A MACHINE LEARNING-BASED INFORMATION RETRIEVAL FRAMEWORK FOR MOLECULAR MEDICINE PREDICTIVE MODELS

By

Firas Hazem Wehbe

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Biomedical Informatics

May, 2011

Nashville, Tennessee

Approved:

Professor Cynthia S. Gadd

Professor Constantin F. Aliferis

Professor Steven H. Brown

Professor Pierre P. Massion

Professor Daniel R. Masys

Professor Hua Xu

DEDICATION

To my beautiful wife Maggie,

our unborn son and his future siblings,

my parents Hazem and Fadya,

and my brother and best friend

Bashar and his wife Madiha.

ACKNOWLEDGEMENTS

This dissertation would not have been possible without my advisors: Dr. Gadd's invaluable weekly guidance impacted almost every element of this work. Her honest support and encouragement helped me persevere and finish the last phases of my long PhD education. Ten years (and one month) ago, I was in the office of Dr. Aliferis, then director of graduate studies, making my case for joining the DBMI graduate program. Ever since, he has influenced my life irreversibly and taught me valuable lessons as a teacher, research mentor, passionate scientist, and friend.

I also want to thank the members of my PhD committee (in alphabetical order): Dr. Brown taught me a lot about the practice of informatics. His influence goes back to my very early years in the department when he gave me my first reading assignment as a graduate student. Little did I know then how relevant Ogden's semiotic triangle would be to my dissertation. It has been a privilege to benefit from Dr. Masys' vast informatics experience and insight both during his guidance of my PhD dissertation and while working with him on the CCASANET project. Dr. Massion was always ready to provide his perspective for this work both as a clinician and a researcher. I want to thank him for his valuable help and time with article annotation and for allowing me to present and discuss this project with his research team. Dr. Xu shared his valuable practical experience with me. His regular advice and guidance helped shape many of the applied experimental aspects of this research.

I am grateful to Dr. Yin Aphinyanaphongs for sharing his feature extraction and machine learning code which facilitated the implementation of many of the experiments

conducted for this work. Similarly, Dr. Josh Denny and Ms. Lisa Bastarache provided valuable practical assistance when using KnowledgeMap for feature extraction. For many years, Dr. Lawrence Fu has been a great friend (and exemplary roommate). He provided practical advice on text categorization and helped with the pilot article annotations. I also want to thank my friend Dr. Fouad Boulos for providing the breast cancer journal set and for helping with the pilot article annotations.

Independent expert annotations, a crucial part of this work, were made possible by financial support from the Vanderbilt Graduate School (Dissertation Enhancement Grant) and the Vanderbilt Institute for Clinical and Translational Research (VICTR voucher #VR1017). I want to thank the anonymous expert annotators for their valuable time and feedback.

During my many years in the DBMI graduate program, Ms. Rischelle Jenkins has been an infinite source of cheer and help for me and for many other students. Thank you Rischelle!

On a personal note, I want to thank my wife Maggie for her love and support that carried me through this journey. She is my best friend and the love of my life. I am also eternally indebted to my parents who endured so much and were willing to sacrifice everything during the dark days of the Lebanese civil war to provide the best possible education for me and my brother. Bashar, Madiha, Rami, Ramzi, Susan, Tom, Christiane, and Jeremiah – thank you for your cheerleading and for reminding me of the big picture and how great life can be beyond grad school.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Clinical bioinformatics research relies on molecular biology techniques to inform the clinical management of individual patients. Computational models can predict clinical outcomes, such as prognosis or response to treatment, based on the results of high throughput molecular assays. The large number of such models reported in the literature is growing at a pace that overwhelms the human ability to manually assimilate this information. For advances in molecular medicine to translate into clinical results, clinicians and translational researchers need to have up-to-date access to high-quality predictive models. Therefore the important problem of retrieving and organizing the vast amount of published information within this domain needs to be addressed. The inherent complexity of this domain and the fast pace of scientific discovery make this problem particularly challenging.

In this dissertation, I will discuss the limitations of existing tools for solving this problem and propose an information retrieval framework for organizing and retrieving clinical bioinformatics predictive models. This framework will need to adequately represent the complex attributes that characterize bioinformatics predictive models such as their purpose, their underlying methodology and source of data. A knowledgebase in which these models are stored and their attributes indexed for effective retrieval will be an integral component of this information retrieval framework. It is also self-evident that information within this knowledgebase has to be up-to-date and comprehensive. The

methods used to populate this repository will need to be scalable to the pace and volume of scientific discovery in this field.

Based on these observations, the specific aims of this dissertation are:

1. Propose and validate an information retrieval framework based on a semantic analysis of clinical bioinformatics.
2. Build and evaluate reproducible scalable automated filters for identifying relevant clinical bioinformatics papers using the MEDLINE database.
3. Build and evaluate reproducible scalable automated or semi-automated methods for annotating and indexing relevant papers for the supporting knowledgebase.

Aim 1

The first aim of this dissertation is to ***propose and validate an information retrieval framework based on a semantic analysis of clinical bioinformatics.*** Formal knowledge representation of the domain is needed to support the underlying computation. It will also inform the design of the overall information retrieval framework.

1a. Identify types of information that are relevant to clinical bioinformatics predictive models. What objects are related to predictive models in this domain? What is the information needed to annotate models and related objects to support their effective retrieval? For example, the development and validation of predictive models requires datasets of molecular patient data and associated known clinical outcomes. To support queries for models in this domain, it will be useful to collect information on the source of molecular data used by specific predictive models and on the type of clinical outcomes that these models predict.

1b. Construct and validate a semantic model (ontology). The datasets described above are essential for training and/or testing of predictive models and for determining their scope and quality. Therefore, a dataset is an object that is functionally relevant to a clinical predictive model. The first step is to define, based on use cases, an ontology of objects and relationships in the domain of clinical bioinformatics that are relevant to the proposed information retrieval framework. This ontology's expressiveness for the domain of clinical bioinformatics will be analyzed.

1c. Design an information retrieval framework based on the semantic model. The main purpose of the information retrieval framework is to retrieve models and related objects in response to clinical bioinformatics queries. A set of attributes can be used to characterize the objects, and the objects can then be annotated by assigning values to their attributes. Queries specify the values of these attributes and the objects that match query predicates will be retrieved. The classes of attributes that will be chosen to annotate objects in the knowledgebase will be based on the semantic model (ontology) in Aim 1b. The choice of these attributes, i.e. annotation scheme, needs to balance expressiveness - ability to correctly represent the domain – with support for efficient indexing and retrieval.

Aim 2

The second aim of this dissertation is to ***build and evaluate reproducible scalable automated methods for identifying relevant clinical bioinformatics papers using the MEDLINE database.*** Published articles are the primary source of information about biomedical research. Building a knowledgebase supporting the desired information

retrieval framework requires a set of papers that describe clinical bioinformatics predictive models. Manually identifying these relevant papers from the literature from the large number of related articles is a tedious task. The scalability of building and maintaining an up-to-date collection of relevant papers can be achieved via automated filters. Within this framework, filters are text classifiers that flag relevant papers based on the text content of their MEDLINE record. Statistical machine learning models have been shown in the past to reliably replicate human classification tasks for MEDLINE article retrieval. The contents of the MEDLINE record is converted into numerical features that can be used by machine learning to compute paper classification. During the feature extraction step, tokenization breaks the stream of text into words and/or symbols. In addition, feature extraction may rely on linguistic and semantic transformations such as shallow parsing, word stemming or stop-word removal. Other types of feature extraction steps exist and may depend, for example, on the location of terms within the MEDLINE record (title, abstract, MeSH term, etc.)

2a. The first research question that will be investigated is: ***can existing or modified feature extraction transformations be used to train machine learning filters that can identify relevant papers from MEDLINE?*** The machine learning filters will be Support Vector Machine (SVM) based. The performance of these filters will be measured by comparison against a ***gold standard of labeled articles from the domains of lung cancer and bioinformatics*** using area under the receiver operator curve (AUC). This task requires prior reliable manual assignment of labels to the requisite gold standard datasets. The training and gold standard datasets will need to include a mixture of labels that is similar in composition to the results of routine MEDLINE queries in this domain. The

goal is to find, from the different permutations of existing feature extraction transformations, the feature extraction steps that will produce machine learning models with the highest possible AUC cross validation performance in the lung cancer dataset.

2b. If the performance of machine learning filters is sensitive to the domain (disease) of the papers in a dataset, then new filters will need to be trained using gold standards built for all possible diseases, a tedious task. Therefore it is important to find among the models that satisfy the previous hypothesis, those that can successfully filter relevant papers that describe other diseases. Specifically we want to show that it is possible to train such filters. Therefore, the second research question will investigate ***whether there exist from among the filters with favorable performance on the lung cancer gold standard, filters that identify relevant papers in other domains – specifically breast cancer.*** This step will test the generalizability of the clinical bioinformatics filters across multiple medical specialties. The performance of the filters will be benchmarked using, as ***gold standard, a dataset derived from MEDLINE articles from the domain of breast cancer***.

2c. The training and testing datasets used for testing the hypotheses above are based on one annotator's attempt to consistently apply labels about the relevance of these papers to clinical bioinformatics. Even if these hypotheses are true, one cannot infer that the performance of the filters will generalize to annotations judged by other domain experts. It will important to assess generalizability along a different dimension: annotation by a different set of experts. Therefore the third research question is: ***can filters trained to identify relevant papers in the domains of bioinformatics and lung***

cancer using annotation by one expert, identify relevant papers in the same domain as judged by other experts whose annotations were not used to train those filters?

Aim 3

Build and evaluate reproducible scalable automated or semi-automated methods for annotating and indexing papers for the supporting knowledgebase. The next step in building the supporting knowledgebase is the semantic annotation of relevant papers according to the set of attributes defined as part of Aim 1c. The semantic annotation of papers is more complex than assigning a binary "relevant" label as in Aim 2. Annotation requires the identification of more granular attributes of a given paper such as determining the biological source of data used in a model and the type of molecular assays used. Furthermore, the eligibility of a given article for semantic annotation may depend on more than one conditional "relevance" criteria. For example, there may be a predictive model that relies on molecular data; however, the outcome that this model predicts may not conform to the semantic definition of clinical outcome as defined in this information retrieval framework. In this case this model will only qualify for a subset of the annotations. The approach followed for automatic semantic annotation in this aim will be similar to the approach used in Aim 2. The problem will be cast as a machine learning classification problem. Due to the semantic complexity of the attributes used for annotation, there may be a need to expand the feature extraction steps to include Natural Language Processing (NLP) techniques.

