Methods

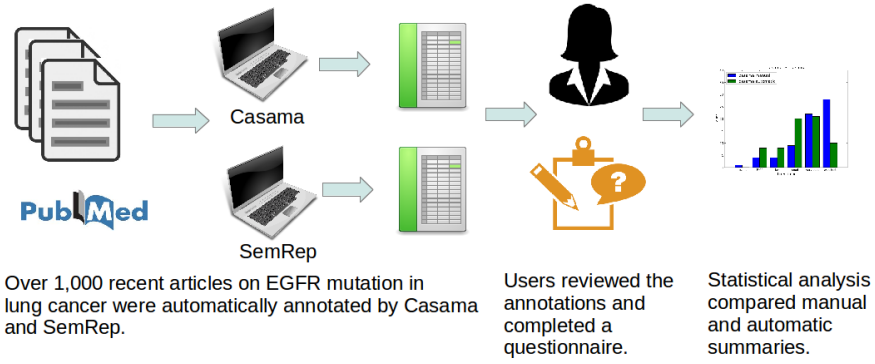
The design for this user evaluation consists of the following steps. First, articles on a variety of topics were selected to form a gold standard. These articles were annotated both manually and automatically according to the Casama and SemRep representations. Then, users were recruited to review the articles and associated summaries. A questionnaire was composed that enables users to rate the summarization quality of Casama and SemRep on a number of topics.

A statistical analysis was performed to discover whether one system rated significantly higher than the other. This analysis was performed both on individual articles and in aggregate. A separate analysis examined how summaries produced by manual annotation compared to those generated automatically.

A diagrammatic overview of this process is given in Figure 8.1.



(a)



(b)

Figure 8.1: Overview of the evaluation pipeline for (a) manually-annotated relations and (b) automatically-extracted relations.

8.2.1 Article collection and annotation

To test Casama's generalizability outside of its target domain, a variety of articles spanning multiple topics were selected for this study. UpToDate and Medscape, two sources of human-curated summaries on a variety of clinical topics, were searched for articles on "driver mutations in cancer" and "targeted therapies in cancer." The top ten relevant articles were selected to form the summarization gold standard in this evaluation.

The articles were numbered and organized into the following families:

Driver mutations in lung cancer (UpToDate):

1. Anaplastic lymphoma kinase (ALK) fusion oncogene positive non-small cell lung cancer

2. Systemic therapy for advanced non-small cell lung cancer with an activating mutation in the epidermal growth factor receptor

Driver mutations in cancers other than lung (UpToDate):

3. Systemic treatment for HER2-positive metastatic breast cancer

4. Anti-angiogenic and molecularly targeted therapy for advanced or metastatic clear-cell renal cell carcinoma

5. Molecularly targeted therapy for metastatic melanoma

Lung cancer, other topics (UpToDate):

6. Systemic therapy for the initial management of advanced non-small cell lung cancer without a driver mutation

7. Advanced non-small cell lung cancer: Subsequent therapies for previously treated patients

8. Personalized, genotype-directed therapy for advanced non-small cell lung cancer

Driver mutations in cancer (Medscape):

9. Genetics of Non-Small Cell Lung Cancer

10. Breast Cancer and HER2

To discover how well the Casama and SemRep representations were able to capture the knowledge expressed in each article, I used the brat rapid annotation tool [SPT12] to manually annotate each of these articles twice: first, for Casama relations and contexts; second, for SemRep relations. To mitigate the bias inherent in performing the annotations myself, I adhered to a set of annotation guidelines in both cases. The Casama guidelines for annotating relations and context can be found in Appendix B. Three sources were used for annotating SemRep relations: SemRep annotation guidelines detailed in [KRF11], the existing SemRep gold standard [Bio13], and the output of the SemRep relation extraction program.

The annotations were subsequently exported to spreadsheets containing each relation, the semantic types of its subject and object, the sentence in which the relation was found, and for Casama, the contexts in which the relation was found. On average, the articles contained instances of 244 Casama relations and 338 SemRep relations.

To evaluate the summarization quality of automatically extracted relations and contexts, a document set consisting of recent articles on EGFR mutation in lung cancer was automatically annotated by both Casama and SemRep. The document set used was identical to the collection described in Chapter 7, Section 7.2.1: 1,340 articles containing "EGFR" and "lung" in the title/abstract, published between September 1, 2013 and December 15, 2015.

The automatic annotations were exported to the same spreadsheet format described above. SemRep extracted many more relations compared to Casama (>7,000 relations for SemRep vs. 318 relations for Casama). This was due to 1) a greater number of relation types targeted by SemRep (29 for SemRep vs. 6 for Casama) and 2) significant repetition of general relations (e.g., 600+ instances of "Non-small cell carcinoma **process of** Human"). Only unique relations were exported to the spreadsheet to ease the examination process by users. The number of instances of each relation was also exported to the spreadsheet.

8.2.2 User assignments

Seven users were recruited: three graduate students with 2-3 years of experience in medical informatics, and four researchers with 1-3 years of experience in a lung cancer clinic. Article assignments were allocated such that each article and its associated spreadsheets were viewed by 3-4 users.

To minimize variability among the users, three of the researchers evaluated the same set of articles (Group A); the three graduate students evaluated the remaining articles (Group B). The groups of articles were selected such that each group contained 1-2 articles from each family described in Section 8.2.1.

The seventh evaluator, a researcher, was assigned a set of higher difficulty articles and the spreadsheets of automatically extracted relations. "Higher difficulty" articles included those in domains for which Casama was not tailored (i.e., "driver mutations in non-lung cancers" and "other topics in lung cancer."). As a result, the number of users viewing these articles was maximized to improve the probability of seeing a significant difference between SemRep and Casama.

The final user assignments were as follows:

Group A: Articles 1, 3, 4, 6, 9

Group B: Articles 2, 5, 7, 8, 10, automatic annotations

Seventh evaluator: Articles 2, 3, 4, 7, automatic annotations

8.2.3 Questionnaire

I composed a user questionnaire to measure the quality of the SemRep and Casama relations from various perspectives. As in the annotation step, bias is introduced by designing the questionnaire myself, rather than through an individual not involved with Casama. Thus, I aimed to include a wide variety of topics that are covered by SemRep and/or Casama to discover the overlap between systems and the relative advantages of each system. During the annotation step, I noted the set of topics covered by either SemRep or Casama. The

users rated the quality of the SemRep relations and Casama contextualized relations with respect to these topics on a 5-point Likert scale (5=excellent, 4=very good, 3=good, 2=fair, 1=poor). The topics were:

Identification of drugs/treatments

Effectiveness of drugs/treatments

Clinical guidelines for drugs/treatments

Side effects of drugs/treatments

Identification of genes/biomarkers

Prognostic effects of genes/biomarkers

Clinical characteristics of genes/biomarkers

Biochemical characteristics of genes/biomarkers

Diagnostic tests/detection methods

Strength of evidence

Additionally, the users rated the overall summarization quality, comprehensibility, and usefulness of SemRep and Casama for several high-level applications: clinical decision support, precision medicine, evidence based medicine, meta-analysis, and general biomedical research.

In a free-text portion of the questionnaire, users were asked to state what relevant information was missing from the SemRep and Casama summaries. Additional free-text comments on any topic relating to SemRep and Casama were also encouraged.

8.2.4 Analysis

The Wilcoxon rank sums test, a non-parametric test commonly used to analyze ordinal data, determined whether one representation tended to have higher scores than another. A significance threshold of 0.05 was selected; however, due to the large number of hypothesis tests (one for every topic and article), Bonferroni correction was applied, resulting in various p-value thresholds for each group of tests [McD09].

	ALK (lung)	HER2 (breast)	VEGF (renal)	EGFR wild-type (lung)	Medscape (lung)
Identification of drugs	0.66	0.31	0.56	0.51	0.51
Effectiveness of drugs	0.05	0.021	0.030	0.05	0.13
Clinical guidelines	0.08	0.11	0.030	0.05	0.05
Side effects	0.51	0.19	0.47	1.0	1.0
Identification of genes	0.28	0.15	0.77	0.28	0.38
Prognostic effects of genes	0.08	0.15	0.15	0.38	0.05
Clinical characteristics of genes	0.13	0.39	0.39	1.0	0.05
Biochemical characteristics of genes	0.13	0.77	1.0	0.38	0.66
Diagnostic tests	0.13	0.56	1.0	1.0	0.13
Strength of evidence	0.05	0.021	0.021	0.05	0.05
Overall summarization quality	0.05	0.021	0.030	0.38	0.05
Comprehensibility	0.19	0.15	0.39	0.05	0.19
Clinical decision support	0.05	0.03	0.083	0.05	0.05
Precision medicine	0.13	0.061	0.061	0.13	0.05
Evidence based medicine	0.05	0.11	0.043	0.081	0.081
Meta-analysis	0.080	0.083	0.083	0.081	0.081
General biomedical research	0.13	0.25	0.15	0.081	0.081

Table 8.1: Group A: P-values for Wilcoxon rank sums, testing the hypothesis that Casama's scores tend to be higher than SemRep's. Significance threshold was 0.0006 due to Bonferroni correction.

First, scores were compared per topic per article (p-value threshold = 0.05 / (17 questions × 5 articles) = 0.0006). As each article was reviewed by only 3-4 users, the sample size for each observation was quite small. With this number of observations, significant results would only be expected if very large effect sizes were observed. To achieve greater statistical power for each topic, responses were then aggregated over all articles (p-value threshold = 0.05 / 17 = 0.003). Finally, the ratings for automatic extraction of relations by SemRep and Casama were compared to that of manual annotation (p-value threshold = 0.05 / 17 = 0.003).

8.3 Results

8.3.1 SemRep vs. Casama per topic per article

Tables 8.1 and 8.2 show the p-values produced by the Wilcoxon rank sums test. As expected, the sample size was too small to be statistically significant for any topic.

	EGFR (lung)	BRAF (melanoma)	Treatment history (lung)	Targeted therapy (lung)	Medscape (breast)
Identification of drugs	0.083	0.51	0.39	0.51	0.51
Effectiveness of drugs	0.043	0.081	0.021	0.05	0.081
Clinical guidelines	0.030	0.081	0.021	0.05	0.05
Side effects	0.67	0.28	0.083	1.0	0.05
Identification of genes	0.25	0.38	0.89	0.38	1.0
Prognostic effects of genes	0.11	0.05	0.15	0.81	0.05
Clinical characteristics of genes	0.021	0.13	0.89	0.05	0.38
Biochemical characteristics of genes	1.0	0.83	0.77	1.0	1.0
Diagnostic tests	0.67	0.081	1.0	0.05	1.0
Strength of evidence	0.021	0.05	0.021	0.05	0.05
Overall summarization quality	0.021	0.05	0.030	0.05	0.05
Comprehensibility	0.19	0.081	0.043	0.05	0.05
Clinical decision support	0.043	0.05	0.083	0.081	0.13
Precision medicine	0.083	0.05	0.083	0.13	0.081
Evidence based medicine	0.021	0.05	0.030	0.05	0.05
Meta-analysis	0.043	0.13	0.030	0.05	0.19
General biomedical research	0.060	0.081	0.15	0.081	0.05

Table 8.2: Group B: P-values for Wilcoxon rank sums, testing the hypothesis that Casama's scores tend to be higher than SemRep's. Significance threshold was 0.0006 due to Bonferroni correction.

8.3.2 SemRep vs. Casama per topic over all articles

Tables 8.3 and 8.4 present the median scores for each topic when aggregated over all articles and the p-values for the Wilcoxon rank sums test.

Groups A and B rated Casama similarly to SemRep for identification of drugs/treatments, side effects, identification of genes/biomarkers, clinical characteristics of genes/biomarkers, biochemical characteristics of genes/biomarkers, and diagnostic tests. For all remaining topics, Casama received significantly high scores compared to SemRep.

8.3.3 Automatic extraction

Table 8.5 presents the p-values for the Wilcoxon rank sums test comparing SemRep's automatically extracted relations to Casama's automatically extracted relations and contexts. No individual topic showed significantly higher Casama scores; however, Casama did outperform SemRep when aggregated over all topics (p = 4.7 e-05).

Topic	Median (SemRep)	Median (Casama)	p-value
Identification of drugs	4	4	0.079
Effectiveness of drugs	2	4	**2.8e-06**
Clinical guidelines	2	4	**6.5e-06**
Side effects	3	4	0.16
Identification of genes	3	3	0.044
Prognostic effects of genes	2	3	**0.00032**
Clinical characteristics of genes	2	3	0.027
Biochemical characteristics of genes	2	3	0.22
Diagnostic tests	2	3	0.55
Strength of evidence	2	4	**6.5e-07**
Overall summarization quality	2	4	**1.5e-05**
Comprehensibility	2	4	**0.00089**
Clinical decision support	2	4	**4.3e-06**
Precision medicine	2	3	**1.8e-05**
Evidence-based medicine	2	4	**1.7e-05**
Meta-analysis	2	4	**3.6e-05**
General biomedical research	2	4	**0.00032**

Table 8.3: Group A: Median scores and p-values for Wilcoxon rank sums when aggregated over all articles, testing the hypothesis that Casama's scores tend to be higher than SemRep's. Bold: $p \leq 0.003$ (significant after Bonferroni correction).

Topic	Median (SemRep)	Median (Casama)	p-value
Identification of drugs	4	5	0.026
Effectiveness of drugs	2	4	**7.5e-07**
Clinical guidelines	2	5	**2.2e-06**
Side effects	2	3	0.014
Identification of genes	4	4	0.26
Prognostic effects of genes	1	5	**4.8e-05**
Clinical characteristics of genes	1	3	0.013
Biochemical characteristics of genes	1	1	0.89
Diagnostic tests	1	3	0.11
Strength of evidence	1	5	**6.5e-07**
Overall summarization quality	2	4	**1.1e-06**
Comprehensibility	2	3.5	**7.6e-05**
Clinical decision support	2	5	**6.0e-05**
Precision medicine	2	4	**4.8e-05**
Evidence-based medicine	2	5	**1.4e-06**
Meta-analysis	2	4	**2.5e-05**
General biomedical research	2	4	**8.0e-05**

Table 8.4: Group B: Median scores and p-values for Wilcoxon rank sums when aggregated over all articles, testing the hypothesis that Casama's scores tend to be higher than SemRep's. Bold: $p \leq 0.003$ (significant after Bonferroni correction).

Topic	p-value
Identification of drugs	0.89
Effectiveness of drugs	0.043
Clinical guidelines	0.56
Side effects	0.15
Identification of genes	0.19
Prognostic effects of genes	0.083
Clinical characteristics of genes	0.043
Biochemical characteristics of genes	0.89
Diagnostic tests	0.89
Strength of evidence	0.11
Overall summarization quality	0.043
Comprehensibility	0.08
Clinical decision support	0.15
Precision medicine	0.06
Evidence-based medicine	0.25
Meta-analysis	0.06
General biomedical research	0.47
All topics	**4.7e-05**

Table 8.5: P-values for Wilcoxon rank sums, testing the hypothesis that Casama's scores for automatically extracted relations and contexts tend to be higher than SemRep's. Bold: $p \leq 0.003$ (significant after Bonferroni correction).

Finally, manual annotations were compared with automatically extracted relations for both SemRep and Casama (Table 8.6). SemRep's automatic extraction of relations on EGFR mutation in lung cancer was not significantly different from manual annotations on the same topic, either individually or in aggregate.

Similarly for Casama, no significant effects were seen for the individual topics. However, when aggregating Casama scores over all topics, manual annotation of Casama relations/contexts significantly outperformed automatic extraction ($p = 0.00061$).

8.3.4 Free-text comments

The free-text comments were generally favored Casama over SemRep. Four users noted a larger number of concept and relation types with SemRep, resulting in more difficulty in finding the desired information. In contrast, users described Casama's representation as "concise" and "easy to search." Three users felt that the relations targeted by SemRep (such as "chemotherapy **treats** human patients") were "very broad" and "not useful," whereas Casama

Topic	Manual vs. automatic (SemRep)	Manual vs. automatic (Casama)
Identification of drugs	0.89	0.25
Effectiveness of drugs	1.0	0.56
Clinical guidelines	0.25	0.19
Side effects	0.89	0.03
Identification of genes	0.56	0.56
Prognostic effects of genes	0.67	0.31
Clinical characteristics of genes	0.56	0.25
Biochemical characteristics of genes	1.0	0.89
Diagnostic tests	0.021	0.31
Strength of evidence	1.0	0.25
Overall summarization quality	1.0	0.15
Comprehensibility	0.22	0.38
Clinical decision support	0.77	0.083
Precision medicine	1.0	0.39
Evidence-based medicine	0.89	0.39
Meta-analysis	0.67	0.39
General biomedical research	0.89	0.31
All topics	0.67	**0.00061**

Table 8.6: P-values for Wilcoxon rank sums, comparing manual annotation to automatic extraction. Bold: $p \leq 0.003$ (significant after Bonferroni correction).

was described as "detailed" and "focused." Two users stated that the SemRep relations were "repetitive" or "redundant," which one user wrote might be useful for users unfamiliar with the knowledge space.