3a. The first research question is: ***can existing or modified feature transformations be used to train text classifiers that can replicate human semantic***

annotation of the gold standard? The performance of the classifiers will be measured using the average AUC, and will be evaluated using N-fold cross validation. The aim is to identify and select methods that achieve high (AUC) against the gold standard in a given domain (lung cancer and bioinformatics). This hypothesis will be tested independently for all the attributes used for semantic annotation i.e. unique classifiers will be trained and tested for every attribute used for semantic annotation. There is no a priori assurance that the performance of the classifiers will be uniform across all attributes. If the performance of the text classifier is not uniform, then the causes of this variability will be investigated including the inherent suitability of the individual semantic attribute for automatic annotation.

3b. The second research question will test ***whether modifying the feature extraction transformations used for training semantic classifiers in Aim 3a to include natural language processing (NLP) will alter their performance.*** Specifically, this step will measure the effect of adding the frequency of occurrence of unique UMLS concepts (CUI) within the MEDLINE record to the set of features used to train and test the machine learning dataset. KnowledgeMap, a natural language processing tool that can extract UMLS CUIs in biomedical text, will be used. Similar to Aim 3a, this will be tested for all the annotation attributes, and the cause of variation in performance across the different attributes, if present, will be investigated. Some annotation attributes (e.g. clinical outcome) may be related on the occurrence of CUIs present in traditionally epidemiological text whereas other attributes (e.g. those relating to molecular data) may be related to the occurrence of CUIs that stem from molecular named entities with different coverage in the UMLS.

3c. The third research question is: ***can text classifiers trained for semantic annotation of relevant papers in the domains of bioinformatics and lung cancer using annotation by one expert, replicate the semantic annotation of independent papers in the same domain by other experts?*** Similar to Aim 2c, this will test the generalizability of the classifiers to annotations by different annotators. Similar to Aims 3a and 3b, this will be repeated for all the annotation attributes. If variability was found in the performance of the classifiers corresponding to different attributes, the causes of such variability will be investigated.

CHAPTER II

BACKGROUND AND SIGNIFICANCE

Clinical Bioinformatics

The goal of molecular medicine is to diagnose and find treatments for human diseases by the application of tools of molecular and cell biology[1]. In recent years, researchers have begun to link tissue molecular profiles – such as gene expression information – of individual patients to relevant disease outcomes such as diagnosis[2], prognosis[3], and response to treatment[4]. Knowledge discovered from large scale genomic and molecular biology data is already being put to clinical use[5] and several clinical studies are in the development or validation phase[6]. In a typical scenario, a molecular assay is performed on tissue obtained from a patient. Then, a decision model computes, based on the assay results, the predicted clinical outcome of the patient's disease. Therefore, clinical bioinformatics predictive models rely on molecular and clinical data obtained from a single patient to compute a "decision" or outcome that is used for the clinical management of the patient, for example to help determine the choice of effective therapy. In February 2007, the U.S. Food and Drug Administration approved[7] the first molecular test, MammaPrint™, to predict the recurrence of breast cancer within five to ten years. MammaPrint™[8] and other genomic profiling tests like Oncotype Dx[9] compute clinical outcomes using assay measurements from multiple genes (70 for MammaPrint™ and 21 for Oncotype Dx). Clinical bioinformatics models can be classified based on the type of molecular information used as input. (1) "Genomic" tests measure the in-vivo

activity of a complex set of genes in diseased tissue. (2) "Genetic" tests look at inherited genetic characteristics that are passed from one generation to another. Inherited genetic traits may predispose to certain diseases or may affect an individual's response to pharmacologic therapy. (3) "Proteomic" tests measure the local signal of a set of proteins, the end-product of complex genomic interactions. Clinical bioinformatics models can also be classified based on the type of the clinical outcome that they compute. The type of clinical outcome and its relation to the personalized clinical management of patients has policy and regulatory implications. For example, prognostic and diagnostic genomic kits have been regulated as class II devices by the FDA, requiring less oversight[10]. On the other hand, the FDA requires that the related genotypes and dosing guideline information be included in drugs where genomic correlation is known to affect treatment outcome[11]. The field of pharmacogenomics[12,13] applies whole genome analysis technologies to predict treatment response and adverse drug reaction susceptibility based on individual genetic variability. For example the cancer drug, irinotecan, has side effects that have been linked to an inherited allele that leads to lower expression of a specific drug-clearing enzyme[14]. A listing of drug-related genomic biomarkers is available on the FDA website[11].

Building and validating clinical bioinformatics models is a complicated scientific process that draws from multiple overlapping sources of genetic, genomic, or proteomic data. High throughput experimental methods generate data that can have hundreds or even hundreds of thousands of data-points per sample. Such data are difficult to process manually and require sophisticated computations that draw from a variety of disciplines including biostatistics and machine learning. Furthermore, there is great variability in the

methods that evaluate these predictive models' validity, generalizability, and supporting evidence[6]. Many analyses of statistical shortcomings of current approaches for building and validating predictive models have been published[3,15,16]. Clinical bioinformatics is a complex domain, and there is a clear need to organize the vast amount of information surrounding clinical predictive models and related research information. Due to the fast pace of scientific discovery, there is also a need for tools that provide up-to-date information to clinicians and researchers in this domain.

Related and Similar Resources

Current resources and databases address some of the clinical bioinformatics information needs and can be leveraged when building a system for organizing and retrieving clinical predictive models. Most existing resources store only specialized subsets of predictive models. For example, PharmGKB[17-19] is a database that links genomic variability, mostly accounted for by single nucleotide polymorphisms (SNPs), with phenotypes relating to pharmacokinetics, pharmacodynamics, or therapeutic clinical outcomes. Information is organized in PharmGKB by gene, drug, disease, publications, or datasets. ONCOMINE[20,21], a database and web-based analysis and visualization tools, is restricted to cancer-related gene expression microarray experimental results. Datasets in Oncomine are profiled (annotated) by cancer and tissue types, by experimental methods, and by the types of gene expression differential analysis performed on these datasets, e.g. comparing gene expression differentials across different prognosis groups or across different histological subtypes. Oncomine provides links to the original datasets as well as analysis tools for (clinical) differential analysis of these datasets, but does not

store or classify the applied algorithms or inferred models that were reported in the original publications. The Gene Expression Omnibus (GEO)[22,23], is a resource developed by the NCBI as a MeSH-indexed public repository of microarray and other forms of high-throughput "omics" data submitted by the scientific community. Sources of data in GEO include gene expression microarrays, ArrayCGH, SNP Arrays, Serial Analysis of Gene Expression (SAGE), Massively Parallel Signature Sequencing (MPSS), protein arrays, and mass spectrometry. Information in GEO is organized by series (study-centered data) or by individual genes. Many journals require that gene expression results be submitted using the Minimum Information About a Microarray Experiment (MIAME) format[24] to the GEO prior to publication[25]. Some of the series in GEO are further curated and stored as datasets with more structured annotations (relevant citations, organisms) and the possibility to perform online data analysis. The Biometric Research Branch at the NCI has developed array analysis tools for gene expression data, and provides a hand-curated archive of human cancer gene expression datasets[26]. The Rembrandt repository[27] is highly annotated for clinically-oriented outcomes but is restricted to brain-cancer-related molecular research data. Recently, GeneSigDB – a curated database of gene expression signatures – was revealed[28]. Articles that describe gene expression signatures are manually curated and the gene lists that are reported in the articles are manually transcribed and mapped to standardized gene identifiers in other databases. The emphasis in GeneSigDB is on extracting genes from published papers and indexing these genes across different publications and gene signatures.

Formal Ontologies in Support of Translational Bioinformatics

Ontologies are formal knowledge representations that are computable. Non-vague medical concepts are the building blocks of biomedical ontologies[29]. Ontologies organize concepts in hierarchies and describe relationships that can exist between them. In recent work, Shah, Butte, Musen et al[30], have used ontology-driven indexing of public molecular datasets to support translational bioinformatics. They proposed and evaluated a method to map free text, the predominant method of dataset annotation, to concepts in ontologies such as SNOMED-CT[31] and the NCI Thesaurus[32]. This approach drives the integration of information across various objects that pertain to translational bioinformatics. They aim to index publically available molecular datasets – such as Gene expression datasets, Tissue Micro Array datasets – as well as citations in Pubmed using ontology concepts extracted from free text. Ontology-driven indexing will serve as the basis for a prototype system for information integration that they will offer online along with other tools, such as the "Open Biomedical Annotator"[33], via the National Center for Biomedical Ontology[34]. Ontology-based approaches have been used by genomic and proteomic researchers to ask questions of diverse data repositories. Such cross-database information queries benefit from standard and controlled representation of domain knowledge[35,36]. By standardizing and controlling domain concepts, ontologies such as the NCI Thesaurus[32], the Gene Ontology (GO)[37] and the Clinical Bioinformatics Ontology (REFSEQ)[38] support interoperability between clinical bioinformatics repositories. Other ontology-based frameworks, include the RAD/RAPAD Study Annotator[39], the Functional Genomics Experiment Model[40,41], and the Ontofusion system for biomedical database integration[42,43]. Description logic(DL)-based languages[44], such as the Web Ontology

Language (OWL)[45] are popular means of formal ontology representation. DL is a subset of first order logic that is constrained to guarantee decidable computation when it comes to classification of objects into classes. DLs can be used for conceptual modeling, information integration, and support for semantic query mechanisms. SNOMED-CT is an example of a DL-based ontology. Formal ontologies define unambiguous concepts and provide along with the concepts their semantic types, synonymy information, tree hierarchies and relationships. These attributes make formal ontologies, in theory, the ideal choice for the annotation of resources to support information retrieval.

Manual Annotation: Using Humans to Create Structure

In this proposal, I will use the term "annotation" to refer to the process of mapping a string of unstructured text to a collection of concepts in a terminology or ontology. Unstructured text is annotated by coupling ontology concepts either to the entire text or to specific subsets of the text. An application of the former approach is mapping an entire dataset in GEO based on the textual description that accompanies it to a disease concept in SNOMET-CT[30]. Applications of the latter approach[46-49] identify and highlight every mention of GO concepts within the text of published papers. When building systems that rely on annotation of free text, the choice of a specific annotation scheme (ontology, terminology, or a simple controlled list) is a crucial operational decision. Another important operational decision is the choice of the method used for annotation. In this section, I will discuss the annotation that is entirely performed by humans.