Certain Casama concept and relation types were identified as being most useful: *biomarkers*, *treatments*, and *outcomes* were called out specifically for helping users locate relevant information. In particular, users appreciated that each concept type was unique (as opposed to SemRep, in which multiple semantic types such as "pharmacologic substance" and "therapeutic procedure" both refer to treatment). Users also named the **improves** and **recommended for** relations as very helpful in capturing effectiveness of drugs and associated clinical guidelines.

Casama's contextual elements were also viewed positively by the users, especially contexts related to strength of evidence. Nearly every user expressed appreciation for "clinical trial information" or "whether or not a clinical trial determined the results." One user said of Casama's representation for strength of evidence, "This is where Casama excels." There

were fewer comments related to Casama's representation of patient characteristics, though these were generally favorable as well ("stage of cancer or past treatment is very helpful," "Casama provided more nuance in clinical characteristics of genes/biomarkers.")

Lastly, users gave examples of relevant information that was not targeted by SemRep or Casama. These included: statistical details (p-value thresholds, hazard ratios), knowledge spanning multiple sentences, names of clinical trials, names of agencies with published guidelines, and details of diagnostic tests (clinical recommendations, scoring mechanisms).

8.4 Discussion

This user evaluation showed that Casama outperformed SemRep on many topics and applications, although statistical significance was not reached in the per-topic-per-article evaluation. Positive ratings for Casama were more apparent when aggregating user scores over all articles, providing compelling evidence for the value of Casama's contextualized relations across domains. Indeed, Casama's relation and context types are general enough to be applied in many areas (e.g., specific treatment names may vary for different diseases, but the concept of *treatment history* is very generally applicable).

Comparisons of user ratings between SemRep and Casama demonstrated that there does exist overlap between the representations, particularly in the identification of drugs/treatments and genes/biomarkers. However, Casama did better than SemRep in the representation of strength of evidence, which was highly rated both quantitatively and qualitatively.

One important observation is that SemRep's manual annotations are not significantly different from SemRep's automatic annotations; in contrast, Casama's manual annotations outperformed Casama's automatically extracted relations. Therefore, SemRep's fundamental limitation is its representation rather than its relation extraction method; the opposite is true for Casama. A first priority for future work would be improvement of Casama's relation extraction algorithms (see: Chapter 6, Section 6.4.1). Despite this, Casama's summaries produced via automatic extraction outperformed that of SemRep.

A limitation of this study design is the relatively small number of evaluators and the finite amount of time available to review the articles and summaries. Consequently, no article was reviewed by all evaluators, potentially resulting in variability between Groups A and B. While ratings were consistent between Groups A and B, a larger evaluation with more reviewers per article would be ideal. An evaluation with clinical experts would also reveal further insights about the summarization systems.

While I attempted to be as objective as possible in creating the annotations and designing the questionnaire, it is impossible for me, the designer of the Casama representation, to be perfectly unbiased. A future evaluation should utilize external annotators; the questionnaire should be designed with respect to an expert-curated set of summarization requirements.

8.5 Conclusion

This chapter described an evaluation study in which users rated the summarization quality of Casama and SemRep. While both representations achieved high scores for identification of drugs and genes, Casama outperformed SemRep in capturing knowledge related to strength of evidence, drug effectiveness, clinical guidelines, and more. Casama was also highly rated for overall summarization quality and applications such as evidence-based medicine.

CHAPTER 9

Conclusion

9.1 Introduction

This chapter provides an overall summary of the dissertation, discusses findings and contributions synthesized over multiple chapters, and points to limitations and future work.

9.2 Summary of the dissertation

This dissertation presented Casama, a summarization system based on "contextualized semantic maps," a novel representation for enhancing relation-based summaries of biomedical literature with contextual knowledge about study features and patient population. An instance of the Casama representation was developed in the domain of driver mutations in non-small cell lung cancer (NSCLC), encompassing relations between treatments, outcomes, biomarkers, clinical-pathologic features, and detection methods. Study context included methodological design, cohort size, endpoints, and p-values; study population context included personal attributes, disease features, treatment history, and response.

A variety of methods were used to evaluate Casama as a representation and a summarization system. Casama performed equal to or better than PubMed in classification of study objectives and study designs. Automatic extraction of relations and context showed modest but promising results. The automatically-extracted study population contexts were used to retrieve more relevant articles compared with PubMed (assuming the Casama representation is complete and up-to-date). Finally, a user evaluation of Casama as a summarization system demonstrated significant advantage of Casama over a context-free representation, SemRep.

9.3 Contributions

In addition to Casama's technical contributions (e.g., annotated gold standard, automatic extraction methods), Casama's broader contributions to the research community include: a generalizable method for structuring knowledge as contextualized relations; a representation for strength of evidence and methods for extracting these features accurately; and a representation for patient/population features in cancer.

The system presented here is tailored to the domain of lung cancer; however, the central contribution of this work is a formal method for defining the relations, study contexts, and population contexts that can be applied in any domain of interest. Literature review, existing ontologies, and iterative consultation with experts were all used to define and structure the knowledge space. The "contextualized semantic map" generated by integrating the relations and contexts discovered in biomedical literature was demonstrated to improve information retrieval and summarization compared to established techniques.

Given that the representation developed in this work is domain-specific, it is important to ask whether Casama would generalize well to other domains. The balance between specificity and generalizability is a delicate one: an overly general representation may not capture the richness of the desired knowledge; in contrast, a highly domain-specific representation may prove rigid and unscalable. This issue was addressed at several points in this dissertation. Chapter 5 concluded that Casama's study design classifier could generalize well, given that the classifier was robust to differences in vocabulary and top classification features were not specific to lung cancer. To a lesser extent, this was also true of the study objectives classifier. Chapter 8 evaluated Casama as a summarization system across a variety of topics, including driver mutations in lung cancer, general topics in lung cancer, and driver mutations in non-lung cancers. In this evaluation, user ratings were consistently high when aggregating across all articles types, suggesting that Casama can indeed capture knowledge across various domains within cancer.

In particular, Casama's study context representation performed well across all evaluations. Kappa agreement was excellent in the annotation of study context. Automatic

classification of study design was equal to or better than PubMed; automatic extraction of study context was high for most study context types. Users rated Casama's summarization quality for strength of evidence very favorably (median Likert score of 4-5, indicating very good or excellent). Additionally, several users stated explicitly that including information about clinical trials was very valuable. Thus, the handling of study context is one strength of Casama. Application of contextualized relations to evidence-based decision support is a promising direction for future investigation.

Within the domain of lung cancer, Casama's representation of patient/population features was based on existing lung cancer ontologies and was augmented by experts to include knowledge of driver mutations, targeted therapy, and imaging features. Thus, Casama's representation brings together knowledge that previously has not been captured by a single ontology. This type of contextual knowledge is in accordance with a trajectory in clinical care: precision medicine. One application of such a representation is patient-tailored information retrieval, demonstrated in Chapter 7.

This dissertation has demonstrated that strength of evidence and population contexts are not only useful, but necessary elements in clinical decision support systems that examine biomedical literature. A rich representation, such as the one developed for Casama, is required to enable clinicians to make decisions that are evidence-based and individually tailored. Importantly, this work has presented a method for operationalizing the process of developing, implementing, and evaluating such a representation.

9.4 Limitations

The representation and capture of numerical data is a definite area of improvement. Users evaluating Casama as a summarization system noted that statistical information such as confidence intervals and hazard ratios are relevant to the interpretation of clinical trial results and should be included in the summaries.

Furthermore, in the annotation task described in Chapter 4, Section 4.4, it was decided that numerical comparisons (e.g., "Progression-free survival was 6 months for erlotinib vs.

3 months for docetaxel") would not be annotated. Rather, Casama focuses on relations stated qualitatively (e.g., "Progression-free survival was higher with erlotinib vs. docetaxel") to simplify the automatic extraction process. But because both types of sentences express the same Casama relation ("erlotinib **improves** progression-free survival"), a more robust system would be able to extract the numerical data, perform the appropriate comparison, and distill this into relational form. Ultimately, sentences containing numerical data were still problematic for Casama's automatic extraction system, as relations expressed numerically were not annotated, but the co-occurence between treatments and outcomes were tagged by Casama as valid relations (see: Chapter 6, Section 6.4).

Indeed, Casama would benefit greatly from improved automatic extraction algorithms. Only a subset of all Casama relations and contexts were targeted for automatic extraction; of these, precision was modest and recall was low. The impact of these results on summarization quality was demonstrated in Chapter 8: manual annotations significantly outperformed automatically extracted Casama relations and contexts. Improved lexicons for concept identification and a relation extraction system trained on biomedical literature could yield significant benefits to Casama overall.

Nearly every experiment in this dissertation relied on multiple individuals for annotation, relevance judgments, and user evaluation. While doing so minimizes the amount of bias inherent in using a single individual, it also introduces variability between individuals that must be acknowledged. In Chapter 7, no set of results was reviewed by more than one person. As a result, agreement between reviewers is unknown. It is possible that the judgments made were highly variable and noisy, especially given that the concept of "relevance" was very loosely defined. In Chapter 8, each summary was reviewed by 3-4 people; ideally, all 7 reviewers would have looked at each summary. Unfortunately, this was not possible due the amount of time and effort required. A more robust evaluation would include more users and more overlap between users. Inclusion of clinical experts in the evaluation (rather than graduate students or research staff) would also be highly desirable in future evaluations.

9.5 Future work

Development and evaluation of a visual summarization system for Casama remains an open area of investigation. A preliminary visualization is described in Appendix E. Furthermore, Casama currently only captures relations expressed within a single sentence. Co-reference resolution, an active research area in natural language processing, would enable the capture of relations expressed over multiple sentences.

An open question raised by this research is: how should expert systems deal with the dynamic nature of scientific knowledge? This issue was brought to light in Chapter 7, in which Casama's information retrieval performance was hindered by the incompleteness of the representation, which remained static from the time of development to the time of evaluation. Representations of knowledge must evolve, in content as well as structure, as new knowledge is discovered and new clinical practices are developed. The work presented here captures a moment in scientific knowledge; how representational systems adapt to change will be an intriguing new frontier.

Going beyond the applications described in this dissertation, a representation for population contexts could prove useful at a consumer level as more non-experts search for health-related information online. Expert-curated summaries of disease are available, but consumers may find it difficult to apply this knowledge to themselves or their loved ones without a knowledge model of the disease. Giving consumers access to a structured representation of relevant patient/population attributes can help them target their searches more meaningfully.

The improvement of patient-tailored retrieval with Casama's structured representation suggests that this information should be leveraged in the document indexing process. One way to accomplish this would be to require authors to provide this data as part of the publishing process. This proposition has several advantages: the information would not be subject to errors introduced by automatic extraction; no single person or group of people would bear the burden of annotating a large document set; improved retrieval of one's articles provides significant incentive for authors to participate. To achieve this, authors must adhere to a standardized representation and vocabulary (preferably mediated by a

centralized organization such as the National Library of Medicine). The representation such as Casama could be a first step in this direction.

9.6 Conclusion

This dissertation has described a method for representing biomedical literature that captures the findings of the study (relations) and their associated contexts (study design, study population). This approach was shown to add substantial value to information retrieval and summarization applications, particularly for clinical decision support systems seeking to enable evidence-based and precision medicine.

APPENDIX A

Glossary of abbreviations

ALK - Anaplastic lymphoma kinase

bpref - Binary preference

DCG - Discounted cumulative gain

DFS - Disease free survival

EGFR - Epidermal growth factor receptor

HSDB - Human Studies Database

LSP - Lexico-syntactic pattern

LUCADA - Lung Cancer Database (National Lung Cancer Audit)

NCI - National Cancer Institute

NDCG - Normalized discounted cumulative gain

NSCLC - Non-small cell lung cancer

OCRe - Ontology of Clinical Research

OS - Overall survival

P - Precision

PICO - Problem/Population, Intervention, Comparsion, Outcome

PMID - PubMed Identifier

PFS - Progression-free survival

R - Recall

ROC - Receiver operating characteristic

SVM - Support vector machine

TFIDF - Term frequency, inverse document frequency

TKI - Tyrosine kinase inhibitor

TREC - Text REtrieval Conference

UMLS - Unified Medical Language System

APPENDIX B

Annotation guidelines for document classification

You will be given a spreadsheet containing a set of abstracts on EGFR or ALK mutation in lung cancer.

Abstracts can be found in Column F of the spreadsheet. Please read each abstract carefully and annotate it for STUDY OBJECTIVE and STUDY DESIGN by filling out each row.

You may consult the full-text if available (hyperlink in Column P).

B.1 Study Objective

Please categorize each study into one or more of the following groups by placing a '1' in Column H (characterization), Column I (detection), Column J (treatment), and/or Column K (prognosis). If the abstract does not belong in the group, place a 0 in the cell. Classify abstracts by the purpose of the study. The categories are:

1. Mutation characterization studies

   These types of studies aim to discover clinical-pathologic features of a driver mutation, such as age, sex, smoking status, and histology. Also belonging to this category are mutation prevalence studies and studies that aim to characterize biomarkers for a driver mutation.

2. Mutation detection studies

   These types of studies demonstrate a method (either new or existing) for detecting driver mutations.

3. Treatment studies

   Treatment studies examine the effects, such as response or adverse events, of a drug regimen or other therapy.

4. Prognosis studies

   Prognosis studies associate driver mutations or clinical-pathologic features with outcomes such as survival, tumor response, or adverse events.

Short version using typical examples:

Study identifies prevalence of mutation in a defined population → characterization

Study compares clinical features between mutated and wild-type groups → characterization

Study demonstrates a method for detecting driver mutations → detection

Study examines the effect of a treatment, perhaps comparing it to another treatment → treatment

Study compares outcomes such as survival between mutated and wild-type groups → prognosis

Some studies will not fit into any category. Studies on cell lines or mice, case reports, reviews, meta-analyses, cost analyses, papers not about EGFR, ALK, or NSCLC are considered out of scope. Place a 0 in rows H-K, and an "na" in the study design column.

Sometimes the difference between various document classes can be very subtle. Classify the studies by purpose — mere mentions aren't enough. Ask yourself what the variables or comparison groups are in this study. For example, a study that compares response to gefitinib in EGFR-mutated vs. EGFR-wild type groups is a prognosis study, because the variable in question is presence of driver mutation. In contrast, a study that compares response to gefitinib vs. chemotherapy is a treatment study, because the variable is the type of treatment applied.

Studies can fit in more than one category — for example, a study that compares smoking history between mutated and wild-type groups (thus it is a characterization study) and also

analyzes the clinical features that predicted improved survival (thus it is a prognosis study).

B.2 Study Design

Please annotate each study for STUDY DESIGN by entering the corresponding code into Column O.

1. Experimental studies.

   These types of studies apply some kind of intervention to the patient and observe the results. Clinical trials fall into this category.

   Code: ex

2. Cohort studies.

   Cohort study is a type of observational study (i.e., no intervention is applied).

   Various cohorts (groups of patients differing by the variable in question) are defined and compared. Observations are made at more than one timepoint (thus, temporal outcomes such as survival can be assessed). If possible, differentiate between cohort studies that are prospective (the outcome of the study is not known at the beginning of the study) or retrospective (study looks back on old data where the outcome has already occurred).

   Codes:

   prospective cohort study: pc

   retrospective cohort study: rc

   cohort study, cannot differentiate further: ch

3. Case control studies.

   These studies differ from cohort studies in that patients are selected based on having the outcome/event in question. These "cases" are compared to a group that did not

have the outcome/event (these are the "controls"). The investigators look back in time to determine factors leading to that outcome/event.

Code: cc

4. Case series.

   These studies are descriptive (rather than analytical) and describe the experiences of a group of patients (perhaps who share a common clinical-pathologic feature or treatment history). There is no control group.

   Code: cs

5. Cross-sectional studies.

   These type of studies make an observation of the population at a single timepoint. Prevalence studies will fall into this category.

   Code: xs

APPENDIX C

Annotation guidelines for concept and relation annotation

You will be presented with a set of sentences from abstracts on EGFR and ALK mutations in lung cancer. Your goal is to annotate the key pieces of information that summarize the findings of the study and their associated context. This knowledge is represented in the form of relations (findings) and concepts (context).

Below, we define the 3 types of annotations:

1. Study context (e.g., cohort size, statistical tests, p-values, etc.)
2. Population context
   (a) Relational population context: erlotinib improved PFS in *patients with EGFR mutation*
   (b) Eligibility criteria: the cohort consisted of *Japanese non-smoking patients*
3. Relations (e.g., erlotinib improved PFS)

C.1 Annotation rules

C.1.1 Study context

Annotate context by selecting the text span of the concept phrase and choosing the concept type.

Study design context is usually found in the Background or Methods section of an abstract, although mentions may occur in other sections as well. Study context includes:

stat_ test

Statistical tests are used to determine overall trends and/or statistical significance of results. Commonly used statistical tests are Cox proportional hazards, logistic regression, student's t-test, univariate analysis, multivariate analysis.

phase

Clinical trial phase (i, ii, iii, iv). Annotate as: "phase i" (not "i").

blinding

Whether the people involved in the clinical trial knew whether the drug was being administered or not (open-label, blind, double blind).

endpoint

The final clinical outcomes measured in a study. Common ones are: progression free survival, overall survival, quality of life, safety, toxicity.

cohort_ size

The number of patients included in the study population. If expressed only in terms of the size of 2 or more arms, annotate the size of each arm. Only include the size of cohort included in the analysis!