Manual Annotation Using Formal Ontologies

Formal ontologies are often the first choice due to their desirable properties mentioned above. Manual annotation with ontology concepts by domain experts may provide the highest possible quality; however, researchers have argued[50,51] that many problems are associated with this approach, namely:

1. Evolution: Ontologies change over time. Concepts may be added, removed, replaced, or moved around in the graph. While "graceful evolution"[29] (See Cimino's terminology disederata) can help achieve backward compatibility, completeness cannot be assured in the future unless human annotators go back and re-annotate according to the newer versions of the ontology. New concepts can be missed in the indexing of old documents.
2. Inter-annotator agreement: Manual curation is subjective and may depend on the scientific background, or expertise of the annotators as well as on the process of annotation itself. Research has shown significant variability between annotators[47,52,53]. The variability may arise from the complexity of the annotation scheme, from the ambiguity in annotation instructions, or from intrinsic difference between the individual annotators (education, professional training, annotation experience etc.)
3. Scalability: Human effort cannot keep pace with scientific progress and the volume of new published knowledge.

Despite these drawbacks, ontologies are still the preferred annotation and indexing scheme for many reference databases where the stored information is well structured and characterized[54]. GO and other bioinformatics ontologies are used for

indexing gene products and structural biology and proteomic databases. Specialized ontologies have also been useful for annotating molecular databases for flies and other species[55]. Theoretically, extensive ontologies like SNOMED-CT provide the flexibility not just to represent the stored information, but to also represent flexible query expressions making them appealing for information retrieval systems. However, increased expressiveness leads to increased computational complexity. Using DL databases like SNOMED-CT can be computationally impractical for certain tasks given the current state of DL reasoning technology[56-58].

Manual Annotation Using Ad hoc Schema

Some researchers have designed annotation schema that are tailored to the task at hand. Chapman et al[59] used grounded theory to design an annotation scheme for extracting clinical conditions from emergency department records. They started with a general theory statement and followed an iterative approach to refine the annotation schema. The schema was then validated and evaluated for completeness. Since there is no external reference to act as gold standard, Chapman et al relied on inter-annotator agreement as one measure of the quality of annotation outcome.

Quality of Manual Annotation: Inter-Annotator Agreement

Hripcsak studied inter-annotator agreement (IAA) in biomedical annotations[60,61] and found significant inter-annotation variability when using controlled terminologies to code problem lists[52]. A decade earlier Giuse and Miller reported similar variability[62]. Hripcsak analyzed different metrics that measure IAA, and concluded that the Kappa

concordance statistic is often not suitable to measure agreement when using biomedical ontologies for annotation. The theoretical reason for rejecting Kappa, the classical inter-rater agreement metric, in such cases is that annotators have to select from a large set of concepts. The statistical probability of them agreeing on the true negatives is high. Kappa will then be skewed towards over-optimistic results, and the false agreement rate will be hard to estimate[60]. Hripcsak proposed using another metric, the F-measure, which converges to Kappa when the number of concepts to be picked is small relative to the entire set of concepts in the terminology[61]. Variability in human cognition, knowledge, and understanding of the annotation process are possible explanations of IAA variability. Chapman, Dowling and Hripcsak studied annotation behavior in more detail[63]. They found that training annotators (physicians and non-physicians) to use the given annotation schema helped improve their agreement scores over time.

Despite IAA variability, manual annotations are generally considered as having higher quality yet lower yield than automated methods[50,51,64]. Some studies consider manual curation the gold standard without empirically reporting IAA[65]. For benchmarking purposes, gold standards can rely on expert consensus or on less subjective criteria such as choosing ACP journal club papers as the positive standard for machine learning filters that identify high quality papers in internal medicine[66].

Scalable Annotation

Clinical bioinformatics is characterized by a fast pace of scientific discovery. Manual annotation and curation of text in this domain may not succeed in building

comprehensive and up-to-date databases. Many approaches have been proposed as solutions for this scalability problem.

Crowdsourcing

Crowdsourcing is the act of dividing a large task into multiple small similar tasks and outsourcing those tasks to a large group of people. Even though it relies on manual curation by humans, crowdsourcing relies on technologies (so called "Web 2.0" techniques) that allow the harnessing of mass collaboration. Wikipedia is an example of a massive collaboration by a large number of people, each contributing a relatively small portion of the effort. Some investigators have attempted crowd-sourcing of specialized databases. For example, a protein interaction database, WikiProteins[67], was constructed using a structured semantic wiki and community contribution. Other examples include the BOWiki[68], and ArrayWiki[69]. While Wikipedia remains a success case, the above mentioned databases are relatively recent, and their long term success as comprehensive, up-to-date resources remains to be tested. (BOWiki for example, has not been updated in over a year). Crowdsourcing may be performed via an open call for voluntary participation from the community. For such project to be successful, multiple factors need to be in place, such as a perceived need by the community for the resource, and designing the small task units in a way that mirrors current work being done by the community members. Another crowdsourcing approach is by making micropayments for crowdsource "workers" as compensation for completed units of work. Amazon's™ Mechanical Turk is a resource that matches workers with repetitive so called Human Intelligence Tasks (HIT). Mechanical Turk "requesters" design the HITs and pay

"workers" a small fee for each completed HIT. For example, the Laboratory for Personalized Medicine (LPM) at Harvard Medical School posted hundreds of documents on Mechanical Turk and asked the "workers" to annotate those documents by answering a small set of questions for each document[70].

Assisted Curation

Assisted curation is the transfer from unstructured information (typically text) into structured information (typically databases or ontologies) by human curators, who are assisted by computational methods based on text-mining. Text mining can provide decision support for curators by highlighting semantics types, duplications, or relevant entities within the text. For example, Altman et al used PharmPresso, an adaptation of a text mining tool TextPresso[71], to assist the curation of PharmGKB[65]. PharmPresso extracts pharmacogenomic concepts and relationships from full article text (not just Abstract/MeSH terms) and highlights those concepts in the text of the paper for the human user. Jin et al automatically identified GO annotations in the literature using multi-label classification techniques that utilized the structure of the GO graph[72]. Another approach to assisted curation is using automated methods to label or classify entire documents. The most prominent example is the constant massive need to annotate the PubMed database using MeSH terms. The staff at the National Library of Medicine have been continuously improving the Medical Text Indexer (MTI), a tool that automatically recommends MeSH main headings to NLM indexers. The developers of TMI use NLP, statistical and machine learning based method for producing MeSH recommendations. These methods have been used and evaluated independently and

combined[73]. Aphinyanaphongs et al constructed and validated machine learning filters that identify PubMed papers describing high quality clinical evidence[66,74]. Haynes et al designed filters for high quality clinical studies in the literature by optimizing computerized combinations of search terms[75]. In certain situations, human annotation is intended as the first step. The outcome of the initial human annotation can then be used to train an artificial-intelligence curator that takes over the annotation process if it can replicate the quality of human annotation[76]. Currently, there is no standard approach for assisted curation of biomedical databases; similarly, there is no consensus yet among researchers on the ideal balance between the machine and the human roles in this process[50,51,77].

Leveraging Semantic Technology

Some specialized biomedical search engines like HubMed[78], iHOP[79,80], EBIMed[81], GOPubMed[82], AliBaba[83], TextPresso[71], augment traditional information retrieval frameworks with different semantic enhancement. Their methodologies can be summarized as follows:

- NLP-based techniques are used to identify named entities in the papers. NLP techniques include entity recognition by matching strings to concepts, enumeration of concept co-occurance, concept disambiguation using contextual information, and summarization. The semantic information extracted via NLP is then used to annotate and index the text database. Information extraction may include "mining" new relationships based on semantic types, such as disease-drug, and the co-occurrence of concepts in the text. Examples include: the

BioProspecting approach for mining genetic markers from the New England Journal of Medicine[84]; the Clinical E-Science Framework (CLEF) information extraction approach for building a semantically annotated corpus of clinical text[85,86]; and the PharmGKB approach for automatic identification of drug-genotype-phenotype relationships from the pharmacogenomics literature[65,87,88].

- NLP-based techniques parse semantic entities in queries. This is referred to as query transformation and refinement. The semantic entities obtained from the query are used to match objects in the database that are indexed by those concepts. This approach requires preexisting semantic annotation of the database as described in the previous point.
- The hierarchical structure of bioinformatics ontologies like GO can guide the users as they browse the database. Subsumption reasoning can be leveraged to help the users search the databases at different granularity levels.
- Information can be obtained based on hyperlinks from the database to external sources of knowledge like Wikipedia, or to structured databases like GenBank and GEO. Linked resources can provide additional information. For example TextPresso, and Pharmpresso extract and analyze full paper text from pdf files. Links to protein sequences and motifs can add additional information not provided in paper abstracts.
- Other preprocessing techniques include: highlighting text classified as "evidence" like highlighting words from the query in the results; ranking based on score functions of arbitrary complexity e.g. using Google's PageRank to analyze the

network structure of bibliome citations; interactive queries; pre-calculation and caching of semantic distances.

Semantic Annotation by Content Creators

Automated and semi-automated semantic analysis can benefit from semantic annotations of published papers and data by the authors themselves upon submission of their manuscripts. Proposed frameworks for this approach include: the Structured Digital Abstracts (SDAs) requirements by the Royal Chemical Society[89]; the FEBS letters experiment[90,91]; and SciXML / SciBORG[92]. Some have proposed requiring computationally guided annotation be an integral part of the editorial process[89,93].

Ontology Learning

Finally, text can be annotated using new ontologies constructed directly from text (rather than automatic annotation of text using existing ontology). This approach attempts the automatic discovery of terms, synonyms, concepts, and taxonomic and non-taxonomic relationships[94]. It may be difficult to automatically construct concept synonymy and hierarchical relationships based solely on statistical and natural language processing of biomedical text corpora due to the very large number of medical concepts and the extensive hierarchies and relationships between these concepts[95,96]. Furthermore, the nature of clinical medical knowledge is such that complex semantic manipulations of concepts (e.g. post-coordination) are required for adequate representation[97].