C.1.2 Population context

Annotate context by selecting the text span of the concept phrase and choosing the concept type.

The following clinical features describe the population of a study. Annotate these if they describe the eligibility criteria of a study (usually found in the Methods section) or to add context to a relation (found in Results and Conclusions sections).

Demographics

age

Annotate specific age groups ("> 65 years of age") as well as general categories ("elderly", "younger").

sex

Annotate "men," "women", "males," "females," etc.

race

Annotate ethnicity or nationality of the study population (e.g., Caucasian, Chinese, Western population, etc.)

smoking_history

Annotate terms describing history of smoking, e.g., "never smoking," "light smoking," "current smokers," etc.

Prognostic factors

performance_status

Annotate specific values for performance status ("ECOG 0 or 1") as well as general references to performance status, e.g., "fit," "PS," etc.

comorbidity

Annotate additional conditions of the patient such as other diseases, organ function values, etc.

Tumor features

biomarker

Biomarkers are molecular/genetic features of the tumor, such as EGFR mutation, EGFR wild-type, ALK re-arrangement, etc.

stage

Annotate clinical stage (i-iv) as well as general references to primary tumor or metastatic disease. Annotate "stage i" rather than "i".

histology

Annotate mentions of histology and other cellular features of tumor (e.g. adenocarcinoma, squamous cell carcinoma, poorly-differentiated nodules, etc.)

imaging_feature

Annotate imaging features such as opacity, enhancement, consolidation, etc.

Treatment responses

Annotate mentions of the following clinical response categories as features of the study population.

progression

Progression indicates that the disease worsened (e.g., tumor growth, more tumors). Progression may occur if a patient never responded to treatment, or initially responded and progressed later. Annotate mentions of progression status (if possible, with respect to a treatment history of targeted therapy or chemotherapy).

resistance

Resistance indicates that a patient's tumor is not sensitive to a treatment. Resistance may be de novo or acquired. Annotate mentions of resistance status (if possible, with respect to a treatment history of targeted therapy or chemotherapy).

recurrence

Recurrence indicates that the patient's disease was eradicated completely (for example, in the case of surgical resection or complete response), then returned.

stable_disease

Stable disease means that known tumors did not change in size, and no tumors appeared.

partial_response

Partial response means that there was at least a 50% decrease in tumor volume, but residual disease remains.

complete_response

Complete response means all detectable tumor has disappeared.

Treatment histories

Note: these refer to *history* of treatment in the study population, not the drugs whose effects/outcomes are being studied.

surgery_history

Annotate mentions of a history of surgery (e.g., resection, lobectomy, pneumectomy, etc.)

radiotherapy_history

Annotate mentions of a history of radiotherapy.

chemotherapy_history

Annotate mentions of a history of chemotherapy (e.g., platinum-based chemotherapy, docetaxel, cisplatin, pemetrexed, etc.)

targeted_therapy_history

Annotate mentions of a history of targeted therapy (e.g., EGFR TKIs, erlotinib, afatinib, gefitinib, etc.)

combined_therapy_history

Annotate mentions of combined therapy (e.g., EGFR TKIs with chemotherapy).

treatment_line

Treatment line tells us how many treatment types have been attempted previously (e.g. first-line = no prior therapy).

other_treatment_history

Catch-all for references to treatment history.

other_clinical_feature

Annotate clinical features not included in this representation.

C.2 Relations and relational context

Annotate relations by:

1. selecting the text span of the subject of the relation,
2. choosing a concept type for the subject,
3. selecting the text span of the object of the relation,
4. choosing a concept type for the object,
5. selecting the subject and dragging a relation arrow to the object,
6. selecting the relation type.

C.2.1 Relational context

Annotate context by selecting the text span of the concept phrase and choosing the concept type.

Relation context tells us more about the proposition besides A [relation] B. Most common examples are:

A [relation] B in [population] Example: erlotinib improved progression-free survival in EGFR+ patients

A [relation] B with p-value [p] Example: EGFR mutation was positively correlated with female sex ($p < 0.05$)

A [relation] B with [treatment] Example: EGFR mutation predicted better overall survival with erlotinib Note: use the *treatment* concept type in this case, as erlotinib is the drug being studied. In contrast, use one of the *treatment history* concept types to refer to a past history of treatment.

[treatment] [relation] [outcome] as [treatment_line] Annotate treatment line with respect to a treatment mentioned in the same sentence.

C.2.2 Concepts that participate in relations only

outcome

An outcome is the result of a treatment. Progression-free survival, overall survival, quality of life are commonly-studied outcomes. Outcomes can also be less specific results of a treatment intervention, such as "benefit," "efficacy," "toxicity."

rate

A rate is the numerical frequency of a mutation. (e.g., 10%)

detection_method

Detection methods such as PCR and IHC aim to detect biomarker status in biological samples.

material

Material is the type of biological sample used in a detection test, such as blood or tissue.

Concepts that can participate in relations or provide relational context

treatment

A treatment is any type of intervention intended to treat a condition in a patient. Use this concept type to annotate chemotherapy, targeted therapies, surgeries, etc.

biomarker

Biomarkers are molecular/genetic features of the tumor, such as EGFR mutation, EGFR wild-type, ALK re-arrangement, etc.

Example of *treatment* participating in a relation, with *biomarker* providing context:

"erlotinib improved progression-free survival in EGFR+ patients"

Example of *biomarker* participating in a relation, with *treatment* providing context:

"EGFR+ patients experienced better progression-free survival on erlotinib"

While these examples appear to be conveying the same information, they are actually quite different. The first example came from a treatment study, where the aim was to characterize the effects of erlotinib. In the second example, the study characterized EGFR status with respect to outcome (i.e., a prognosis study).

Any clinical feature (enumerated in the "study population" section above) can participate

in a relation or provide relational context.

C.2.3 Concepts that only provide relational context

pvalue

P-values provide a measure of statistical significance of the results. (e.g., 0.05, 0.01)

C.2.4 Relations

Use the given study objectives to inform you regarding which relations you'll expect to see.

Characterization studies

In characterization studies, clinical-pathologic features are correlated with biomarker status (and, occasionally, other clinical-pathologic features).

clinical_feature positive_ correlation biomarker | clinical_feature indicates that the feature and the biomarker tend to occur together with statistical significance.

clinical_feature negative_ correlation biomarker | clinical_feature indicates that the feature and the biomarker tend to not occur together with statistical significance.

clinical_feature no_ correlation biomarker | clinical_feature indicates that the feature and the biomarker appear to have no relationship.

clinical_feature correlation biomarker | clinical_feature indicates that the feature and the biomarker are related but the directionality is not stated.

Other types of characterization papers study the frequency of a mutation within a population.

biomarker has_ rate rate specifies the numeric value of the frequency of the mutation.

biomarker has_ higher_ rate_ in, has_ lower_ rate_ in, has_ similar_ rate_ in clinical_feature qualitatively describes the mutation rate in a certain population (specified by *clinical_feature*) This relation suggests an observed trend between the biomarker and clinical feature that did not reach statistical significance.

biomarker has_higher_rate_than, *has_lower_rate_than*, *has_similar_rate_to biomarker* compares mutation rates between biomarkers.

Note: only annotate mutation frequencies if the aim of the paper is the characterize a particular population (not simply descriptive statistics of the cohort being studied).

Detection studies

In detection studies, a method for detecting mutation status is demonstrated.

detection_method detects biomarker specifies the name of the method and the biomarker it detects.

biomarker detected_in material indicates the type of biological specimen used to detect the biomarker

detection_method detects_in material indicates the type of biological specimen used in the detection_method.

Treatment studies

In treatment studies, the association between treatments and outcomes is studied.

treatment improves, *does_not_improve outcome* indicates that the treatment did or did not lead to a desired outcome (e.g., longer survival).

treatment worsens, *does_not_worsen outcome* indicates that the treatment did or did not lead to a less desirable outcome.

treatment associated_with, *not_associated_with outcome* indicates an association or no association with the outcome, but *improves* or *worsens* is not appropriate (e.g., *treatment associated_with favorable safety profile*)

treatment recommended_for, *not_recommended_for clinical_feature* states that the treatment is or is not appropriate for a certain population.

Prognosis studies

In prognosis studies, clinical-pathologic features, biomarkers, and detection methods are associated with outcomes.

clinical_feature | *biomarker* | *detection_method predicts_better*, *does_not_predict_better outcome* indicates that a more desirable outcome was or was not predicted by the clinical feature / *biomarker* / *detection_method*.

clinical_feature | *biomarker* | *detection_method predicts_worse*, *does_not_predict_worse outcome* indicates that a less desirable outcome was or was not predicted by the clinical feature / *biomarker* / *detection_method*.

clinical_feature | *biomarker* | *detection_method predicts*, *does_not_predict outcome* indicates that an outcome was or was not predicted, but "better or "worse is not appropriate (e.g., *biomarker* predicts benefit, *biomarker* predicts resistance)

All types of studies

entity compared_with entity This relation is used to to indicate that an entity participating in the relation is being compared to another entity of the same type.

C.3 General instructions

Annotate sentences that *interpret* the findings of the study. Do not annotate raw, numeric results.

Do not annotate general remarks about previous work or descriptive statistics such as baseline demographics.

In some cases, there may be more than one relation and associated context in the sentence (for example: multiple p-values, each associated with a different relation). Use the *nary_relation* annotation type to tie relations to their context. Select the text span of the relation subject, then choose *nary_rel* from the dialog box. Select the nary-relation and drag the arrow to the relation subject, the relation object, and the associated context.

Aim to annotate phrases completely yet minimally (e.g., *advanced* NSCLC with *EGFR wild-type*).

Use the most semantically meaningful relation when possible (e.g., favor *predicts_better* over *predicts*, *improves* over *associated_with*)

Special case for *resistance* and *progression* concepts: should always be annotated with an overlapping treatment_history concept if possible.

APPENDIX D

Relevance judgments

42 year old woman with newly diagnosed stage IV EGFR mutant disease and no prior therapy. What is the best initial therapy for her?				
PMID	Title	Definitely relevant	Potentially relevant	Not relevant
24736073	Comparison of clinical outcomes following gefitinib and erlotinib treatment in non-small-cell lung cancer patients harboring an epidermal growth factor receptor mutation in either exon 19 or 21.	x		
26096453	Epidermal growth factor receptor exon 20 insertions in advanced lung adenocarcinomas: Clinical outcomes and response to erlotinib.	x		
26262682	Post-Progression Survival after EGFR-TKI for Advanced Non-Small Cell Lung Cancer Harboring EGFR Mutations.		x	
25057173	EGFR biomarkers predict benefit from vandetanib in combination with docetaxel in a randomized phase III study of second-line treatment of patients with advanced non-small cell lung cancer.		x	
25288198	Phase Ib study evaluating a self-adjuvanted mRNA cancer vaccine (RNActive) combined with local radiation as consolidation and maintenance treatment for patients with stage IV non-small cell lung cancer.		x	
25261231	A single-arm, multicenter, safety-monitoring, phase IV study of icotinib in treating advanced non-small cell lung cancer (NSCLC).		x	
25349291	Phase II trial of stereotactic body radiation therapy combined with erlotinib for patients with limited but progressive metastatic non-small-cell lung cancer.		x	
24439569	Does KRAS mutational status predict chemoresistance in advanced non-small cell lung cancer (NSCLC)?			x
25202368	A new receptor tyrosine kinase inhibitor, icotinib, for patients with lung adenocarcinoma cancer without indication for chemotherapy.		x	
24263064	First-line gefitinib in Caucasian EGFR mutation-positive NSCLC patients: a phase-IV, open-label, single-arm study.	x		

Table D.1: Casama results for Case 1.

63 year old woman with an EGFR mutation. Received erlotinib, followed by carboplatin and pemetrexed at progression followed by afatinib at progression. Now again progressing. Should she undergo a biopsy to evaluate whether she has a T790M mutation?				
PMID	Title	Definitely relevant	Potentially relevant	Not relevant
25841484	Patient reported outcomes from LUX-Lung 3: first-line afatinib is superior to chemotherapy-would patients agree?			x
26354527	Afatinib in Non-Small Cell Lung Cancer Harboring Uncommon EGFR Mutations Pretreated With Reversible EGFR Inhibitors.		x	
25517450	[Efficacy of first-line afatinib versus chemotherapy in EGFR mutation positive pulmonary adenocarcinoma].			x
25242668	Activity of the EGFR-HER2 dual inhibitor afatinib in EGFR-mutant lung cancer patients with acquired resistance to reversible EGFR tyrosine kinase inhibitors.	x		
25232040	Experience with afatinib in patients with non-small cell lung cancer progressing after clinical benefit from gefitinib and erlotinib.		x	
23912954	Rare and complex mutations of epidermal growth factor receptor, and efficacy of tyrosine kinase inhibitor in patients with non-small cell lung cancer.			x
26349474	EGFR-TKI rechallenge with bevacizumab in EGFR-mutant non-small cell lung cancer.	x		
24789720	The application of real-time PCR technique to detect rare cell clones with primary T790M Substitution of EGFR gene in metastases of non-small cell lung cancer to central nervous system in chemotherapy naive patients.		x	
24493829	The impact of EGFR T790M mutations and BIM mRNA expression on outcome in patients with EGFR-mutant NSCLC treated with erlotinib or chemotherapy in the randomized phase III EURTAC trial.		x	
26309190	Spatiotemporal T790M Heterogeneity in Individual Patients with EGFR-Mutant Non-Small-Cell Lung Cancer after Acquired Resistance to EGFR-TKI.			x

Table D.2: Casama results for Case 2.

67 year old woman with an EGFR mutation. Received erlotinib. Now underwent repeat biopsy. T790M negative, but small cell transformation noted on repeat biopsy. What is the optimal treatment approach for her?				
PMID	Title	Definitely relevant	Potentially relevant	Not relevant
26306902	Randomized Phase II Trial of Erlotinib Beyond Progression in Advanced Erlotinib-Responsive Non-Small Cell Lung Cancer.	x		
26003540	Phase II study of erlotinib in elderly patients with non-small cell lung cancer harboring epidermal growth factor receptor mutations.			x
26474959	Phase 1 study of romidepsin plus erlotinib in advanced non-small cell lung cancer.		x	
25435849	Intercalated chemotherapy and erlotinib for advanced NSCLC: high proportion of complete remissions and prolonged progression-free survival among patients with EGFR activating mutations.	x		
25669662	Phase I dose-escalation study of pilaralisib (SAR245408, XL147), a pan-class I PI3K inhibitor, in combination with erlotinib in patients with solid tumors.	x		
25170013	A phase I/II study combining erlotinib and dasatinib for non-small cell lung cancer.		x	
25450874	Prospective assessment of pemetrexed or pemetrexed plus platinum in combination with gefitinib or erlotinib in patients with acquired resistance to gefitinib or erlotinib: a phase II exploratory and preliminary study.		x	
25349291	Phase II trial of stereotactic body radiation therapy combined with erlotinib for patients with limited but progressive metastatic non-small-cell lung cancer.			x
26174465	A prospective, multicentre phase II trial of low-dose erlotinib in non-small cell lung cancer patients with EGFR mutations pretreated with chemotherapy: Thoracic Oncology Research Group 0911.		x	
25841484	Patient reported outcomes from LUX-Lung 3: first-line afatinib is superior to chemotherapy-would patients agree?		x	

Table D.3: Casama results for Case 3.

62 year old EGFR mutant man status post frontline carboplatin, paclitaxel and bevacizumab with maintenance bevacizumab and erlotinib, not progressing and rebiopsied. Noted to have a T790M mutation. Would this patient benefit from a change in therapy from his current erlotinib and bevacizumab?				
PMID	Title	Definitely relevant	Potentially relevant	Not relevant
25870087	Phase I/II Study of HSP90 Inhibitor AUY922 and Erlotinib for EGFR-Mutant Lung Cancer With Acquired Resistance to Epidermal Growth Factor Receptor Tyrosine Kinase Inhibitors.		x	
26349474	EGFR-TKI rechallenge with bevacizumab in EGFR-mutant non-small cell lung cancer.	x		
23912954	Rare and complex mutations of epidermal growth factor receptor, and efficacy of tyrosine kinase inhibitor in patients with non-small cell lung cancer.		x	
24478319	Poor response to erlotinib in patients with tumors containing baseline EGFR T790M mutations found by routine clinical molecular testing.	x		
26309190	Spatiotemporal T790M Heterogeneity in Individual Patients with EGFR-Mutant Non-Small-Cell Lung Cancer after Acquired Resistance to EGFR-TKI.	x		
26153496	A randomized, double-blind, placebo-controlled, phase III trial of erlotinib with or without a c-Met inhibitor tivantinib (ARQ 197) in Asian patients with previously treated stage IIIB/IV nonsquamous nonsmall-cell lung cancer harboring wild-type epidermal growth factor receptor (ATTENTION study).	x		
26306902	Randomized Phase II Trial of Erlotinib Beyond Progression in Advanced Erlotinib-Responsive Non-Small Cell Lung Cancer.	x		
26003540	Phase II study of erlotinib in elderly patients with non-small cell lung cancer harboring epidermal growth factor receptor mutations.		x	
24636848	Phase I/II trial of vorinostat (SAHA) and erlotinib for non-small cell lung cancer (NSCLC) patients with epidermal growth factor receptor (EGFR) mutations after erlotinib progression.			x
25923549	AZD9291 in EGFR inhibitor-resistant non-small-cell lung cancer.	x		

Table D.4: Casama results for Case 4.

27 year old woman with newly diagnosed EGFR mutant NSCLC with a T790M mutation and L858R mutation in the EGFR gene. Should she be tested for a germline T790M mutation?				
PMID	Title	Definitely relevant	Potentially relevant	Not relevant
24736066	Hereditary lung cancer syndrome targets never smokers with germline EGFR gene T790M mutations.	x		
26577492	Monitoring EGFR T790M with plasma DNA from lung cancer patients in a prospective observational study.			x
26572169	Next-generation sequencing of tyrosine kinase inhibitor-resistant non-small-cell lung cancers in patients harboring epidermal growth factor-activating mutations.		x	
26514492	Correlation between EGFR Gene Mutations and Lung Cancer: a Hospital-Based Study.		x	
24724747	Routine implementation of EGFR mutation testing in clinical practice in Flanders: 'HERMES' project.			x
25355724	Small-cell lung cancer detection in never-smokers: clinical characteristics and multigene mutation profiling using targeted next-generation sequencing.			x
25722667	Lung Adenocarcinoma with Pulmonary Miliary Metastases and Complex Somatic Heterozygous EGFR Mutation.	x		
26362141	EGFR mutation status in Middle Eastern patients with non-squamous non-small cell lung carcinoma: A single institution experience.		x	
25450875	Clinical likelihood of sporadic primary EGFR T790M mutation in EGFR-mutant lung cancer.		x	
26609535	The Impact of Sequence of Chemotherapy and EGFR-TKI Treatment on Different EGFR Mutation Lung Adenocarcinoma.		x	

Table D.5: Casama results for Case 5.

42 year old woman with newly diagnosed EGFR mutant disease and no prior therapy. What is the best initial therapy for her?				
PMID	Title	Definitely relevant	Potentially relevant	Not relevant
25702091	Statins augment efficacy of EGFR-TKIs in patients with advanced-stage non-small cell lung cancer harbouring KRAS mutation.			x
24994038	Mutations of EGFR or KRAS and expression of chemotherapy-related genes based on small biopsy samples in stage IIIB and IV inoperable non-small cell lung cancer.		x	
26639246	Radiotherapy effects on brain/bone metastatic adenocarcinoma lung cancer and the importance of EGFR mutation test.			x
25514801	Dynamic plasma EGFR mutation status as a predictor of EGFR-TKI efficacy in patients with EGFR-mutant lung adenocarcinoma.			x
24263064	First-line gefitinib in Caucasian EGFR mutation-positive NSCLC patients: a phase-IV, open-label, single-arm study.	x		
26047516	Classification of Epidermal Growth Factor Receptor Gene Mutation Status Using Serum Proteomic Profiling Predicts Tumor Response in Patients with Stage IIIB or IV Non-Small-Cell Lung Cancer.		x	
25936883	[Clinical Research of EGFR and KRAS Mutation in Fine Needle Aspiration Cytology Specimens of Non-small Cell Lung Carcinoma].			x
24682604	Features and prognostic impact of distant metastasis in patients with stage IV lung adenocarcinoma harboring EGFR mutations: importance of bone metastasis.			x
25589191	Afatinib versus cisplatin-based chemotherapy for EGFR mutation-positive lung adenocarcinoma (LUX-Lung 3 and LUX-Lung 6): analysis of overall survival data from two randomised, phase 3 trials.	x		
25538894	EGFR Mutation Positive Stage IV Non-Small-Cell Lung Cancer: Treatment Beyond Progression.		x	

Table D.6: PubMed results for Case 1.

63 year old woman with an EGFR mutation. Received erlotinib, followed by carboplatin and pemetrexed at progression followed by afatinib at progression. Now again progressing. Unknown T790M status. Should she undergo a biopsy to evaluate whether she has a T790M mutation?				
PMID	Title	Definitely relevant	Potentially relevant	Not relevant
25450875	Clinical likelihood of sporadic primary EGFR T790M mutation in EGFR-mutant lung cancer.		x	
24768581	Incidence of T790M mutation in (sequential) rebiopsies in EGFR-mutated NSCLC-patients.		x	
25939061	Acquired EGFR C797S mutation mediates resistance to AZD9291 in non-small cell lung cancer harboring EGFR T790M.		x	
26267891	The epidermal growth factor receptor (EGFR / HER-1) gatekeeper mutation T790M is present in European patients with early breast cancer.			x
26577492	Monitoring EGFR T790M with plasma DNA from lung cancer patients in a prospective observational study.		x	
25560642	Usefulness of nanofluidic digital PCR arrays to quantify T790M mutation in EGFR-mutant lung adenocarcinoma.		x	
24789720	The application of real-time PCR technique to detect rare cell clones with primary T790M Substitution of EGFR gene in metastases of non-small cell lung cancer to central nervous system in chemotherapy naive patients.		x	
26309190	Spatiotemporal T790M Heterogeneity in Individual Patients with EGFR-Mutant Non-Small-Cell Lung Cancer after Acquired Resistance to EGFR-TKI.			x
25091415	Structural signature of the G719S-T790M double mutation in the EGFR kinase domain and its response to inhibitors.		x	
24737599	Clinical outcome according to the level of preexisting epidermal growth factor receptor T790M mutation in patients with lung cancer harboring sensitive epidermal growth factor receptor mutations.		x	

Table D.7: PubMed results for Case 2.

67 year old woman with an EGFR mutation. Received erlotinib. Now underwent repeat biopsy. T790M negative, but small cell transformation noted on repeat biopsy. What is the optimal treatment approach for her?				
PMID	Title	Definitely relevant	Potentially relevant	Not relevant
26400668	Small cell lung cancer transformation and T790M mutation: complimentary roles in acquired resistance to kinase inhibitors in lung cancer.	x		
26557922	Histological transformation from non-small cell to small cell lung carcinoma after treatment with epidermal growth factor receptor-tyrosine kinase inhibitor.	x		
26187428	Emergence of RET rearrangement co-existing with activated EGFR mutation in EGFR-mutated NSCLC patients who had progressed on first- or second-generation EGFR TKI.			x
26473643	Mechanisms of Acquired Resistance to AZD9291, a Mutation-Selective, Irreversible EGFR Inhibitor.			x
26424310	An Autopsy Case of Two Distinct, Acquired Drug Resistance Mechanisms in Epidermal Growth Factor Receptor-mutant Lung Adenocarcinoma: Small Cell Carcinoma Transformation and Epidermal Growth Factor Receptor T790M Mutation.			x
24768581	Incidence of T790M mutation in (sequential) rebiopsies in EGFR-mutated NSCLC-patients.	x		
26152920	Shades of T790M: Intratumor Heterogeneity in EGFR-Mutant Lung Cancer.		x	
24457237	Small-cell carcinoma in the setting of pulmonary adenocarcinoma: new insights in the era of molecular pathology.		x	
24828667	Small-cell lung cancers in patients who never smoked cigarettes.	x		
25826094	Constitutive asymmetric dimerization drives oncogenic activation of epidermal growth factor receptor carboxyl-terminal deletion mutants.	x		

Table D.8: PubMed results for Case 3.

62 year old EGFR mutant man status post frontline carboplatin, paclitaxel and bevacizumab with maintenance bevacizumab and erlotinib, not progressing and rebiopsied. Noted to have a T790M mutation. Would this patient benefit from a change in therapy from his current erlotinib and bevacizumab?				
PMID	Title	Definitely relevant	Potentially relevant	Not relevant
26577492	Monitoring EGFR T790M with plasma DNA from lung cancer patients in a prospective observational study.			x
26309190	Spatiotemporal T790M Heterogeneity in Individual Patients with EGFR-Mutant Non-Small-Cell Lung Cancer after Acquired Resistance to EGFR-TKI.	x		
24768581	Incidence of T790M mutation in (sequential) rebiopsies in EGFR-mutated NSCLC-patients.	x		
25560642	Usefulness of nanofluidic digital PCR arrays to quantify T790M mutation in EGFR-mutant lung adenocarcinoma.			x
25450875	Clinical likelihood of sporadic primary EGFR T790M mutation in EGFR-mutant lung cancer.			x
24789720	The application of real-time PCR technique to detect rare cell clones with primary T790M Substitution of EGFR gene in metastases of non-small cell lung cancer to central nervous system in chemotherapy naive patients.			x
24737599	Clinical outcome according to the level of preexisting epidermal growth factor receptor T790M mutation in patients with lung cancer harboring sensitive epidermal growth factor receptor mutations.		x	
26267891	The epidermal growth factor receptor (EGFR / HER-1) gatekeeper mutation T790M is present in European patients with early breast cancer.			x
24478319	Poor response to erlotinib in patients with tumors containing baseline EGFR T790M mutations found by routine clinical molecular testing.	x		
25405807	Quantification and dynamic monitoring of EGFR T790M in plasma cell-free DNA by digital PCR for prognosis of EGFR-TKI treatment in advanced NSCLC.		x	

Table D.9: PubMed results for Case 4.

27 year old woman with newly diagnosed EGFR mutant NSCLC with a T790M mutation and L858R mutation in the EGFR gene. Should she be tested for a germline T790M mutation?				
PMID	Title	Definitely relevant	Potentially relevant	Not relevant
24736066	Hereditary lung cancer syndrome targets never smokers with germline EGFR gene T790M mutations.	x		

Table D.10: PubMed results for Case 5.

APPENDIX E

Visualization

E.1 Introduction

This appendix describes a tool developed for visualizing contextualized semantic maps. The Casama visualization suite includes features such as vocabulary standardization, use of color and size, and filtering. A custom plugin was developed that, inspired by force-based layouts, arranges nodes and edges in a semantically meaningful way.

E.2 Creating a contextualized semantic map

A contextualized semantic map was produced by loading the manually-annotated relations and contexts into a graph structure using the Python library networkx [SS08]. Each relation (edge) has a set of attribute-value pairs corresponding to the contextual types in the representation and their instances from the annotated document set. Gephi [BHJ09], an open-source, Java-based network visualization tool was used to render and manipulate the network (Figure E.1).

E.2.1 Vocabulary standardization

Because the raw annotations include a variety of expressions referring to the same concept (e.g., EGFR+, EGFR positive, EGFR mutation), vocabulary standardization is performed prior to creating the graph structure. In doing so, synonymous nodes are consolidated to minimize redundancy. Common abbreviations were detected by regular expression and replaced with their normalized forms. Table E.1 provides the complete list of standardized

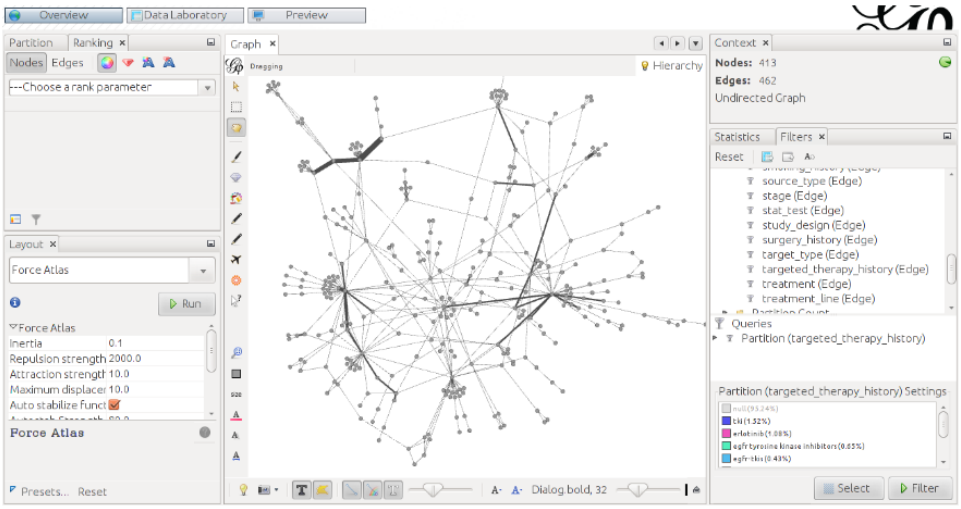


Figure E.1: Screenshot of Gephi interface. The central panel depicts the entire contextualized semantic map, generated from manual annotations. The left panel contains parameters for adjusting layout. The right panel contains the filtering interface.

Term	Normalized form
<biomarker> + mutation, mut, postive, +	<biomarker> mutation
<biomarker> + negative, wild, without	<biomarker> wild-type
<biomarker> + expression, deletion, insertion, rearrangement, translocation, amplification	<biomarker> + modifying text
progression free survival, progression-free survival, pfs	(no surrounding words) progression-free survival
overall survival, os	overall survival
disease-free survival, disease free survival, dfs	disease-free survival
mean survival time, mst	mean survival time
tyrosine kinase inhibitors, tki	tyrosine kinase inhibitors

Table E.1: Common terms and their normalized forms.

terms.

E.2.2 Basic functions: filtering, color, and size

Filtering is the easiest way to reduce the overall size of the graph. The contextualized semantic map produced from the manual annotations contains 570 nodes and 591 edges (Figure E.2). The user can create a filter, for example, that displays just the **predicts** relations with respect to overall survival. Filtering is performed by accessing a list of potential filters corresponding to the contextual types in the Casama representation (e.g., relation type, subject of the relation). When a filter is selected, the user is then shown a list of annotated

instances of this type of context. The user selects which instances to filter in and out (e.g., relation type = **predicts**, subject of the relation = overall survival). Filtering reduces the size of the graph and produces subgraphs the facilitate knowledge discovery relevant to the user's information need. Filtering may be also be applied iteratively, as results from the first filter help the user modify his or her query.

Gephi also provides mechanisms for depicting the statistical and semantic properties of nodes and edges using color and size. For example, the size of a node can be scaled by its number of associated edges. Similarly, the width of an edge is scaled by the number of times the relation was found in this document set. In this way, statistically "important" nodes and edges draw the eye.

Nodes and edges may also be color coded by their semantic properties. Coloring nodes by semantic type (e.g., *biomarker, treatment, outcome*) can help users quickly pick out concepts and relations of interest. Edges may be colored by any type of edge metadata, including relation type (e.g., **improves**, **predicts**), study design context (e.g., *study objective, study design*), or study population context (e.g., *biomarker, treatment history*).

Figure E.3 illustrates a filtered contextualized semantic map with node and edge scaling applied. It is easily observed that the relation between overall surival and biomarkers, especially EGFR mutation, KRAS mutation, and exon 19 deletion, is well-studied in this data set.

E.2.3 Semantic force layout

A custom Gephi plugin was developed for using the semantic knowledge provided by Casama to inform the layout of the graph. The plugin is based on Force Atlas, a layout style included in Gephi. Force-based layouts utilize a physical model of attraction and repulsion. Nodes repel, but connected nodes attract. After multiple iterations of attraction and repulsion, the graph settles into a balanced layout in which edge crossings are minimized.

The plugin developed for Casama is based on a similar principle of attraction and repulsion. Nodes repel, connected nodes attract, and nodes or edges containing similar semantic

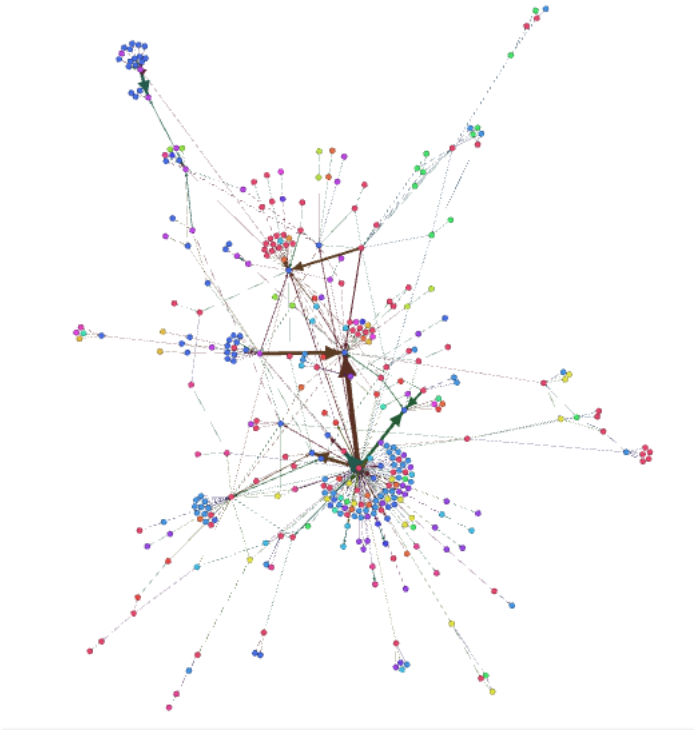


Figure E.2: The entire contextualized semantic map: 570 nodes and 591 edges.