Evaluation of Annotation and the Inescapable "Gold Standard"

Regardless of the methodology used to achieve scalability of document annotation, there will be a need for an evaluation study of the annotation process. Furthermore, reliable manual (human) annotation of a set of papers will be integral to any evaluation of these methods as evident in all of the studies referenced above. Consider the broad categories of annotation methodologies:

- NLP-based information extraction (fully automated, or semi-automated): Such methods were typically validated using "gold standard" annotations of concepts within documents by human experts. In NLP-based methods, the F-measure[61] is typically used to measure agreement of the outcome of NLP with expert annotations, because the facts that are extracted from the text by NLP belong to a very large set of concepts. When NLP methods are utilized as classifiers of a binary outcome (i.e. to infer whether a label is present or absent), information retrieval metrics such as recall, precision and sensitivity are used to evaluate their performance[98].
- Statistical and machine learning based methods (fully automated label assignment, or semi automated use of filters that reduce the work of human indexers): Building and validating supervised machine learning models requires "gold standard" training and testing datasets compiled by human annotators. Recall and precision can also be calculated based on the gold standard and used as indicators of the performance of the resultant classifiers. However, the output of many machine learning models is a real number that can be used to generate a ranked result (as opposed to an unordered result set). The ordered ranking allows

for variation of the decision cut-off points for class assignment and the construction of a precision vs recall curve. The precision vs recall curve can be transformed into the receiver operator curve (ROC). The area under the ROC is a common metric for evaluating the performance of machine learning text classifiers[99].

- Manual Annotation by Humans: Biomedical databases can be completely annotated by trained curators, by crowdsourcing, or by community volunteers. As mentioned earlier, many researchers have pointed to inter-annotator variability in biomedical databases[52,53,60,62,63]. Higher variability may result when annotation relies on the users' conceptual model of the given domain and on their individual understanding of the meaning in unstructured text. The reliability of the gold standard can be empirically evaluated by assigning an overlapping subset of documents to multiple annotators and measuring a concordance metric (e.g. Kappa) of their label assignment.

Building methodologies and tools (such as text annotation workbenches[100]) for constructing reliable biomedical text corpora that can serve as gold standard is an active area of research[50,63,86,100-102].

CHAPTER III

A NOVEL INFORMATION RETRIEVAL MODEL FOR HIGH-THROUGHPUT MOLECULAR MEDICINE MODALITIES

Note

This chapter consists of the content of the published article: Wehbe FH, Brown SH, Massion PP, Gadd CS, Masys DR, Aliferis CF. A novel information retrieval model for high-throughput molecular medicine modalities. *Cancer Inform. 2009 Feb 9;8:1-17.* The content of the "Related Work" section of the article is subsumed by Chapter II of this dissertation and is omitted from this chapter to avoid redundancy. The "Appendix" of the published article is attached to this dissertation as Appendix A.

Introduction

The goal of Molecular Medicine is to diagnose and find treatments for human diseases by the application of tools of molecular and cell biology[1]. In recent years, researchers have begun to link tissue molecular profiles—such as gene expression information—of individual patients to relevant disease outcomes such as diagnosis[2], prognosis[3], and response to treatment[4]. Knowledge discovered from large-scale genomic and molecular biology data is already being put to clinical use[5] and several clinical studies are in the development or validation phase[6].

The field of pharmocogenomics, for example, applies whole genome analysis technologies to predict drug treatment response and adverse drug reaction susceptibility based on individual genetic variability[12,13]. For instance, an inherited genetic trait places

some individuals at risk for adverse drug reactions (diarrhea, neutropenia) to the antineoplastic drug irinotecan[14,103,104]. Individuals with the most common variant allele (UGT1A1*28) have lower expression levels of an enzyme that deactivates irinotecan. The FDA requires that the related genotype and dosing guideline information be included in the irinotecan package insert[11]. Other mutations are associated with a good clinical prognosis[105] and positive response to certain classes of drugs[106]. A listing of drug-related genomic biomarkers is available on the FDA website[11].

In a typical scenario, a molecular assay is performed on tissue obtained from a patient. Then, a decision model computes, based on the assay results, the "predicted" clinical outcome of the patient's disease. For example, the U.S. Food and Drug Administration approved in February of 2007 the first high-dimensional molecular test to predict the recurrence of breast cancer within five to ten years. Many similar tests are expected to follow[107].

Discovering clinically significant knowledge from large-scale genome and molecular biology information is a complicated scientific process that draws from multiple overlapping sources of data describing complex interactions at the genomic, proteomic, or other "omic" levels. High throughput "omic" experimental methods generate data that can have hundreds or even hundreds of thousands of data-points per sample. Such data are difficult to process manually and require sophisticated computation. Decision models that process the resulting data are also complex and draw from a variety of disciplines including biostatistics and machine learning. Furthermore, there is great variability in the methods that evaluate these predictive models' validity, generalizability, and supporting evidence[6].

For advances in molecular medicine to come to clinical fruition, it is crucial for clinical and translational researchers to have access to relevant, up-to-date, and correct information about known molecular medicine modalities[108], such as research datasets, research methods, known and validated decision models, and related evidence. Therefore the important problem of retrieving and organizing the vast amount of information issued from molecular medicine research needs to be addressed. The inherent complexity of this domain and the fast pace of scientific discovery make this problem particularly challenging.

Problem Statement

Our goal is to develop a general purpose information retrieval system that satisfies the following two requirements:

1. The system should be able to index, retrieve and organize most methods of molecular profiling, most forms of predictive computational models, many types of clinical outcome, as well as supporting evidence and computational resources.
2. The knowledgebase needs to be comprehensive and up to date. This requires simple, cheap, fast, and scalable methods to build the knowledge base and to keep it current. To keep up with the rapid pace of discovery in clinical bioinformatics, these methods have to be automated or semi- automated in the worst case. For this system to support the first requirement, its underlying knowledge representation formalism has to convey the semantic complexity of the clinical bioinformatics domain; on the other hand, the underlying formalism has to be simple enough to support the second requirement of relying on scalable

automated methods. The problem, therefore, is to develop a framework and semantic model that balance these two requirements.

This system will also have to accommodate a wide range of query types. Consider the following query examples to be posed by clinicians and/or clinical and translational researchers:

- **Example Query 1**: *"What models exist that predict the response to the chemotherapy regiment (CHOP) in patients with diffuse large B-cell lymphoma (DLBCL)?"* In this query, the following entities are specified: "disease" is specified as "DLBCL"; "clinical outcome" is specified as "response to CHOP". Notice that this question leaves the specific method of "molecular profiling" open. This query might be posed by an oncologist looking for up-to-date knowledge to guide her choice of treatment strategy for her DLBCL patient.
- **Example Query 2**: *"What models exist that predict response to the chemotherapy regiment (CHOP) based on gene expression profile?"* This query does not specify the type of cancer, it does, on the other hand, restrict all desired models to those based on gene expression data. This query may be posed by a researcher in pharmacogenomics looking to correlate the expression of specific genes with the biological function of specific drugs.
- **Example Query 3**: *"What papers have compared multiple supervised learning methods for the prediction of cancer diagnosis based on gene expression data using a cross validation method?"* This query could be posed by a clinical researcher in possession of a gene expression dataset who is looking for proven methods to build and validate models for diagnosing prospective cancer patients

using gene expression microarrays. Notice that in this query, the specific disease and the specific outcome are not specified. Only the type of outcome is specified as "diagnosis". Also notice that this query specifies classes of algorithms ("supervised learning") and validation methods ("crossvalidation") rather than individual methods.

- **Example Query 4**: *"What datasets originating from breast tumor samples contain mass spectrometry data and contain clinical survival data?"* This is a specific query by someone who is interested in building and testing models that predict survival in breast cancer based on raw mass spectrometry data.

These queries require the search and retrieval of a multiplicity of molecular medicine modality object types including but not limited to documents, which are the focus of traditional information retrieval problems. Our envisioned system is intended to represent and retrieve four different types of objects relevant to clinical bioinformatics:

- **Papers**: A published paper is the primary unit of scientific communication. Individual papers or groups of papers describe the methods and results of high throughput molecular medicine research.
- **Datasets**: In many cases, researchers publish their data in the public domain[109]. Often, that data is utilized by other researchers seeking to develop new and improved analysis methods, to test novel hypotheses, or simply to reproduce or validate the published results.
- **Algorithms/Software**: Research laboratories that develop data analysis methods often publish implementation of the algorithms that they have developed and applied[110].

- **Models**: Predictive computational models are produced by the application of algorithms on research datasets. Predictive computational models provide a "decision" based on molecular assays and clinical data obtained from a single patient. The predictive computational model's decision (output) may then be used for the clinical management of the respective patient, for example to help determine the choice of effective therapy. Ideally the process of decision model formation includes rigorous statistical validation to ensure that the utility of a given decision model can generalize to a wider population.

Model Formulation and Proof of Concept

Model: Objects, indexing scheme, and queries

We developed an information retrieval model to support our intended system by examining use cases that mimic the queries introduced above in the domains of diffuse large B-cell lymphoma (DLBCL) and breast cancer. The model is described in the context of the task of retrieving research information from the semantically complex clinical bioinformatics domain of gene expression microarrays in the diagnosis and treatment of DLBCL.

Initially, we conducted manual literature reviews for papers that describe this domain. We noted the different objects that were described in the papers that were reviewed, i.e. by identifying *Algorithms*, *Datasets*, or *Models* described in each *Paper*. Conceptually, the objects in the knowledgebase are all the *Papers*, and the union of all

Algorithms, *Datasets*, and *Models* that are described by the Papers. An *Algorithm*, a *Dataset*, or a *Model* can be referenced in more than one *Paper*.

Further examination of these objects revealed that each can be described by at least one *Context* that specifies the following elements in a tuple: <Disease, Population, Purpose, and Modality>. For example in the Paper by Wright et al.[111], a Model that predicts the molecular subtype of DLBCL was produced and validated by applying the *Algorithm* "Bayes Classifier" on two gene expression *Datasets*. The five objects (1 *Paper*, 1 *Algorithm*, 2 *Datasets*, and 1 *Model*) can each be annotated with the following *Context*: (*Disease* = DLBCL, *Population* = Human Patients, *Purpose* = Predict Molecular Subtype, *Modality* = Gene Expression Microarray).

A query to the knowledgebase should then return a subset of the objects in the knowledgebase. A simple enumeration of *Papers, Algorithms, Datasets*, and *Models* that relate to gene expression microarrays in the context of DLBCL is shown in the left side of Figure 1. We also realized that a query can be represented as a partial or complete *Context*. For example, the *Contexts* represented by the example queries above are shown in Table 1. Queries 1–3 specify partial *Contexts*, and Query 4 specifies a complete *Context*. A quick and simple indexing scheme can be achieved by using a set of canonical terms for each of the *Context* elements, and then indexing each of the objects with at least one complete *Context* tuple. Objects are retrieved when their *Context* elements match the *Context* elements specified in the query.

We conducted a broad search for DLBCL gene-expression-related objects, by placing a query as in Figure 1 that specified the following *Context*: (*Disease* = DLBCL, *Modality* = Genomic). In the following section we will discuss three clinical

bioinformatics scenarios that involve a subset of DLBCL gene-expression-related objects. The scenarios were encountered when we analyzed the set of manually collected objects that satisfied this *Context*. Figures 2–4 will provide a pictorial representation of these scenarios.

Table 1. Contexts partially or completely specified by the example queries in the problem statement section above

Query #	Disease	Population	Purpose	Modality
1	DLBCL	Human Patients	Response to CHOP Regimen	-
2	-	-	Response to CHOP Regimen	Gene Expression
3	-	-	Diagnosis	Gene Expression
4	Breast Cancer	Human Patients	Predict Survival	Mass Spectrometry

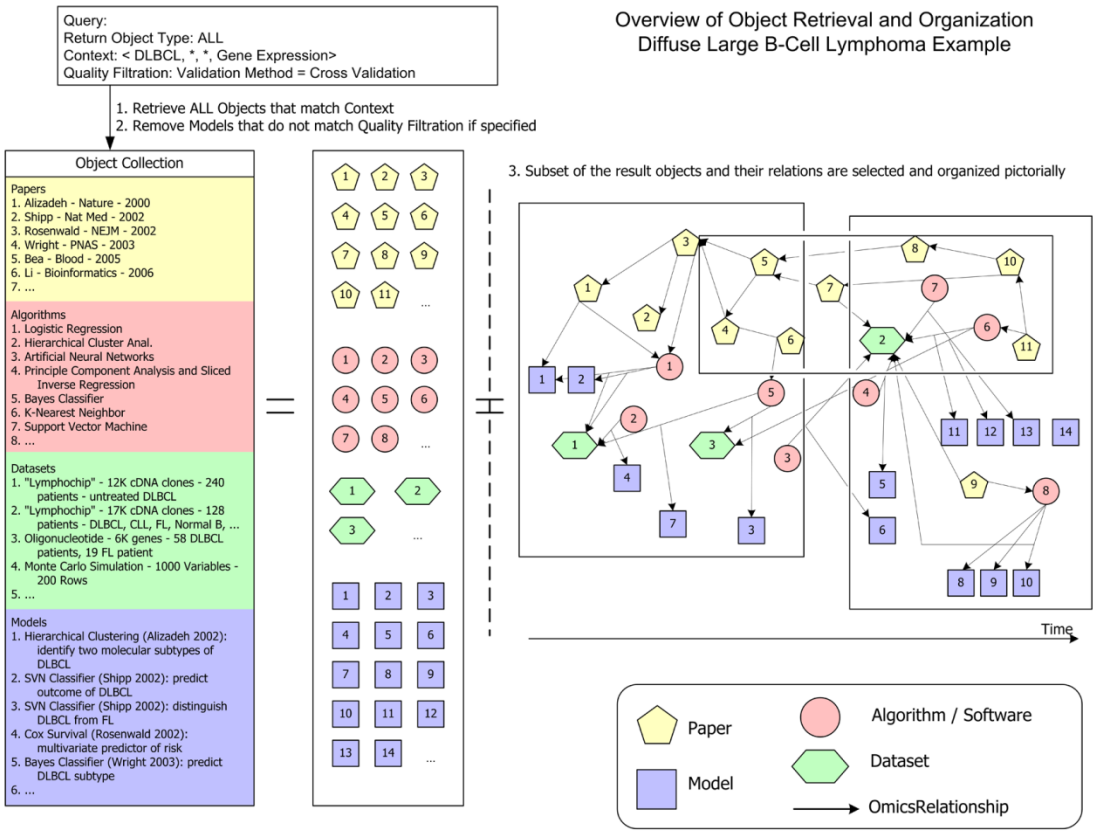


Figure 1. An overview of how the information retrieval model will be applied to the DLBCL use case. Left side: After specifying the desired query parameters (Context, Quality Filtration), the system will return a potentially large result set of molecular medicine modality objects. This enumerated set of objects is the raw result. Please refer to the subsection "Model: Objects, Indexing Scheme and Queries," last two paragraphs. Right side: One or more subsets of the raw result may then be selected by the user for visualization and organization based on the relationships between these objects. The subsection "Model: Object Relationships and Quality Filters" elaborates on this process. The full details of the DLBCL use case are mentioned in the subsection "Proof of Concept: Diffuse Large B C-Cell Lymphoma". Three subsets of objects from the DLBCL domain along with their relationships are organized pictorially according to our model in Figures 2, 3 and 4.

Proof of concept: Diffuse large B-cell lymphoma

DLBCL is the most common form of non-hodgkins lymphoma in adults. Historically, less than half of DLBCL patients are cured by chemotherapy[112]. It was suggested early on that DLBCL actually comprises several diseases that differ in responsiveness to chemotherapy. A pioneering paper by Alizadeh et al. in 2000[113]

applied bioinformatics methods to investigate this hypothesis. They measured gene expression levels in lymphoid tissue collected from a variety of healthy and sick individuals. The microarray platform used, called "lymphochip," measured mRNA levels by hybridization on cDNA spots. The cDNA gene library on the lymphochip was deliberately designed to include genes known to be expressed in lymphoid tissue. The resultant *Dataset*, which consisted of around 17 thousand gene expression analytes for 128 samples, was analyzed using an unsupervised hierarchical clustering *Algorithm*. Based on the hierarchical clustering results, multiple decision *Models* were generated that either related to the biological behavior of DLBCL or to the clinical outcome of patients suffering from DLBCL (See Fig. 2). In the former category, the decision Models seemed consistent with the following hypotheses: (1) That DLBCL can be distinguished based on gene expression data from follicular lymphoma (FL), another form of lymphoma; (2) That there are two molecular subtypes of DLBCL; and (3) That one subtype's molecular signature resembles that of activated peripheral B-cells (APB-like) whereas the other's signature resembles that of B-cells found in the germinal centers of lymph nodes (GC-like). The resultant clinical decision *Model* of this study was that DLBCL samples that clustered in the GC-like category had better survival than those that clustered in the APC-like category.

Two subsequent studies attempted to further investigate and validate the hypotheses that were reported in the Alizadeh *Paper*. See Figure 2 for a graphical view of the objects and relationships that were reported in these three *Papers*. Rosenwald et al. used the same microarray platform, the lymphochip, to collect data from 240 patients with DLBCL[114]. In this study, two *Algorithms* were used. An unsupervised hierarchical

clustering *Algorithm* was used in a similar way to that described in the Alizadeh paper. However, three resultant hierarchical clusters (molecular subtypes) were found and labeled: "Activated B- Cell-like", "GC-B-Cell-like", and "Type 3". The second *Algorithm* relied on multivariate regression techniques to construct a clinical survival prediction *Model* based on (so-called) gene expression scores. The decision *Model* was derived from a *Dataset* of 160 patients and was validated on the remaining 80 patients. This decision *Model* instance was compared to another widely used clinical predictive *Model*, the "International Prognostic Index" (IPI)[115], that predicts clinical outcome based only on clinical parameters. Molecular and clinical data were reported as independent factors in predicting clinical outcomes.

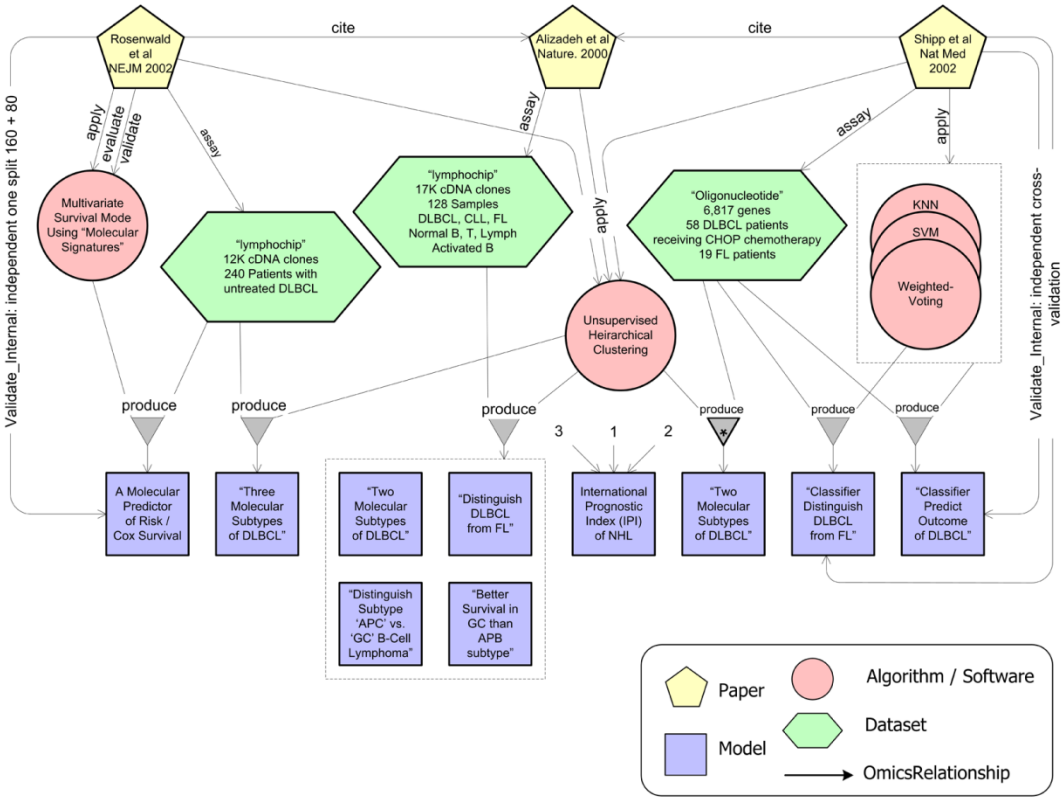


Figure 2. A pictorial representation of the first three widely cited Papers relevant to the DLBCL use case along with the Datasets, Algorithms, and Models that were described in these Papers. Identifying and presenting relationships between these objects is important for the semantic organization of this domain. These relationships are represented by edges connecting the different objects. For example, the three Papers each describe how Algorithms were applied to Datasets to produce decision Models. We identify this class of ternary relationship as Run_ on_Produce (Produce in the figure for simplification). Furthermore, the Shipp (Shipp and others, 2002) and the Rosenwald (Rosenwald and others, 2002) Papers describe how the rightmost and leftmost predictive Models (respectively) were validated using the Datasets that they had assayed. This scenario is detailed in the subsection "Proof of Concept: Diffuse Large B-cell Lymphoma," paragraphs 1–3.