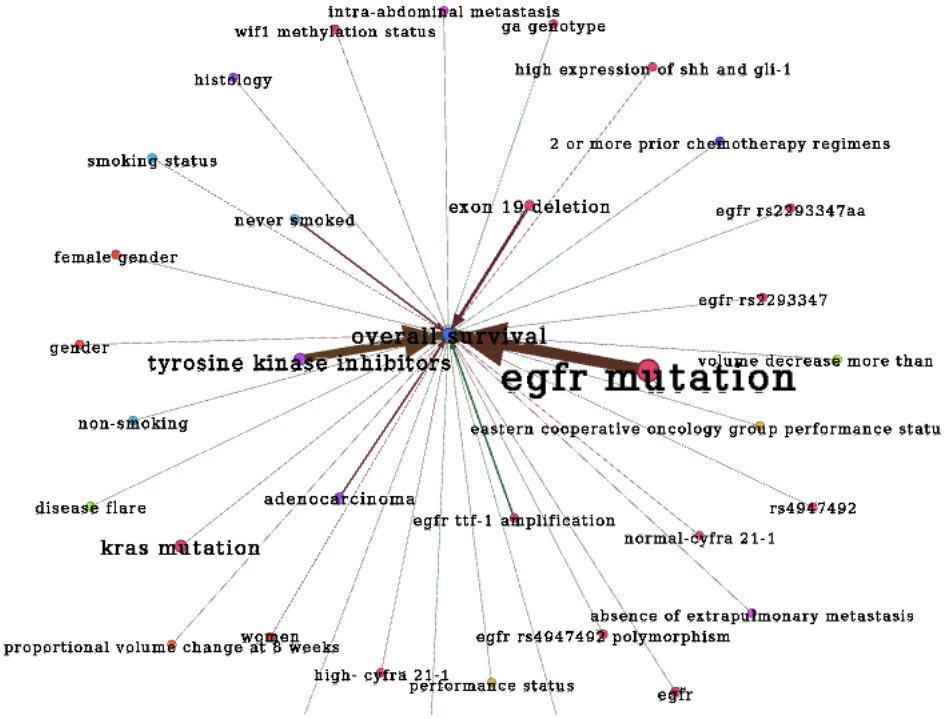


Figure E.3: The contextualized semantic map after 1) coloring nodes and edges by type, 2) scaling nodes by degree and edges by frequency, and 3) applying filters (relation_type=**predicts**, subject_of_relation=overall survival).

knowledge attract. The user may select a node attribute and an edge attribute on which to arrange a "semantic force layout." The plugin iterates over each pair of nodes and edges, adding attractive force if they share the same metadata for the user-selected attribute. An example of semantic force layout is illustrated in Figure E.4.

E.2.4 Spreadsheet view

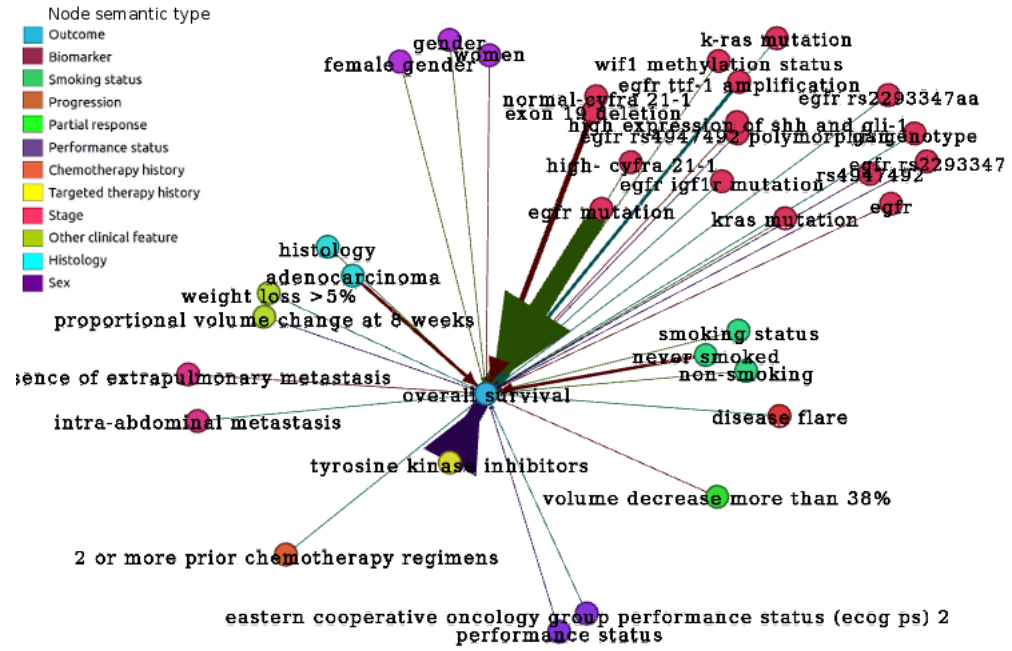
The contextualized semantic map may also be viewed in spreadsheet form, allowing the examination of relations and contexts in a tabular form. Spreadsheet view includes relations in their raw, unaggregated form (as opposed to the graphical view which combines multiple instances of a relation into a single edge). The user may sort and filter on raw concept mentions, normalized forms, relation types, and each type of study design and study population context. The spreadsheet can also be searched for specific terms of interest.

Spreadsheet view also includes article metadata such as authors, journal, and publication date. Importantly, the raw abstract is also available in spreadsheet view, allowing the user to trace relations back to their sources.

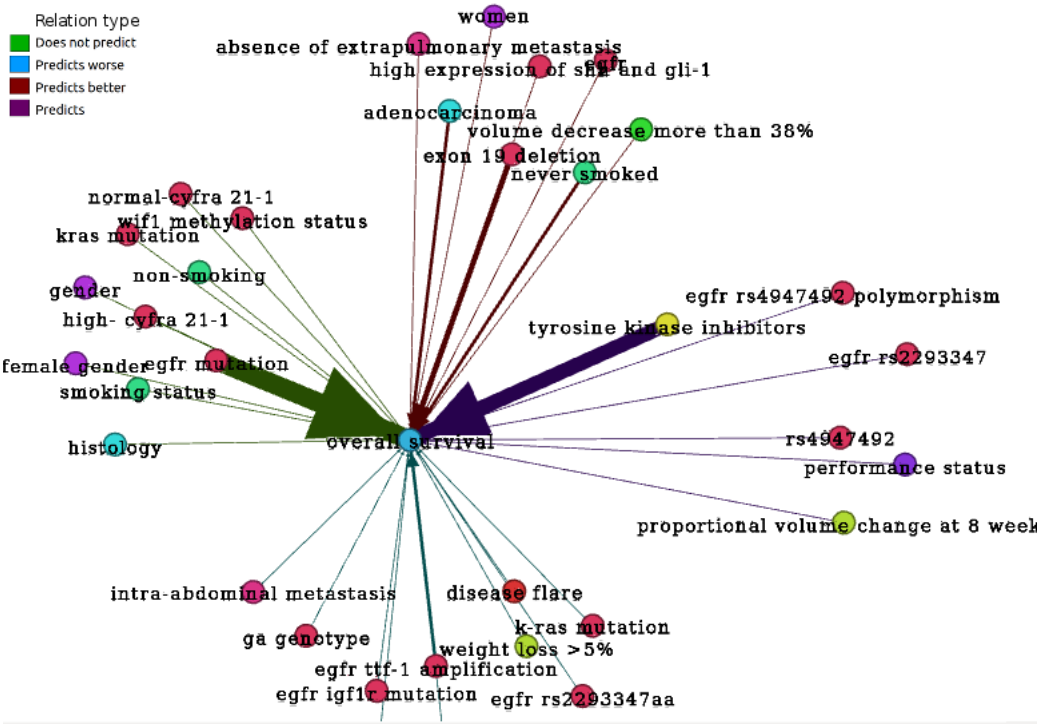
E.2.5 Filtering on study design context

Figures E.6 and E.7 are fragments of the contextualized semantic map examining treatment-oriented relations (**improves**, **associated_with**, and **recommended_for**). Figure E.6a has been filtered to include only relations found in experimental studies such as clinical trials. Erlotinib, docetaxel, and an erlotinib-pemetrexed sequence are identified as treatments associated with positive outcomes at the highest level of evidence. Newer treatments afatinib and matuzumab with paclitaxel are also identified.

If the user wishes to include more information, they may broaden the filter to include relations from prospective and retrospective studies (Figure E.6b). A new treatment node, "tyrosine kinase inhibitors", appears, in addition to its associated outcomes. The user may also observe that erlotinib is a recommended treatment after gefitinib failure, and that afatinib is associated with improved quality of life.

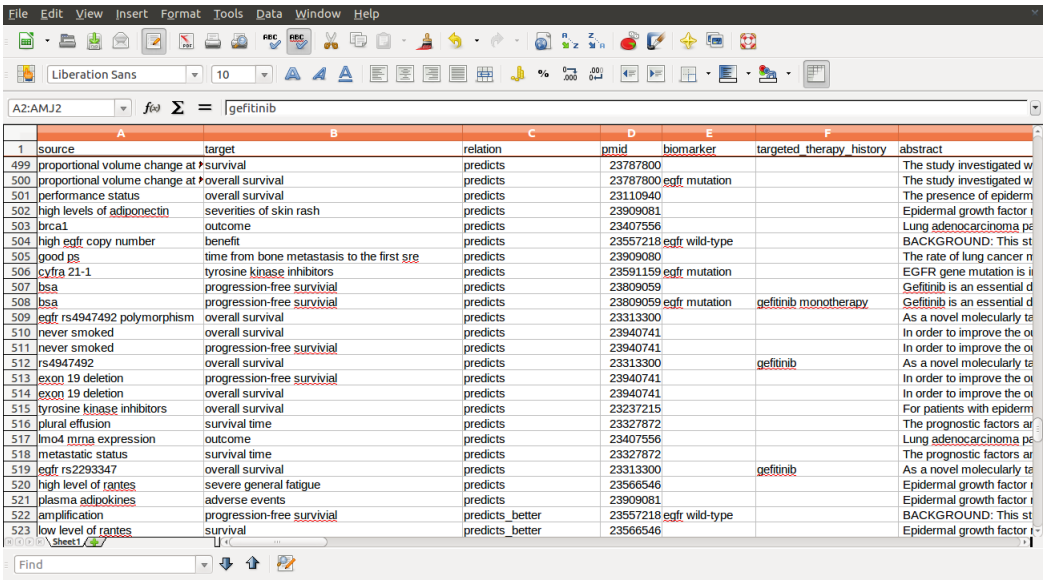


(a) Nodes clustered by semantic type.



(b) Edges clustered by relation type.

Figure E.4: Semantic force layout.

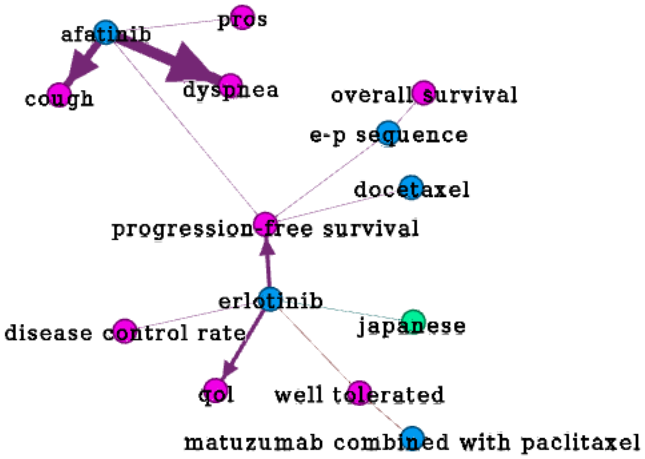


	A	B	C	D	E	F	
1	source	target	relation	pmid	biomarker	targeted_therapy_history	abstract
499	proportional volume change at	survival	predicts	23787800			The study investigated w
500	proportional volume change at	overall survival	predicts	23787800	egfr mutation		The study investigated w
501	performance status	overall survival	predicts	23110940			The presence of epiderm
502	high levels of adiponectin	severities of skin rash	predicts	23909081			Epidermal growth factor r
503	brca1	outcome	predicts	23407556			Lung adenocarcinoma pa
504	high egfr copy number	benefit	predicts	23557218	egfr wild-type		BACKGROUND: This st
505	good ps	time from bone metastasis to the first sre	predicts	23909080			The rate of lung cancer n
506	cyfra 21-1	tyrosine kinase inhibitors	predicts	23591159	egfr mutation		EGFR gene mutation is i
507	bsa	progression-free survivial	predicts	23809059			Gefitinib is an essential d
508	bsa	progression-free survivial	predicts	23809059	egfr mutation	gefitinib monotherapy	Gefitinib is an essential d
509	egfr rs4947492 polymorphism	overall survival	predicts	23313300			As a novel molecularly ta
510	never smoked	overall survival	predicts	23940741			In order to improve the ou
511	never smoked	progression-free survivial	predicts	23940741			In order to improve the ou
512	rs4947492	overall survival	predicts	23313300		gefitinib	As a novel molecularly ta
513	exon 19 deletion	progression-free survivial	predicts	23940741			In order to improve the ou
514	exon 19 deletion	overall survival	predicts	23940741			In order to improve the ou
515	tyrosine kinase inhibitors	overall survival	predicts	23237215			For patients with epiderm
516	plural effusion	survival time	predicts	23327872			The prognostic factors ar
517	lmo4 mrna expression	outcome	predicts	23407556			Lung adenocarcinoma pa
518	metastatic status	survival time	predicts	23327872			The prognostic factors ar
519	egfr rs2293347	overall survival	predicts	23313300		gefitinib	As a novel molecularly ta
520	high level of rantes	severe general fatigue	predicts	23566546			Epidermal growth factor r
521	plasma adipokines	adverse events	predicts	23909081			Epidermal growth factor r
522	amplification	progression-free survivial	predicts_better	23557218	egfr wild-type		BACKGROUND: This st
523	low level of rantes	survival	predicts_better	23566546			Epidermal growth factor r

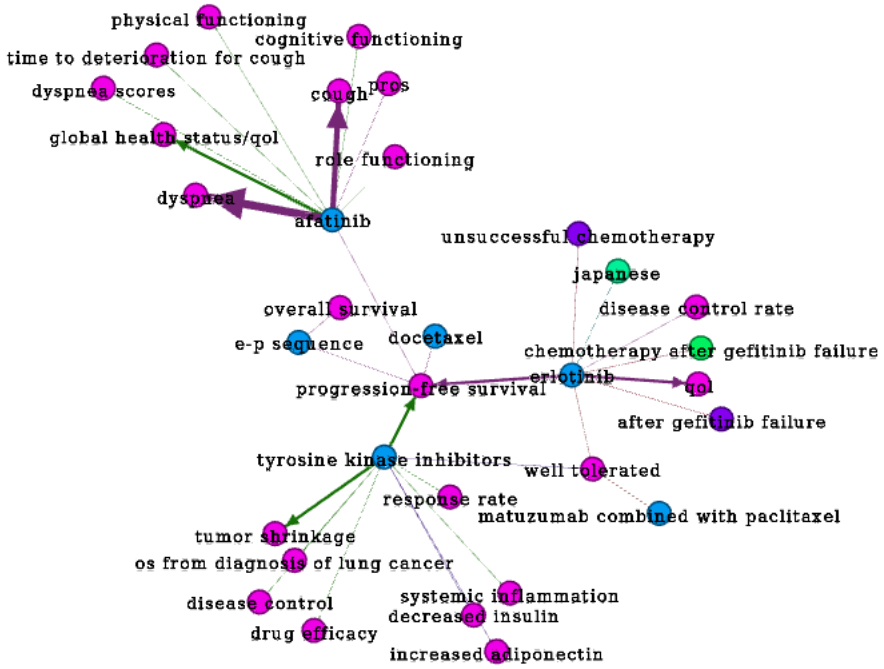
Figure E.5: Spreadsheet view, sorted by relation type.

E.2.6 Filtering on study population context

Consider a query pertaining to a patient with a prior history of EGFR tyrosine kinase inhibitors (TKIs). Figure E.7a is a fragment of the graph depicting the factors that influence overall survival (filter: subject_of_relation=overal survival). A variety of factors are displayed, including treatments, biomarkers, and demographic information such as smoking history and gender. However, when a sub-filter is applied to focus the graph on relations from study population with a history of EGFR-TKIs (filter: targeted_therapy_history=TKI, erlotinib, gefitinib), a more specific picture emerges (Figure E.7b). Certain biomarkers predict improved survival, whereas weight loss, presence metastases, poor performance status, and heavy pretreatment predict worse overall survival in a study population similar to the patient/query.