In a third study, by Shipp et al.[116], gene expression was measured in tumor samples from 58 DLBCL patients receiving the CHOP chemotherapy protocol, and from 19 FL patients. In this study, however, oligonucleotide-based microarrays were used instead of the cDNA-based lymphochip. Supervised learning methods (*Algorithms*) were used to construct two predictive classifiers (decision *Models*): one associated with the biological hypothesis that DLBCL can be distinguished from FL based on gene

expression data, and another associated with the clinical hypothesis that gene expression data can predict the clinical outcome of DLBCL. The latter decision *Model* was also compared to the IPI clinical predictive *Model*, and in this study as well, molecular and clinical data were found to be independent factors in predicting outcomes. A more rigorous cross validation method was used to validate the models produced by this study. In this paper, the previous claims about molecular sub-types were put to test. The same unsupervised hierarchical clustering *Algorithm* was applied on their dataset[1] to cluster the samples. Two molecular subtypes did emerge, and they did show "APB-" and "GC-" B-cell-like expression patterns. However, survival was *not* found to be different between the two groups.

[1] Notice that the oligonucleotide sequences on the microarrays platform of this study were matched through their annotations to the cDNA genes in the "lymphochip" platform used in the other studies. Only the sequences that matched were used in this clustering technique. That's why the ternary relationship apply-on-to-produce has an asterisk in Figure 2.

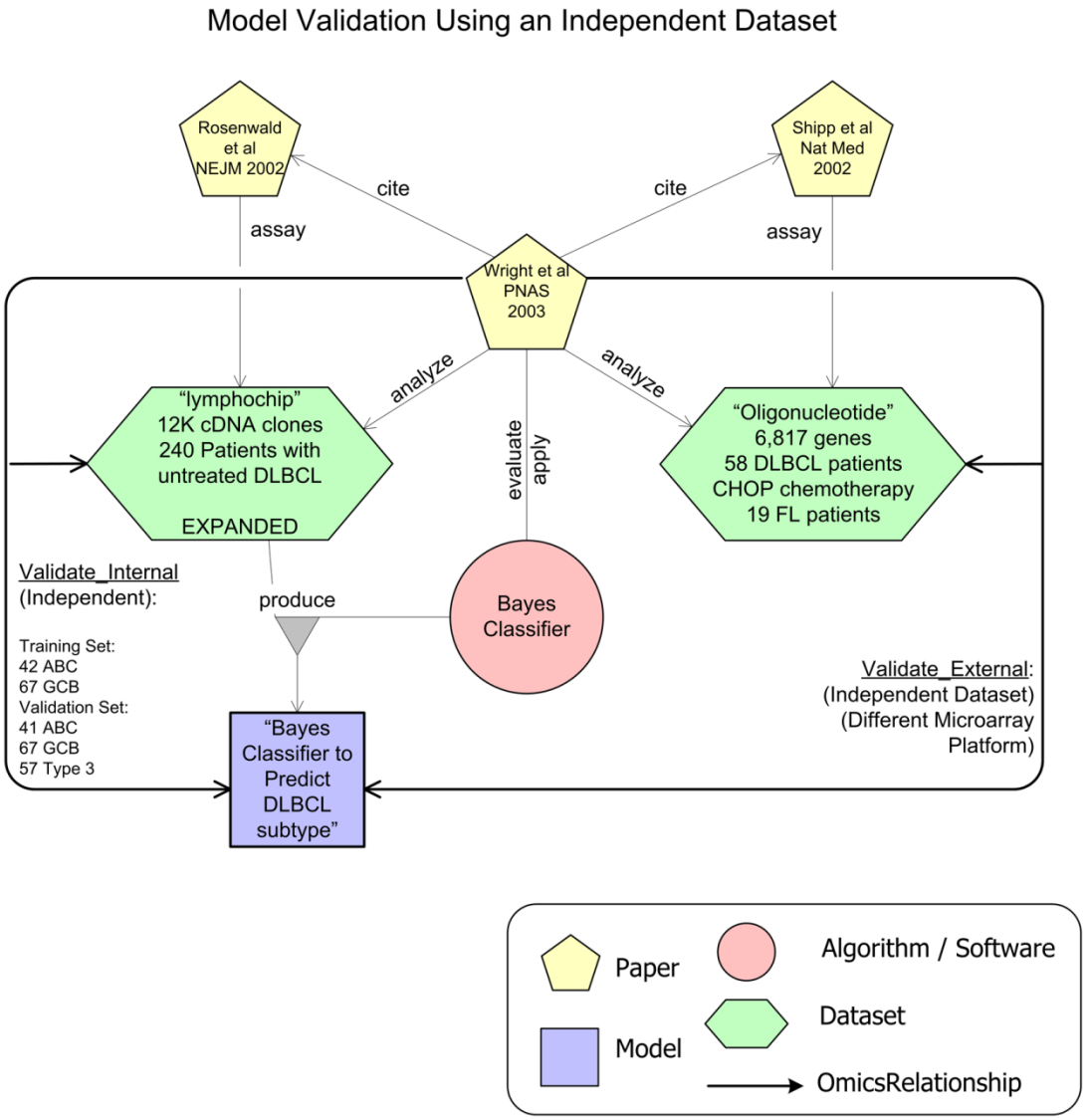


Figure 3. This figure shows the objects and relationships that surround the production and external validation of a Bayes-classifier Model as described in the Wright et al. (Wright and others 2003) Paper and explained in the subsection "Proof of Concept: Diffuse Large B-Cell Lymphoma", paragraph 4. The Model (bottom center) was produced by applying the Bayes-classifier Algorithm to the lymphochip Dataset (left). The Model was internally validated (left side arc) using that Dataset which was split into independent training and testing sets. It was then externally validated (right side arc) using another independent Dataset that was assayed and described in a previous Paper (right). It is important to represent and identify this type of scenario in which higher quality Models are produced, i.e. Models that generalize across different Datasets and, in this case, across different molecular assay platforms (oligonucleotide vs. cDNA).

Wright et al.[111] wanted to reconcile the results from the last two studies (See Fig. 3). They developed a Bayes classifier (i.e. a decision *Model*) to predict molecular sub-

type and clinical outcome. It was trained and validated on the Rosenwald *Dataset* that used the lymphochip platform. The classifier was then independently validated on the *Dataset* produced by the Shipp group, again using sequence annotations to reconcile the cDNA sequences with the oligonucleotide sequences. This seems to support the biological hypothesis that the "two molecular subtypes" in DLBCL correlate with different biological and clinical behavior. The semantics of the relationship between this *Model* and these two *Datasets* is reflected through the visual description and organization in this figure.

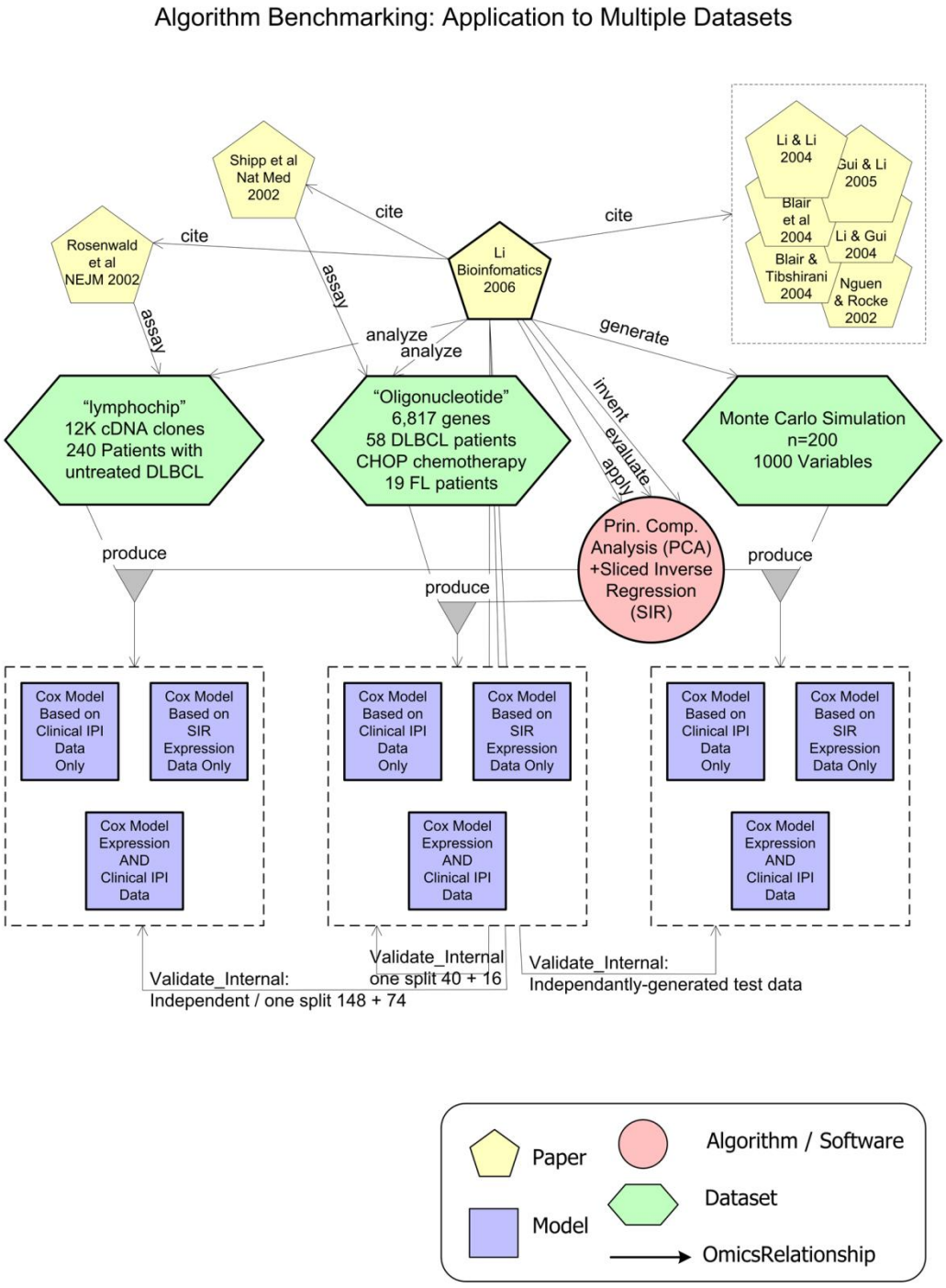


Figure 4. This figure describes how an Algorithm (PCA + SIR) was described by the Li et al. (Li, 2006) Paper. This Algorithm was benchmarked using two independent Datasets that were assayed and described by previous Papers, and one Dataset produced by Monte Carlo simulation. The Models that were produced by the application of this Algorithm on these Datasets were validated internally using one independent split of the respective Datasets. This scenario is commonly encountered in methodological research aimed at developing and benchmarking new classification Algorithms. Please refer to subsection "Proof of Concept: Diffuse Large B-Cell Lymphoma," paragraph 5.

On the other hand, the more recent paper by Li et al.[117] describes a study that develops and evaluates a specific data-analysis method (i.e. *Algorithm*) (See Fig. 4). This *Algorithm*, "Principle Component Analysis and Sliced Inverse Regression", was applied to both the Rosenwald and Shipp *Datasets*, as well as to a *Dataset* produced by a Monte Carlo Simulation. Decision *Models* were generated and they were validated on an independent subset obtained through one split of the data (148 training samples, 74 training samples). This figure focuses on one algorithm in this *Context* and relates all the objects (and relationships) that are relevant to the evaluation of this *Algorithm*.

Model: Object relationships and quality filters

These examples demonstrate that the figures and their underlying complex semantics cannot be conveyed by simple retrieval and enumeration of objects returned by *Context*, i.e. as in the left side of Figure 1. A potentially large number of returned objects need to be organized and displayed intuitively. One aspect of object organization relates to the relationships between the different object types. Such relationships were indicated by edges in the figures. For example, a *Paper* can describe how an *Algorithm* is used to *Analyze* a *Dataset*. A *Model* is *Produced* by running an *Algorithm* on a *Dataset*. *Models* are *Validated* using more than one *Dataset*. Grouping objects in annotated relationships can be leveraged in post-retrieval organization and display to provide semantic information about the objects.

All the predictive *Models* mentioned above underwent some form of validation, expressed via the *Validate* relationships in the respective figures. The *Validate*

relationship is further specialized via the *Validate External* and *Validate Internal* subclasses. Please see the section on evidence annotation in the appendix. As molecular predictive *Models* mature and get closer to routine clinical practice, it is important to consider the evidence supporting their validity and generalizability. As described by Pepe et al.[118], clinical bioinformatics predictive models typically go through multiple stages of validation before being accepted in standard practice. Therefore, our envisioned system will need to filter different objects based on the strength of supporting evidence. For example, these query results can be narrowed to include only high quality models by appending the following requirements to the query *"[get models that …], have been developed using datasets with sample size (n) larger than 200 patients, and that have been validated using an independent dataset."*

The concepts mentioned so far that will support the information retrieval model are described in more detail in the appendix. Now we can revisit Figure 1 in its entirety. It gives an overview of how a query is intended to be processed: A query sets the desired object types, specifies a partial or complete *Context(s)*, and sets conditions for quality filtration. The process is decomposed into three steps: (1) returning objects that are indexed by *Context* tuples that match the query's *Context*, (2) filtering out objects based on quality of evidence, and (3) selecting smaller sets of objects by the user and organization of these objects along with their relationships in an intuitive way.

Proof of concept: Molecular prognostic test for breast cancer—MammaPrint®

The same semantic representation and organizational principles of *Papers*, *Datasets*, *Algorithms*, and *Models* that relate to MammaPrint®, the first commercial Breast Cancer molecular prognostic test, are shown in Figure 5 and explained below.

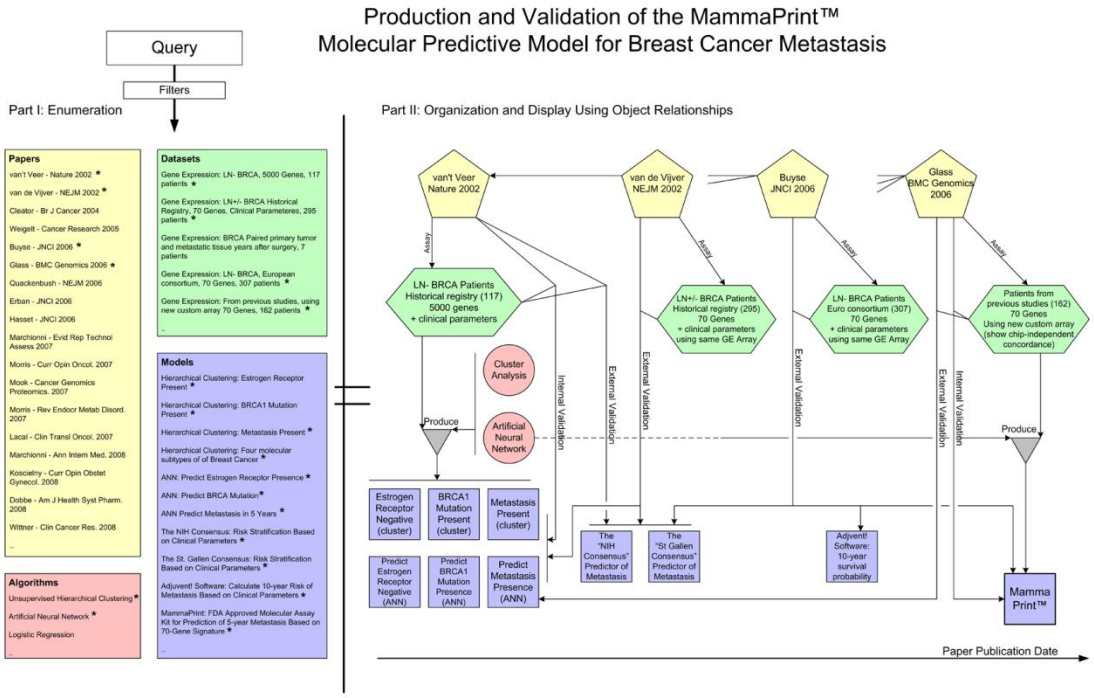


Figure 5. This figure depicts objects and object relationships that span the development and evolution of the MammaPrint™ Model from its earlier versions. The figure also represents the validation of MammaPrint™ across multiple Datasets and its comparison to other Models. Notice that the other clinical predictive models are classical models that do not incorporate molecular data. The information retrieval framework will incorporate classical (non-molecular) clinical predictive Models only when they are relevant to the validation of molecular prediction Models. Otherwise classical Models will not be indexed or stored. Similar to the process described in Figure 1, a query to this domain will return a raw set of objects (Part I, left side). A subset of the raw result may be selected for visual organization and display (right side) of the objects and their relationships (Part II, right side). The detailed prose description of this scenario is presented in the subsection "Proof of Concept: Molecular Prognostic Test for Breast Cancer—MammaPrint®".

Researchers in the Netherlands[5] analyzed historical breast cancer tissues using a 25,000 sequence oligonucleotide microarray. Seventy genes were found to be predictive of 5-year metastasis in Lymph Node (LN)-negative female patients under the age 55. Unsupervised hierarchical clustering (*Algorithm*) distinguished the following three

characteristics: Estrogen-receptor negative (i.e. cannot be treated with the drug Tamoxifen), having BRCA1 germline mutation, and metastasis within 5 years. In other words, three *Models* were Produced using the hierarchical clustering *Algorithm*. A supervised machine learning method, Artificial Neural Network (ANN, another *Algorithm*), was used to construct a classifier (*Model*), using a "70-gene signature", that predicts these characteristics. This predictive *Model* was *Validated Internally* using a leave-one-out approach. The researchers also showed that this molecular predictive *Model* was an independent predictor of metastasis from other well-known decision *Models* that relied solely on clinical parameters (the NIH Consensus and the St. Gallen Consensus). In that paper, not only did the molecular decision *Model* improve clinical outcome prediction, but it also predicted the same number of patients who had metastasis with fewer false positives. This is important given the morbidity and economic costs associated with adjuvant chemotherapy[119,120]. The 70-"gene signature" *Model* was *Externally Validated*[121] using 295 consecutive historical patients in a *Dataset* that is different from the *Dataset* that was used to *Produce* that signature. It also provided[122] the correct decision outcome, i.e. *Externally Validated*, on primary tumor tissue from 7 patients and on matched metastatic tissue obtained years later from the same patients (not shown in Fig. 5). This validation was not of a clinical, but of a biological hypothesis that: molecular subtype determines the metastatic potential early in the disease as opposed to invasiveness resulting from cumulative mutations[2].

A spin-off commercial company, Agendia™, developed a custom kit that measured gene expression and contained a similar 70-"gene signature" *Model*, now called

[2]That same study Validated a decision Model described elsewhere (also not shown in Fig. 5) that used unsupervised clustering to separate Breast Cancer samples into four molecular subtypes. All matched primary tumors and metastatic tissue belonged to the same molecular subtype.

MammaPrint®. MammaPrint® was also *Produced* using the ANN *Algorithm* and *Internally Validated*[8]. The new platform was shown to be concordant with the previous 25,000 oligonucleotide chip[8] (thus *Externally Validating* that *Dataset's* corresponding *Model*). MammaPrint® was *Externally Validated* through multi-center European consortium study[123]. It was also compared to known clinical decision *Models*, including one based on a software, Adjuvant!, that calculates 10-year survival probability based on clinical parameters.