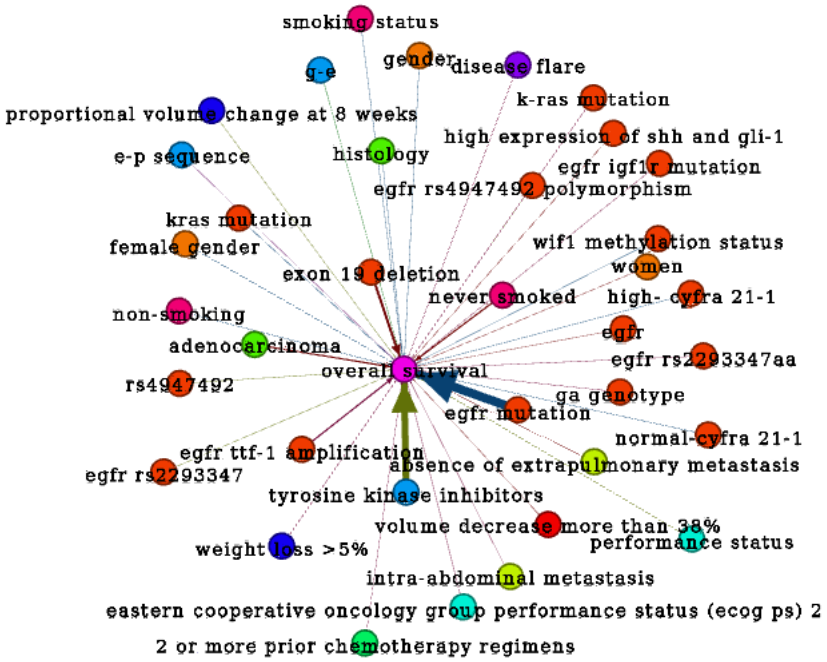


(a) Experimental studies only.

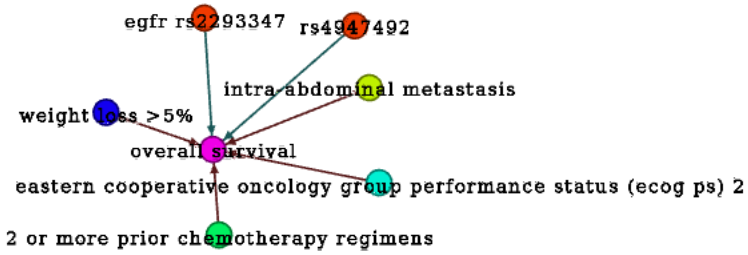


(b) Experimental studies, prospective cohort studies, and retrospective cohort studies.

Figure E.6: A fragment of the contextualized semantic map, examining treatment-oriented relations (improves, associated_with, recommended_for).



(a) No additional filters.



(b) targeted_therapy_history=EGFR tyrosine kinase inhibitors.

Figure E.7: A fragment of the contextualized semantic map examining factors that influence overall survival.

References

[AFD07] Caroline B Ahlers, Marcelo Fiszman, Dina Demner-Fushman, François-Michel Lang, and Thomas C Rindflesch. "Extracting semantic predications from Medline citations for pharmacogenomics." In *Pacific Symposium on Biocomputing*, volume 12, pp. 209–220. World Scientific, 2007.

[ANC13] Maria E Arcila, Khedoudja Nafa, Jamie E Chaft, Natasha Rekhtman, Christopher Lau, Boris A Reva, Maureen F Zakowski, Mark G Kris, and Marc Ladanyi. "EGFR exon 20 insertion mutations in lung adenocarcinomas: prevalence, molecular heterogeneity, and clinicopathologic characteristics." *Molecular cancer therapeutics*, **12**(2):220–229, 2013.

[ATA14] Noha Alnazzawi, Paul Thompson, and Sophia Ananiadou. "Building a semantically annotated corpus for congestive heart and renal failure from clinical records and the literature." *Proceedings of Louhi*, **14**:69–74, 2014.

[ATS05] Yindalon Aphinyanaphongs, Ioannis Tsamardinos, Alexander Statnikov, Douglas Hardin, and Constantin F. Aliferis. "Text categorization models for high-quality article retrieval in internal medicine." *Journal of the American Medical Informatics Association: JAMIA*, **12**(2):207–216, April 2005.

[BBC07] Brandon Beamer, Suma Bhat, Brant Chee, Andrew Fister, Alla Rozovskaya, and Roxana Girju. "UIUC: A knowledge-rich approach to identifying semantic relations between nominals." In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pp. 386–389. Association for Computational Linguistics, 2007.

[BCS07] Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. "Open information extraction for the web." In *IJCAI*, volume 7, pp. 2670–2676, 2007.

[BDS08] Markus Bundschus, Mathaeus Dejori, Martin Stetter, Volker Tresp, and Hans-Peter Kriegel. "Extraction of semantic biomedical relations from text using conditional random fields." *BMC bioinformatics*, **9**(1):207, 2008.

[BEE12] Michael Bada, Miriam Eckert, Donald Evans, Kristin Garcia, Krista Shipley, Dmitry Sitnikov, William A Baumgartner, K Bretonnel Cohen, Karin Verspoor, Judith A Blake, et al. "Concept annotation in the CRAFT corpus." *BMC bioinformatics*, **13**(1):161, 2012.

[BGK05] Razvan Bunescu, Ruifang Ge, Rohit J Kate, Edward M Marcotte, Raymond J Mooney, Arun K Ramani, and Yuk Wah Wong. "Comparative experiments on learning information extractors for proteins and their interactions." *Artificial intelligence in medicine*, **33**(2):139–155, 2005.

[BHG09] Jari Björne, Juho Heimonen, Filip Ginter, Antti Airola, Tapio Pahikkala, and Tapio Salakoski. "Extracting Complex Biological Events with Rich Graph-based Feature Sets." In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task*, BioNLP '09, pp. 10–18, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[BHJ09] Mathieu Bastian, Sebastien Heymann, Mathieu Jacomy, et al. "Gephi: an open source software for exploring and manipulating networks." *ICWSM*, **8**:361–362, 2009.

[Bio13] Lister Hill National Center for Biomedical Communications. "Semantic Knowledge Representation." `https://semrep.nlm.nih.gov/GoldStandard.html`, 2013. Accessed: 2016-06-06.

[BKL09] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, Beijing ; Cambridge Mass., 1 edition edition, July 2009.

[BL15] Catherine Blake and Ana Lucic. "Automatic endpoint detection to support the systematic review process." *Journal of Biomedical Informatics*, **56**:42 – 56, 2015.

[BMZ13] Joshua Bauml, Rosemarie Mick, Yu Zhang, Christopher D Watt, Anil Vachani, Charu Aggarwal, Tracey Evans, and Corey Langer. "Frequency of EGFR and KRAS mutations in patients with non small cell lung cancer by racial background: do disparities exist?" *Lung Cancer*, **81**(3):347–353, 2013.

[Bod04] Olivier Bodenreider. "The unified medical language system (UMLS): integrating biomedical terminology." *Nucleic acids research*, **32**(suppl 1):D267–D270, 2004.

[Bre03] Patrick Brézillon. "Focusing on context in human-centered computing." *IEEE Intelligent Systems*, **18**(3):62–66, 2003.

[BS11] Jari Björne and Tapio Salakoski. "Generalizing Biomedical Event Extraction." In *Proceedings of the BioNLP Shared Task 2011 Workshop*, BioNLP Shared Task '11, pp. 183–191, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[BS13] Jari Björne and Tapio Salakoski. "TEES 2.1: Automated annotation scheme learning in the BioNLP 2013 shared task." In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pp. 16–25, 2013.

[BSN10] Florian Boudin, Lixin Shi, and Jian-Yun Nie. "Improving medical information retrieval with PICO element detection." In *Advances in Information Retrieval*, pp. 50–61. Springer, 2010.

[BV04] Chris Buckley and Ellen M Voorhees. "Retrieval evaluation with incomplete information." In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 25–32. ACM, 2004.

[BWC09] Thorsten Barnickel, Jason Weston, Ronan Collobert, Hans-Werner Mewes, and Volker Stümpflen. "Large scale application of neural network based semantic role labeling for automated relation extraction from biomedical texts." *PLoS One*, **4**(7):e6393, 2009.

[CBH01] Wendy W Chapman, Will Bridewell, Paul Hanbury, Gregory F Cooper, and Bruce G Buchanan. "A simple algorithm for identifying negated findings and diseases in discharge summaries." *Journal of biomedical informatics*, **34**(5):301–310, 2001.

[CC07] Grace Y Chung and Enrico Coiera. "A study of structured clinical abstracts and the semantic classification of sentences." In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pp. 121–128. Association for Computational Linguistics, 2007.

[CEZ97] James J Cimino, Gai Elhanan, and Qing Zeng. "Supporting infobuttons with terminological knowledge." In *Proceedings of the AMIA Annual Fall Symposium*, p. 528. American Medical Informatics Association, 1997.

[CFH14] Zhao Chen, Christine M Fillmore, Peter S Hammerman, Carla F Kim, and Kwok-Kin Wong. "Non-small-cell lung cancers: a heterogeneous set of diseases." *Nature Reviews Cancer*, **14**(8):535–546, 2014.

[CFZ13] G Chen, J Feng, C Zhou, Y-L Wu, X-Q Liu, C Wang, S Zhang, J Wang, S Zhou, S Ren, et al. "Quality of life (QoL) analyses from OPTIMAL (CTONG-0802), a phase III, randomised, open-label study of first-line erlotinib versus chemotherapy in patients with advanced EGFR mutation-positive non-small-cell lung cancer (NSCLC)." *Annals of oncology*, p. mdt012, 2013.

[Chu09] Grace Yuet-Chee Chung. "Towards identifying intervention arms in randomized controlled trials: Extracting coordinating constructions." *Journal of biomedical informatics*, **42**(5):790–800, 2009.

[CL13] Md Faisal Mahbub Chowdhury and Alberto Lavelli. "FBK-irst: A multi-phase kernel based approach for drug-drug interaction detection and classification that exploits linguistic information." *Atlanta, Georgia, USA*, **351**:53, 2013.

[CMB09] Carlos Cano, Thomas Monaghan, Armando Blanco, Dennis P Wall, and Leonid Peshkin. "Collaborative text-annotation resource for disease-centered relation extraction from biomedical text." *Journal of biomedical informatics*, **42**(5):967–977, 2009.

[CMS06] David Chen, Hans-Michael Müller, and Paul W Sternberg. "Automatic document classification of biological literature." *BMC bioinformatics*, **7**(1):1, 2006.

[CRY12] Sungbin Choi, Borim Ryu, Sooyoung Yoo, and Jinwook Choi. "Combining Relevancy and Methodological Quality into a Single Ranking for Evidence-based Medicine." *Inf. Sci.*, **214**:76–90, December 2012.

[CSE10] Janara Christensen, Stephen Soderland, Oren Etzioni, et al. "Semantic role labeling for open information extraction." In *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading*, pp. 52–60. Association for Computational Linguistics, 2010.

[CSG10] Adrien Coulet, Nigam H Shah, Yael Garten, Mark Musen, and Russ B Altman. "Using text to build semantic networks for pharmacogenomics." *Journal of biomedical informatics*, **43**(6):1009–1019, 2010.

[CWC13] Guoping Cai, Rebecca Wong, David Chhieng, Gillian H Levy, Scott N Gettinger, Roy S Herbst, Jonathan T Puchalski, Robert J Homer, and Pei Hui. "Identification of EGFR mutation, KRAS mutation, and ALK gene rearrangement in cytological specimens of primary and metastatic lung adenocarcinoma." *Cancer cytopathology*, **121**(9):500–507, 2013.

[DCK08] Berry De Bruijn, Simona Carini, Svetlana Kiritchenko, Joel Martin, and Ida Sim. "Automated information extraction of key trial design elements from clinical trial publications." In *AMIA Annual Symposium Proceedings*, volume 2008, p. 141. American Medical Informatics Association, 2008.

[DL07] Dina Demner-Fushman and Jimmy Lin. "Answering clinical questions with knowledge-based and statistical techniques." *Computational Linguistics*, **33**(1):63–103, 2007.

[DLW13] Shujun Dai, Youru Liu, Lin Wang, Baohui Han, and Liyan Jiang. "Analysis of Prognostic Factors in Patients with Stage IV Lung Adenocarcinoma after Failure of Second-line EGFR-TKIs Therapy." *Chinese Journal of Lung Cancer*, **16**(1), 2013.

[DMD03] Ian Donaldson, Joel Martin, Berry De Bruijn, Cheryl Wolting, Vicki Lay, Brigitte Tuekam, Shudong Zhang, Berivan Baskin, Gary D Bader, Katerina Michalickova, et al. "PreBIND and Textomy–mining the biomedical literature for protein-protein interactions using a support vector machine." *BMC bioinformatics*, **4**(1):11, 2003.

[Don06] Kevin Donnelly. "SNOMED-CT: The advanced terminology and coding system for eHealth." *Studies in health technology and informatics*, **121**:279, 2006.

[DP09] Centers for Disease Control and Prevention. "NHIS - Adult Tobacco Use - Glossary." `http://www.cdc.gov/nchs/nhis/tobacco/tobacco_glossary.htm`, 2009. Accessed: 2015-09-18.

[DPS07] Martin Dawes, Pierre Pluye, Laura Shea, Roland Grad, Arlene Greenberg, and Jian-Yun Nie. "The identification of clinically important elements within medical journal abstracts: Patient_Population_Problem, Exposure_Intervention, Comparison, Outcome, Duration and Results (PECODR)." *Journal of Innovation in Health Informatics*, **15**(1):9–16, 2007.

[EKK05] Noemie Elhadad, M-Y Kan, Judith L Klavans, and KR McKeown. "Customization in a unified framework for summarizing medical literature." *Artificial intelligence in medicine*, **33**(2):179–198, 2005.

[ETB09] EA Eisenhauer, Patrick Therasse, Jan Bogaerts, LH Schwartz, D Sargent, Robert Ford, J Dancey, S Arbuck, S Gwyther, M Mooney, et al. "New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1)." *European journal of cancer*, **45**(2):228–247, 2009.

[Evi11] OCEBM Levels of Evidence Working Group et al. "The Oxford 2011 levels of evidence.", 2011.

[EZM13] Gillian Ellison, Guanshan Zhu, Alexandros Moulis, Simon Dearden, Georgina Speake, and Rose McCormack. "EGFR mutation testing in lung cancer: a review of available methods and their use for analysis of tumour tissue and cytology samples." *Journal of clinical pathology*, **66**(2):79–89, 2013.

[FBS10] Marcelo Fiszman, Bruce E Bray, Dongwook Shin, Halil Kilicoglu, Glen C Bennett, Olivier Bodenreider, and Thomas C Rindflesch. "Combining relevance assignment with quality of the evidence to support guideline development." *Studies in Health Technology and Informatics*, **160**(Pt 1):709–713, 2010.

[FDK09] Marcelo Fiszman, Dina Demner-Fushman, Halil Kilicoglu, and Thomas C Rindflesch. "Automatic summarization of MEDLINE citations for evidence-based medical treatment: A topic-oriented evaluation." *Journal of biomedical informatics*, **42**(5):801–813, 2009.

[FLD11] Yu-Ching Fang, Po-Ting Lai, Hong-Jie Dai, and Wen-Lian Hsu. "MeInfoText 2.0: gene methylation and cancer relation extraction from biomedical literature." *BMC bioinformatics*, **12**(1):471, 2011.

[Flo92] Valerie Florance. "Medical knowledge for clinical problem solving: a structural analysis of clinical questions." *Bulletin of the Medical Library Association*, **80**(2):140, 1992.

[FM86] John W Finney and Rudolf H Moos. "Matching patients with treatments: Conceptual and methodological issues." *Journal of studies on alcohol*, **47**(2):122–134, 1986.

[FPF13] Ondrej Fiala, Milos Pesek, Jindrich Finek, Lucie Benesova, Zbynek Bortlicek, and Marek Minarik. "Sequential treatment of advanced-stage lung adenocarcinoma harboring wild-type EGFR gene: second-line pemetrexed followed by third-line Erlotinib versus the reverse sequence." *Anticancer research*, **33**(8):3397–3402, 2013.

[FRB04] Sherrilynne S Fuller, Debra Revere, Paul F Bugni, and George M Martin. "A knowledgebase system to enhance scientific discovery: Telemakus." *Biomedical Digital Libraries*, **1**(1):2, 2004.

[FRK04] Marcelo Fiszman, Thomas C Rindflesch, and Halil Kilicoglu. "Abstraction summarization for managing the biomedical research literature." In *Proceedings of the HLT-NAACL workshop on computational lexical semantics*, pp. 76–83. Association for Computational Linguistics, 2004.

[GCZ12] J Gao, JQ Chen, L Zhang, and ZY Liang. "[Relationship between EGFR and KRAS mutations and prognosis in Chinese patients with non-small cell lung cancer: a mutation analysis with real-time polymerase chain reaction using scorpion amplification refractory mutation system]." *Zhonghua bing li xue za zhi Chinese journal of pathology*, **41**(10):652–656, 2012.

[Geo96] Stephen L George. "Reducing patient eligibility criteria in cancer clinical trials." *Journal of Clinical Oncology*, **14**(4):1364–1370, 1996.

[Giu93] Fausto Giunchiglia. "Contextual reasoning." *Epistemologia, special issue on I Linguaggi e le Macchine*, **16**:345–364, 1993.

[GLR06] Claudio Giuliano, Alberto Lavelli, and Lorenza Romano. "Exploiting shallow linguistic information for relation extraction from biomedical literature." In *EACL*, volume 18, pp. 401–408. Citeseer, 2006.

[GM02] Roxana Girju, Dan I Moldovan, et al. "Text Mining for Causal Relations." In *FLAIRS Conference*, pp. 360–364, 2002.

[GMB13] Marina Chiara Garassino, Olga Martelli, Massimo Broggini, Gabriella Farina, Silvio Veronese, Eliana Rulli, Filippo Bianchi, Anna Bettini, Flavia Longo, Luca Moscetti, et al. "Erlotinib versus docetaxel as second-line treatment of patients with advanced non-small-cell lung cancer and wild-type EGFR tumours (TAILOR): a randomised controlled trial." *The lancet oncology*, **14**(10):981–988, 2013.

[GMH14] Jean I Garcia-Gathright, Frank Meng, and Will Hsu. "UCLA at TREC 2014 Clinical Decision Support Track: Exploring Language Models, Query Expansion, and Boosting." In *Proceedings of the 2014 Text Retrieval Conference*. National Institute of Standards and Technology, 2014.

[GNN07] Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. "Semeval-2007 task 04: Classification of semantic relations between nominals." In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pp. 13–18. Association for Computational Linguistics, 2007.

[GSB12a] Martin Gerner, Farzaneh Sarafraz, Casey M Bergman, and Goran Nenadic. "BioContext: an integrated text mining system for large-scale extraction and contextualization of biomolecular events." *Bioinformatics*, **28**(16):2154–2161, 2012.

[GSB12b] Binod Gyawali, Thamar Solorio, and Yassine Benajiba. "Grading the Quality of Medical Evidence." In *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, BioNLP '12, pp. 176–184, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[HCS15] Zhe He, Simona Carini, Ida Sim, and Chunhua Weng. "Visual aggregate analysis of eligibility features of clinical trials." *Journal of biomedical informatics*, **54**:241–255, 2015.

[Hea92] Marti A Hearst. "Automatic acquisition of hyponyms from large text corpora." In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pp. 539–545. Association for Computational Linguistics, 1992.

[HG08] Julian PT Higgins, Sally Green, et al. *Cochrane handbook for systematic reviews of interventions*, volume 5. Wiley Online Library, 2008.

[HKC13] JT Hartmann, C Kollmannsberger, I Cascorbi, F Mayer, MM Schittenhelm, S Heeger, and C Bokemeyer. "A phase I pharmacokinetic study of matuzumab in combination with paclitaxel in patients with EGFR-expressing advanced non-small cell lung cancer." *Investigational new drugs*, **31**(3):661–668, 2013.

[HKK09] Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. "SemEval-2010 Task 8: Multi-way Classification of Semantic Relations Between Pairs of Nominals." In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, DEW '09, pp. 94–99, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[HLD06] Xiaoli Huang, Jimmy Lin, and Dina Demner-Fushman. "PICO as a Knowledge Representation for Clinical Questions." In *AMIA 2006 Symposium Proceedings*, pp. 359–363. Citeseer, 2006.

[HMW05] R Brian Haynes, K Ann McKibbon, Nancy L Wilczynski, Stephen D Walter, Stephen R Werre, and Hedges Team. "Optimal search strategies for retrieving scientifically strong studies of treatment from Medline: analytical survey." *BMJ (Clinical research ed.)*, **330**(7501):1179, May 2005.

[HR05] George Hripcsak and Adam S Rothschild. "Agreement, the f-measure, and reliability in information retrieval." *Journal of the American Medical Informatics Association*, **12**(3):296–298, 2005.

[HRB14] Tianyong Hao, Alexander Rusanov, Mary Regina Boland, and Chunhua Weng. "Clustering clinical trials with similar eligibility criteria features." *Journal of biomedical informatics*, **52**:112–120, 2014.

[HRH16] Zhe He, Patrick Ryan, Julia Hoxha, Shuang Wang, Simona Carini, Ida Sim, and Chunhua Weng. "Multivariate analysis of the population representativeness of related clinical studies." *Journal of biomedical informatics*, **60**:66–76, 2016.

[HS13] Thomas A Hensing and Ravi Salgia. "Molecular biomarkers for future screening of lung cancer." *Journal of surgical oncology*, **108**(5):327–333, 2013.

[IB11] Ashwin Ittoo and Gosse Bouma. "Extracting explicit and implicit causal relations from sparse, domain-specific texts." In *Natural Language Processing and Information Systems*, pp. 52–63. Springer, 2011.

[Ins15] National Cancer Institute. "Levels of Evidence: Adult and Pediatric Treatment Studies." `http://www.cancer.gov/publications/pdq/levels-evidence/treatment`, 2015. Accessed: 2015-09-18.

[JAE01] Peter Juni, Douglas G Altman, and Matthias Egger. "Assessing the quality of controlled clinical trials." *British Medical Journal*, **323**(7303):42, 2001.

[Joa02] Thorsten Joachims. *Learning to Classify Text Using Support Vector Machines*. Springer, Boston, 2002 edition edition, April 2002.

[JSC13] Melissa L Johnson, Camelia S Sima, Jamie Chaft, Paul K Paik, William Pao, Mark G Kris, Marc Ladanyi, and Gregory J Riely. "Association of KRAS and EGFR mutations with survival in patients with advanced lung adenocarcinomas." *Cancer*, **119**(2):356–362, 2013.

[KBC10] Svetlana Kiritchenko, Berry de Bruijn, Simona Carini, Joel Martin, and Ida Sim. "ExaCT: automatic extraction of clinical trial characteristics from journal publications." *BMC medical informatics and decision making*, **10**(1):56, 2010.

[KC14] Seunghee Kim and Jinwook Choi. "An SVM-based high-quality article classifier for systematic reviews." *Journal of biomedical informatics*, **47**:153–159, 2014.

[KCA15] Jan A Kors, Simon Clematide, Saber A Akhondi, Erik M van Mulligen, and Dietrich Rebholz-Schuhmann. "A multilingual gold-standard corpus for biomedical concept recognition: the Mantra GSC." *Journal of the American Medical Informatics Association*, p. ocv037, 2015.

[KDR09] Halil Kilicoglu, Dina Demner-Fushman, Thomas C. Rindflesch, Nancy L. Wilczynski, and R. Brian Haynes. "Towards automatic recognition of scientifically rigorous clinical research evidence." *Journal of the American Medical Informatics Association: JAMIA*, **16**(1):25–31, February 2009.

[KJB14] Mark G. Kris, Bruce E. Johnson, Lynne D. Berry, David J. Kwiatkowski, A. John Iafrate, Ignacio I. Wistuba, Marileila Varella-Garcia, Wilbur A. Franklin, Samuel L. Aronson, Pei-Fang Su, Yu Shyr, D. Ross Camidge, Lecia V. Sequist, Bonnie S. Glisson, Fadlo R. Khuri, Edward B. Garon, William Pao, Charles Rudin, Joan Schiller, Eric B. Haura, Mark Socinski, Keisuke Shirai, Heidi Chen, Giuseppe Giaccone, Marc Ladanyi, Kelly Kugler, John D. Minna, and Paul A. Bunn. "Using multiplexed assays of oncogenic drivers in lung cancers to select targeted drugs." *JAMA*, **311**(19):1998–2006, May 2014.

[KLR04] Krishna Kummamuru, Rohit Lotlikar, Shourya Roy, Karan Singal, and Raghu Krishnapuram. "A hierarchical monothetic document clustering algorithm for summarization and browsing search results." In *Proceedings of the 13th international conference on World Wide Web*, pp. 658–665. ACM, 2004.

[KMC11] Su N Kim, David Martinez, Lawrence Cavedon, and Lars Yencken. "Automatic classification of sentences to support Evidence Based Medicine." *BMC bioinformatics*, **12**(Suppl 2):S5, 2011.

[KOT08] Jin-Dong Kim, Tomoko Ohta, and Jun'ichi Tsujii. "Corpus annotation for mining biomedical events from literature." *BMC bioinformatics*, **9**(1):10, 2008.

[KRF11] Halil Kilicoglu, Graciela Rosemblat, Marcelo Fiszman, and Thomas C Rindflesch. "Constructing a semantic predication gold standard from the biomedical literature." *BMC bioinformatics*, **12**(1):486, 2011.

[Lan06] Curtis P Langlotz. "RadLex: A new method for indexing online educational materials 1." *Radiographics*, **26**(6):1595–1597, 2006.

[LCN13] Su Man Lee, Jin Eun Choi, Yeon Kyung Na, Eun Jin Lee, Won Kee Lee, Yi Young Choi, Ghil Suk Yoon, Hyo-Sung Jeon, Dong Sun Kim, and Jae Yong Park. "Genetic and epigenetic alterations of the LKB1 gene and their associations with mutations in TP53 and EGFR pathway genes in Korean non-small cell lung cancers." *Lung Cancer*, **81**(2):194–199, 2013.

[Len95] Douglas B Lenat. "CYC: A large-scale investment in knowledge infrastructure." *Communications of the ACM*, **38**(11):33–38, 1995.

[LFB13] V Ludovini, A Flacco, F Bianconi, M Ragusa, J Vannucci, G Bellezza, R Chiari, V Minotti, L Pistola, FR Tofanetti, et al. "Concomitant high gene copy number and protein overexpression of IGF1R and EGFR negatively affect disease-free survival of surgically resected non-small-cell-lung cancer patients." *Cancer chemotherapy and pharmacology*, **71**(3):671–680, 2013.

[Lin04] Chin-Yew Lin. "Rouge: A package for automatic evaluation of summaries." In *Text summarization branches out: Proceedings of the ACL-04 workshop*, volume 8, 2004.

[LJL11] Zhihui Luo, Stephen B Johnson, Albert M Lai, and Chunhua Weng. "Extracting temporal constraints from clinical research eligibility criteria using conditional random fields." In *AMIA annual symposium proceedings*, volume 2011, p. 843. American Medical Informatics Association, 2011.

[LK77] J Richard Landis and Gary G Koch. "The measurement of observer agreement for categorical data." *biometrics*, pp. 159–174, 1977.

[LMW13] Zhihui Luo, Riccardo Miotto, and Chunhua Weng. "A human–computer collaborative approach to identifying common data elements in clinical trial eligibility criteria." *Journal of biomedical informatics*, **46**(1):33–39, 2013.

[LS13] Anthony Landreth and Alcino J. Silva. "The Need for Research Maps to Navigate Published Work and Inform Experiment Planning." *Neuron*, **79**(3):411–415, July 2013.

[LSJ15] Jiao Li, Yueping Sun, R Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. "Annotating chemicals, diseases, and their interactions in biomedical literature." In *Proceedings of the fifth BioCreative challenge evaluation workshop, Sevilla, Spain*, 2015.

[Luh58] Hans Peter Luhn. "The automatic creation of literature abstracts." *IBM Journal of research and development*, **2**(2):159–165, 1958.

[MAG14] Hamid Mousavi, Maurizio Atzori, Shi Gao, and Carlo Zaniolo. "Text-mining, structured queries, and knowledge management on web document corpora." *ACM SIGMOD Record*, **43**(3):48–54, 2014.

[MB05] Raymond J Mooney and Razvan C Bunescu. "Subsequence kernels for relation extraction." In *Advances in neural information processing systems*, pp. 171–178, 2005.

[MBF14] Rashmi Mishra, Jiantao Bian, Marcelo Fiszman, Charlene R Weir, Siddhartha Jonnalagadda, Javed Mostafa, and Guilherme Del Fiol. "Text summarization in the biomedical domain: a systematic review of recent research." *Journal of biomedical informatics*, **52**:457–467, 2014.

[MC12] Dana Movshovitz-Attias and William W Cohen. "Bootstrapping biomedical ontologies for scientific text using nell." In *Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pp. 11–19. Association for Computational Linguistics, 2012.

[McC93] John McCarthy. "Notes on formalizing context." 1993.

[McD09] John H McDonald. *Handbook of biological statistics*, volume 2. Sparky House Publishing Baltimore, MD, 2009.

[MEG08] Laura Plaza Morales, Alberto Díaz Esteban, and Pablo Gervás. "Concept-graph based biomedical automatic summarization using ontologies." In *Proceedings of the 3rd textgraphs workshop on graph-based algorithms for natural language processing*, pp. 53–56. Association for Computational Linguistics, 2008.

[MJW13] Riccardo Miotto, Silis Jiang, and Chunhua Weng. "eTACTS: a method for dynamically filtering clinical trial search results." *Journal of biomedical informatics*, **46**(6):1060–1067, 2013.

[MMM97] Andrew Merlino, Daryl Morey, and Mark Maybury. "Broadcast news navigation using story segmentation." In *Proceedings of the fifth ACM international conference on Multimedia*, pp. 381–391. ACM, 1997.

[MOP13] Claudiu Mihăilă, Tomoko Ohta, Sampo Pyysalo, and Sophia Ananiadou. "BioCause: Annotating and analysing causality in the biomedical domain." *BMC bioinformatics*, **14**(1):2, 2013.

[MRS08] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. "Evaluation in information retrieval." *Introduction to information retrieval*, pp. 151–175, 2008.

[MS12] Diego Mollá and María Elena Santiago-Martínez. "Creation of a corpus for evidence based medicine summarisation." *The Australasian Medical Journal*, **5**(9):503–506, September 2012.

[MSA01] David Moher, Kenneth F Schulz, and Douglas G Altman. "The CONSORT statement: revised recommendations for improving the quality of reports of parallel group randomized trials." *BMC Medical Research Methodology*, **1**(1):2, 2001.

[MW13] Riccardo Miotto and Chunhua Weng. "Unsupervised mining of frequent tags for clinical eligibility text indexing." *Journal of biomedical informatics*, **46**(6):1145–1151, 2013.

[NB14] Victoria Nebot and Rafael Berlanga. "Exploiting semantic annotations for open information extraction: an experience in the biomedical domain." *Knowledge and information Systems*, **38**(2):365–389, 2014.

[NBK13] Claire Nédellec, Robert Bossy, Jin-Dong Kim, Jung-Jae Kim, Tomoko Ohta, Sampo Pyysalo, and Pierre Zweigenbaum. "Overview of BioNLP shared task 2013." In *Proceedings of the BioNLP Shared Task 2013 Workshop*, pp. 1–7, 2013.

[Nen05] Ani Nenkova. "Automatic text summarization of newswire: Lessons learned from the document understanding conference." In *AAAI*, volume 5, pp. 1436–1441, 2005.

[NG84] Joseph D Novak and D Bob Gowin. *Learning how to learn*. Cambridge University Press, 1984.

[NJH01] Stuart J Nelson, W Douglas Johnston, and Betsy L Humphreys. "Relationships in medical subject headings (MeSH)." In *Relationships in the organization of knowledge*, pp. 171–184. Springer, 2001.

[NMT15] Nhung TH Nguyen, Makoto Miwa, Yoshimasa Tsuruoka, Takashi Chikayama, and Satoshi Tojo. "Wide-coverage relation extraction from MEDLINE using deep syntax." *BMC bioinformatics*, **16**(1):107, 2015.

[Nov77] Joseph Donald Novak. "A theory of education." 1977.

[NZK11] Stuart J Nelson, Kelly Zeng, John Kilbourne, Tammy Powell, and Robin Moore. "Normalized names for clinical drugs: RxNorm at 6 years." *Journal of the American Medical Informatics Association*, **18**(4):441–448, 2011.

[NZR12] Feng Niu, Ce Zhang, Christopher Ré, and Jude W Shavlik. "DeepDive: Web-scale Knowledge-base Construction using Statistical Learning and Inference." *VLDS*, **12**:25–28, 2012.

[PGH07] Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Björne, Jorma Boberg, Jouni Järvinen, and Tapio Salakoski. "BioInfer: a corpus for information extraction in the biomedical domain." *BMC bioinformatics*, **8**(1):50, 2007.

[PKT06] Hyung Paek, Yacov Kogan, Prem Thomas, Seymour Codish, and Michael Krauthammer. "Shallow semantic parsing of randomized controlled trial reports." In *AMIA Annual Symposium Proceedings*, volume 2006, p. 604. American Medical Informatics Association, 2006.

[PNR05] Nalini Polavarapu, Shamkant B Navathe, Ramprasad Ramnarayanan, Abrar Ul Haque, Saurav Sahay, and Ying Liu. "Investigation into biomedical literature classification using support vector machines." In *Computational Systems Bioinformatics Conference, 2005. Proceedings. 2005 IEEE*, pp. 366–374. IEEE, 2005.

[PP08] Patrick Pantel and Marco Pennacchiotti. "Automatically harvesting and ontologizing semantic relations." *Ontology learning and population: Bridging the gap between text and knowledge*, pp. 171–198, 2008.

[Pro16] UpToDate Marketing Professional. "Evidence-Based Clinical Decision Support at the Point of Care | UpToDate." `http://www.uptodate.com/home`, 2016. Accessed: 2016-06-06.

[PSP06] Conrad Plake, Torsten Schiemann, Marcus Pankalla, Jörg Hakenberg, and Ulf Leser. "AliBaba: PubMed as a graph." *Bioinformatics*, **22**(19):2444–2445, 2006.

[PVG11] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. "Scikit-learn: Machine Learning in Python." *J. Mach. Learn. Res.*, **12**:2825–2830, November 2011.

[RF00] Dragomir R Radev and Weiguo Fan. "Automatic summarization of search engine hit lists." In *Proceedings of the ACL-2000 workshop on Recent advances in natural language processing and information retrieval: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 11*, pp. 99–109. Association for Computational Linguistics, 2000.

[RFL05] Thomas C Rindflesch, Marcelo Fiszman, and Bisharah Libbus. "Semantic interpretation for the biomedical research literature." In *Medical informatics*, pp. 399–422. Springer, 2005.

[RGH07] Angus Roberts, Robert Gaizauskas, Mark Hepple, Neil Davis, George Demetriou, Yikun Guo, Jay Subbarao Kola, Ian Roberts, Andrea Setzer, Archana Tapuria, et al. "The CLEF corpus: semantic annotation of clinical text." In *AMIA Annual Symposium Proceedings*, volume 2007, p. 625. American Medical Informatics Association, 2007.

[RH04] Barbara Rosario and Marti A Hearst. "Classifying semantic relations in bioscience texts." In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, p. 430. Association for Computational Linguistics, 2004.

[RH05] Barbara Rosario and Marti A Hearst. "Multi-way relation classification: application to protein-protein interactions." In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 732–739. Association for Computational Linguistics, 2005.

[RH10] Bryan Rink and Sanda Harabagiu. "UTD: Classifying Semantic Relations by Combining Lexical and Semantic Resources." In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, pp. 256–259, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[RJB00] Dragomir R Radev, Hongyan Jing, and Malgorzata Budzikowska. "Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies." In *Proceedings of the 2000 NAACL-ANLP Workshop on Automatic summarization*, pp. 21–30. Association for Computational Linguistics, 2000.

[RKF11] Thomas C Rindflesch, Halil Kilicoglu, Marcelo Fiszman, Graciela Rosemblat, Dongwook Shin, H Kilicoglu, M Fiszman, G Rosemblat, and D Shin. "Semantic MEDLINE: An advanced information management application for biomedicine." *Information Services & Use*, **31**(1-2):15–21, 2011.

[RLH03] Thomas C Rindflesch, Bisharah Libbus, Dimitar Hristovski, Alan R Aronson, and Halil Kilicoglu. "Semantic relations asserting the etiology of genetic diseases." In *AMIA Annual Symposium Proceedings*, volume 2003, p. 554. American Medical Informatics Association, 2003.

[RSD14] Kirk Roberts, Matthew Simpson, Dina Demner-Fushman, Ellen Voorhees, and William Hersh. "State-of-the-art in biomedical literature retrieval for clinical cases: a survey of the TREC 2014 CDS track." *Information Retrieval Journal*, pp. 1–36, 2014.

[RSS87] Glenn D Rennels, Edward H Shortliffe, Frank E Stockdale, and Perry L Miller. "A computational model of reasoning from the clinical literature." *Computer methods and programs in biomedicine*, **24**(2):139–149, 1987.

[RTS11] Anna L Rich, Laila J Tata, Rosamund A Stanley, Catherine M Free, Michael D Peake, David R Baldwin, and Richard B Hubbard. "Lung cancer in England: information from the National Lung Cancer Audit (LUCADA)." *Lung cancer*, **72**(1):16–22, 2011.

[RTW00] Thomas C Rindflesch, Lorraine Tanabe, John N Weinstein, Lawrence Hunter, et al. "EDGAR: extraction of drugs, genes and relations from the biomedical literature." In *Pac Symp Biocomput*, volume 5, pp. 514–25. World Scientific, 2000.

[RWN95] W Scott Richardson, Mark C Wilson, Jim Nishikawa, and Robert S Hayward. "The well-built clinical question: a key to evidence-based decisions." *Acp j club*, **123**(3):A12–3, 1995.

[SAB11] Rodney L Summerscales, Shlomo Argamon, Shangda Bai, Jordan Huperff, and Alan Schwartz. "Automatic summarization of results from clinical trials." In *Bioinformatics and Biomedicine (BIBM), 2011 IEEE International Conference on*, pp. 372–377. IEEE, 2011.

[SAM10] Kenneth F Schulz, Douglas G Altman, David Moher, et al. "CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials." *BMC medicine*, **8**(1):18, 2010.

[SB04] Luciano Serafini and Paolo Bouquet. "Comparing formal theories of context in AI." *Artificial intelligence*, **155**(1):41–67, 2004.

[SCH07] Nicholas Sioutos, Sherri de Coronado, Margaret W Haber, Frank W Hartel, Wen-Ling Shaiu, and Lawrence W Wright. "NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information." *Journal of biomedical informatics*, **40**(1):30–43, 2007.

[SCT10] Ida Sim, Simona Carini, Samson Tu, Rob Wynden, Brad H Pollock, Shamim A Mollah, Davera Gabriel, Herbert K Hagler, Richard H Scheuermann, Harold P Lehmann, et al. "The human studies database project: federating human studies design data using the ontology of clinical research." *AMIA Summits on Translational Science Proceedings*, **2010**:51, 2010.

[SCY14] Luca Soldaini, Arman Cohan, Andrew Yates, Nazli Goharian, and Ophir Frieder. "Query reformulation for clinical decision support search." In *Proceedings of the 2014 Text Retrieval Conference*. National Institute of Standards and Technology, 2014.

[SHG84] C Coscarelli Schag, Richard L Heinrich, and PA Ganz. "Karnofsky performance status revisited: reliability, validity, and guidelines." *Journal of Clinical Oncology*, **2**(3):187–193, 1984.

[SKP93] JB Sørensen, M Klee, T Palshof, and HH Hansen. "Performance status assessment in cancer patients. An inter-observer variability study." *British journal of cancer*, **67**(4):773, 1993.

[SKU13] Jos A Stigt, Ageeth J Knol, Steven M Uil, Harry JM Groen, et al. "Pyrosequencing analysis of EGFR and KRAS mutations in EUS and EBUS-derived cytologic samples of adenocarcinomas of the lung." *Journal of Thoracic Oncology*, **8**(8):1012–1018, 2013.

[SKW08] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. "Yago: A large ontology from wikipedia and wordnet." *Web Semantics: Science, Services and Agents on the World Wide Web*, **6**(3):203–217, 2008.

[SMH13] Isabel Segura Bedmar, Paloma Martínez, and María Herrero Zazo. "Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013)." Association for Computational Linguistics, 2013.

[SML09] Lee T Sam, Eneida A Mendonça, Jianrong Li, Judith Blake, Carol Friedman, and Yves A Lussier. "PhenoGO: an integrated resource for the multiscale mining of clinical and biological data." *Bmc Bioinformatics*, **10**(Suppl 2):S8, 2009.

[SMS11] Isabel Segura Bedmar, Paloma Martinez, and Daniel Sánchez Cisneros. "The 1st DDIExtraction-2011 challenge task: Extraction of Drug-Drug Interactions from biomedical texts." 2011.

[SMW07] Zhongmin Shi, Gabor Melli, Yang Wang, Yudong Liu, Baohua Gu, Mehdi M Kashani, Anoop Sarkar, and Fred Popowich. "Question answering summarization of multiple biomedical documents." In *Advances in Artificial Intelligence*, pp. 284–295. Springer, 2007.

[SN12] Ida Sim and Joyce C Niland. "Study Protocol Representation." In *Clinical Research Informatics*, pp. 155–174. Springer, 2012.

[Sow76] John F Sowa. "Conceptual graphs for a data base interface." *IBM Journal of Research and Development*, **20**(4):336–357, 1976.

[SPT12] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. "BRAT: a web-based tool for NLP-assisted text annotation." In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 102–107. Association for Computational Linguistics, 2012.

[SS08] Daniel A Schult and P Swart. "Exploring network structure, dynamics, and function using NetworkX." In *Proceedings of the 7th Python in Science Conferences (SciPy 2008)*, volume 2008, pp. 11–16, 2008.

[SVH14] Matthew S Simpson, Ellen M Voorhees, and William Hersh. "Overview of the trec 2014 clinical decision support track." Technical report, DTIC Document, 2014.

[TBN11] William D Travis, Elisabeth Brambilla, Masayuki Noguchi, Andrew G Nicholson, Kim Geisinger, Yasushi Yatabe, Charles A Powell, David Beer, Greg Riely, Kavita Garg, et al. "International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society: international multidisciplinary classification of lung adenocarcinoma: executive summary." *Proceedings of the American Thoracic Society*, **8**(5):381–385, 2011.

[TCR09] Samson W Tu, Simona Carini, Alan Rector, Peter Maccallum, Igor Toujilov, Steve Harris, and Ida Sim. "OCRe: an ontology of clinical research." In *11th International Protege Conference*, 2009.

[TCS03] Andy Trotti, A Dimitrios Colevas, Ann Setser, Valerie Rusch, David Jaques, Volker Budach, Corey Langer, Barbara Murphy, Richard Cumberlin, C Norman Coleman, et al. "CTCAE v3. 0: development of a comprehensive grading system for the adverse effects of cancer treatment." In *Seminars in radiation oncology*, volume 13, pp. 176–181. Elsevier, 2003.

[TCS07] Richard TH Tsai, Wen-Chi Chou, Ying-Shan Su, Yu-Chun Lin, Cheng-Lung Sung, Hong-Jie Dai, Irene TH Yeh, Wei Ku, Ting-Yi Sung, and Wen-Lian Hsu. "BIOSMILE: a semantic role labeling system for biomedical verbs using a maximum-entropy model with automatically generated template features." *BMC bioinformatics*, **8**(1):325, 2007.

[THT13] Maurine Tong, William Hsu, and Ricky K Taira. "A formal representation for numerical data presented in published clinical trial reports." In *MedInfo*, pp. 856–860, 2013.

[TIM09] Paul Thompson, Syed A Iqbal, John McNaught, and Sophia Ananiadou. "Construction of an annotated corpus to support biomedical information extraction." *BMC bioinformatics*, **10**(1):349, 2009.

[TNM11] Paul Thompson, Raheel Nawaz, John McNaught, and Sophia Ananiadou. "Enriching a biomedical event corpus with meta-knowledge annotation." *BMC bioinformatics*, **12**(1):393, 2011.

[TNS11] Philippe Thomas, Mariana Neves, Illés Solt, Domonkos Tikk, and Ulf Leser. "Relation extraction for drug-drug interactions using ensemble learning." *Training*, **4**(2,402):21–425, 2011.

[TPC11] Samson W Tu, Mor Peleg, Simona Carini, Michael Bobak, Jessica Ross, Daniel Rubin, and Ida Sim. "A practical method for transforming free-text eligibility criteria into computable criteria." *Journal of biomedical informatics*, **44**(2):239–250, 2011.

[TT12] Maurine Tong and Ricky K Taira. "Improving the accuracy of therapy descriptions in clinical trials using a bottom-up approach." In *AMIA Annual Symposium Proceedings*, volume 2012, p. 1393. American Medical Informatics Association, 2012.

[Tur96] Peter Turney. "The identification of context-sensitive features: A formal definition of context for concept learning." 1996.

[TXT05] Lorraine Tanabe, Natalie Xie, Lynne H Thom, Wayne Matten, and W John Wilbur. "GENETAG: a tagged corpus for gene/protein named entity recognition." *BMC bioinformatics*, **6**(Suppl 1):S3, 2005.

[VG05] Anthony J Viera, Joanne M Garrett, et al. "Understanding interobserver agreement: the kappa statistic." *Fam Med*, **37**(5):360–363, 2005.

[VH13] David K Vawdrey and George Hripcsak. "Publication bias in clinical trials of electronic health records." *Journal of biomedical informatics*, **46**(1):139–141, 2013.

[WEM92] Eckart Walther, Henrik Eriksson, and Mark A Musen. *Plug-and-Play: Construction of task-specific expert-system shells using sharable context ontologies.* Citeseer, 1992.

[WFH12] T Elizabeth Workman, Marcelo Fiszman, and John F Hurdle. "Text summarization as a decision support aid." *BMC medical informatics and decision making*, **12**(1):41, 2012.

[WHC13] Shang-Gin Wu, Fu-Chang Hu, Yih-Leong Chang, Yung-Chie Lee, Chong-Jen Yu, Yeun-Chung Chang, Jenn-Yu Wu, Jin-Yuan Shih, and Pan-Chyr Yang. "Frequent EGFR mutations in nonsmall cell lung cancer presenting with miliary intrapulmonary carcinomatosis." *European Respiratory Journal*, **41**(2):417–424, 2013.

[WSC04] Tuangthong Wattarujeekrit, Parantu K Shah, and Nigel Collier. "PASBio: predicate-argument structures for event extraction in molecular biology." *BMC bioinformatics*, **5**(1):155, 2004.

[WTL10] Byron C Wallace, Thomas A Trikalinos, Joseph Lau, Carla Brodley, and Christopher H Schmid. "Semi-automated screening of biomedical citations for systematic reviews." *BMC bioinformatics*, **11**(1):1, 2010.

[WWL11] Chunhua Weng, Xiaoying Wu, Zhihui Luo, Mary Regina Boland, Dimitri Theodoratos, and Stephen B. Johnson. "EliXR: an approach to eligibility criteria extraction and representation." **18 Suppl 1**:i116–124, 2011.

[XSH06] Rong Xu, Kaustubh Supekar, Yang Huang, Amar Das, and Alan Garber. "Combining text classification and hidden markov modeling techniques for structuring randomized clinical trial abstracts." In *AMIA Annual Symposium Proceedings*, volume 2006, p. 824. American Medical Informatics Association, 2006.

[XW13] Rong Xu and QuanQiu Wang. "Large-scale extraction of accurate drug-disease treatment pairs from biomedical literature for drug repurposing." *BMC bioinformatics*, **14**(1):181, 2013.

[YCD08] Wei Yu, Melinda Clyne, Siobhan M Dolan, Ajay Yesupriya, Anja Wulf, Tiebin Liu, Muin J Khoury, and Marta Gwinn. "GAPscreener: an automatic tool for screening human genetic association literature in PubMed using the support vector machine technique." *BMC bioinformatics*, **9**(1):205, 2008.

[YHS13] James Chih-Hsin Yang, Vera Hirsh, Martin Schuler, Nobuyuki Yamamoto, Kenneth J O'Byrne, Tony SK Mok, Victoria Zazulina, Mehdi Shahidi, Juliane Lungershausen, Dan Massey, et al. "Symptom control and quality of life in LUX-Lung 3: a phase III study of afatinib or cisplatin/pemetrexed in patients with advanced lung adenocarcinoma with EGFR mutations." *Journal of Clinical Oncology*, pp. JCO–2012, 2013.

[YP05] Meliha Yetisgen-Yildiz and Wanda Pratt. "The effect of feature representation on MEDLINE document classification." In *AMIA*, 2005.

[YRH11] Katsumasa Yoshikawa, Sebastian Riedel, Tsutomu Hirao, Masayuki Asahara, and Yuji Matsumoto. "Coreference based event-argument relation extraction on biomedical text." *J. Biomedical Semantics*, **2**(S-5):S6, 2011.

[YXT10] Lanlan Yin, Guixian Xu, Manabu Torii, Zhendong Niu, Jose M Maisog, Cathy Wu, Zhangzhi Hu, and Hongfang Liu. "Document classification for mining host pathogen protein–protein interactions." *Artificial intelligence in medicine*, **49**(3):155–160, 2010.

[ZE99] Oren Zamir and Oren Etzioni. "Grouper: a dynamic clustering interface to Web search results." *Computer Networks*, **31**(11):1361–1374, 1999.

[ZFS11] Han Zhang, Marcelo Fiszman, Dongwook Shin, Christopher M Miller, Graciela Rosemblat, and Thomas C Rindflesch. "Degree centrality for semantic abstraction summarization of therapeutic studies." *Journal of biomedical informatics*, **44**(5):830–838, 2011.

[ZFS13] Han Zhang, Marcelo Fiszman, Dongwook Shin, Bartlomiej Wilkowski, and Thomas C Rindflesch. "Clustering cliques for graph-based summarization of the biomedical research literature." *BMC bioinformatics*, **14**(1):182, 2013.

[ZMO13] Wei Zhang, Elizabeth B McQuitty, Randall Olsen, Hongxin Fan, Heather Hendrickson, Fermin O Tio, Keith Newton, Philip T Cagle, and Jaishree Jagirdar. "EGFR mutations in US Hispanic versus non-Hispanic white patients with lung adenocarcinoma." *Archives of Pathology and Laboratory Medicine*, **138**(4):543–545, 2013.

[ZZZ13] Jing Zhao, Jinyin Zhao, Xiao Zhao, Weijun Chen, Wei Zhong, Li Zhang, Longyun Li, and Mengzhao Wang. "Detection of EGFR Gene Mutations in 100 Non-small Cell Lung Cancer Clinical Samples by a Real-time Polymerase Chain Reaction Method Using Amplification Refractory Mutation System Specific Primers and Taqman Fluorescence Probes." *Chinese Journal of Lung Cancer*, **16**(1), 2013.