Discussion and Future Work

Some public resources currently implement some but not all aspects of our intended functionality and not in an integrated retrieval framework as was discussed in this paper. For example, PharmGKB's clinical outcomes are restricted to outcomes of therapy, and exclude diagnostic and prognostic markers. Oncomine's representation and organization of oncology molecular datasets does not cover decision *Models*, the original *Algorithms* by which these models were produced, or their validation methods. *Datasets* and *Papers* are MeSH-indexed in GEO/PubMed, but their relationships to respective *Models*, *Algorithms*, and *Contexts* are not explicit. The proposed framework is designed to complement existing resources and extend current representations to cover molecular clinical predictive models and their related modalities. Our choice to model this domain using an OWL ontology was made with the goal of semantic integration of this framework with existing knowledge sources. Whenever possible we associate objects in our database with their counterparts in external databases, e.g. using PubMed uid for papers and GEO accession numbers for datasets.

Most existing clinical predictive models do not incorporate molecular features. Classical predictive models that are purely based on clinical parameters are outside the scope of this information retrieval framework; however, classical models will be incorporated *only when* they exist within the context of molecular predictive models. For example, we did include the International Prognostic Index model in the DLBCL case study, and the St. Gallen Consensus model in the MammaPrint™ validation case study. Similarly, storing and annotating gene signatures that predict underlying biological behavior without clinical outcomes is outside the scope of this framework. Again, some molecular clinical predictive models incorporate aspects of purely biological signatures, so we will also include those only when they exist within the context of clinical models. For example, the early DLBCL models (Fig. 2) that identified the underlying biological behavior of DLBCL (as APB-like or GC-like) did correlate with clinical outcomes and therefore they were included in the framework. Using molecular signatures that measure (EGF-R) receptor activity for choice of treatment with tyrosine kinase inhibiting drugs is another example (not discussed in this paper) that comes to mind of what will be included in this framework.

The focus of the present paper is the underlying information retrieval model and not the system's implementation and inference mechanisms which will be described elsewhere (please see Appendix). When developing the formalisms described in this paper, we deliberately selected the minimal set of classes and properties that is expressive enough to allow for semantic organization of the domain. This level of simplicity is intended to enable automated methods for building the knowledgebase. Our current research is focused on building and validating machine learning models that can correctly

annotate the *Contexts* described in clinical bioinformatics papers, and that can also correctly identify the validation methods that are employed in those papers.

Conclusion

While clinically-oriented research exploring gene expression microarrays, mass spectrometry, SNP arrays and other high-throughput molecular assays has followed an exponential growth in recent years, to date there is no general purpose system that allows researchers and clinicians to fi nd models, papers, data, and other related information in this emerging field using a unified and friendly interface. In the present paper we propose a framework for such interface and demonstrate the complexity of its required functionality. Our long- term goal is to construct a system that addresses this need. As a significant first step, we developed a formalism that supports storage and retrieval of a multiplicity of clinical bioinformatics objects such as published papers, datasets, decision models, and discovery and inference algorithms. This formalism opens the way for automated methods that support the knowledgebase's creation and annotation. In addition, it allows for a second layer of organization of objects returned by queries based on their (1) interrelationships and (2) strength of methodological validation. We demonstrated the power of this model in the complicated domain of diffuse large B-cell lymphoma. In future work we plan to deploy and test a prototype system based on the model of the present paper applied to biomarker discovery for other malignancies.

CHAPTER IV

MACHINE LEARNING FILTERS FOR RETRIEVING RELEVANT MOLECULAR MEDICINE PUBLICATIONS

Introduction

In the previous chapter, an information retrieval framework for storing and organizing clinical bioinformatics predictive models was proposed. Within this framework, these models are annotated and indexed using a set of attributes (the clinical bioinformatics *Context* described in Chapter III and appendix A) that were determined based on the semantic analysis of this domain and the types of queries that this framework is designed to address. A knowledgebase in which these models are stored is integral to this framework. The information that will be used to populate this knowledgebase will be derived from published articles that describe clinical bioinformatics predictive models. As discussed in Chapter II, the manual retrieval of relevant papers from the published literature and the semantic annotation of models described herein are tedious tasks. Scalable methods are required to ensure that the knowledgebase is comprehensive and up-to-date with the rapid pace of published clinical bioinformatics research. This chapter will describe the building and evaluation of machine learning filters for automated retrieval of relevant papers from MEDLINE (see Figure 6). The next chapter will describe the use of machine learning methods for automated or semi-automated semantic annotation of the relevant papers according to the semantic annotation scheme.

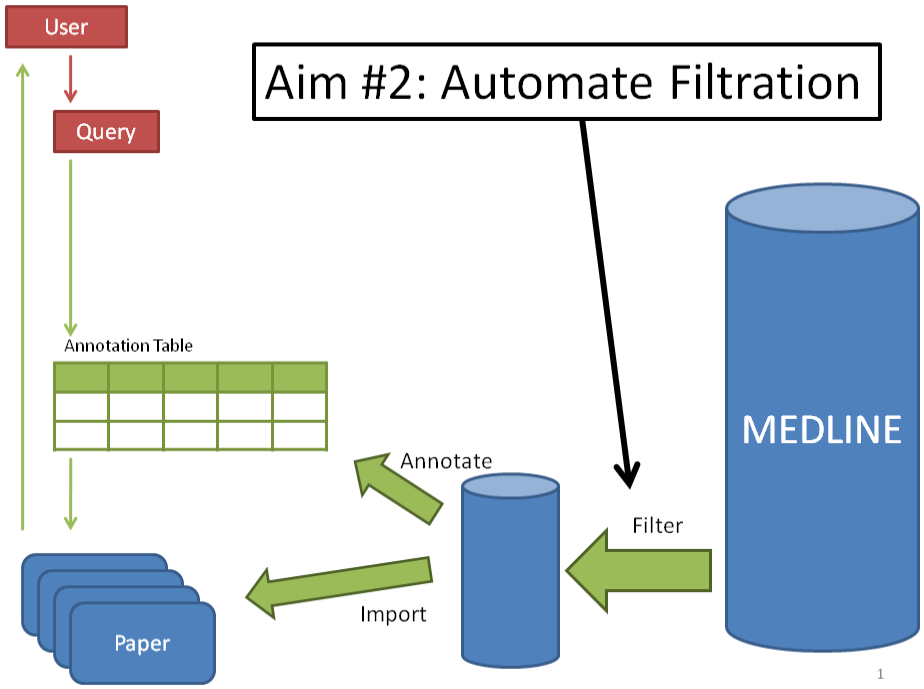


Figure 6 – The second aim of the PhD dissertation is to build and evaluate reproducible scalable automated methods for identifying relevant clinical bioinformatics papers from the MEDLINE database. The set of relevant papers will be annotated in subsequent steps to build the knowledgebase that will support the overall information retrieval framework.

Operational Definition of the "Relevant Articles"

The semantic model and derived annotation scheme described so far were developed based on a thorough analysis of a specialized and relatively small set of papers. These papers were obtained via a focused ad hoc search of MEDLINE. Constructing a large corpus of related papers for the research described in this chapter (and for ultimately building the knowledgebase itself) requires an unambiguous definition of "relevant papers". This definition will need to provide operational guidance to human annotators for making consistent determinations whether articles in MEDLINE are relevant or not. This definition was made after a pilot manual annotation of an expanded

set of MEDLINE articles. It was further refined during multiple discussion with domain experts and annotators. This culminated in a written document (Appendix B – Annotation Guidelines) that was used as an annotation manual by me and by other domain experts who were part of this research. This annotation guideline also provided functional definitions for the concepts in the annotation scheme itself which have also been refined from the clinical bioinformatics *Context* described in Chapter III. The methods section in this chapter will describe the part of the annotation guideline that defined "relevant papers". The following chapter will describe the remaining part of the guideline which provided operational definitions to annotations which were applied once the papers were determined as "relevant".

Automated Filtration Methods

The scalability of building and maintaining an up-to-date collection of relevant papers can be achieved via automated filters (Figure 6). In the context of this framework, filters are defined as text classifiers that assign positive or negative labels to papers based on the text content of their MEDLINE record. Statistical machine learning models have been shown in the past to reliably replicate human classification tasks for MEDLINE article retrieval.

Using machine learning requires the conversion of free unstructured text into numerical features that can be used to compute a given paper's classification. Feature extraction includes counting the frequency of occurrence of words in the given text followed by linguistic and semantic transformations such as word stemming or stop-word removal. Other types of feature extraction steps exist and may depend, for example, on

the location of terms within the PubMed record (title, abstract, MeSH term, etc.) The machine learning filters that will be used will be based on Support Vector Machines (SVM). SVMs are supervised machine learning classifiers that require a training dataset of known classification outcome. The ability of these SVM filters to discriminate relevant papers from other MEDLINE articles will be measured by applying them to a test set of papers of known classification (gold standard) and examining the resulting area under the receiver operating characteristic curve (AUC).

The first research question in this context is:

- *Can existing or modified feature extraction transformations be used to train machine learning filters that can identify relevant papers from MEDLINE?(Aim 2a)*

The dataset that will be used for this question will be selected from the domains of *bioinformatics and lung cancer*. The performance of the SVM filters will be evaluated using N-fold cross validation in which the gold standard is separated into multiple independent training and testing sets.

If the performance of machine learning filters is sensitive to the domain (disease) of the papers in a dataset, then new filters need to be trained using gold standards built for all possible diseases. This can be avoided if the SVM filters have the ability to find relevant articles in other medical specialties. Therefore, the next research question is:

- *Can the filters that were trained using the bioinformatics and lung cancer gold standard, and found to have favorable performance identify relevant papers in other domains?(Aim 2b)*

This will test the generalizability of the clinical bioinformatics filters to other medical specialties and requires an additional gold standard dataset derived from MEDLINE articles from the domain of *breast cancer*.

The training and testing datasets used for research questions above are based on one person's (me) attempt to consistently apply labels about the relevance of these papers to clinical bioinformatics. Therefore the final research question will assess generalizability along a different dimension: annotation by a different set of experts:

- *Can filters trained to identify relevant papers in the domains of bioinformatics and lung cancer using annotation by one expert, identify relevant papers in the same domain as judged by other experts whose annotations were not used to train those filters?(Aim 2c)*

Methods

Defining Relevant Papers and the Annotation Form

As discussed above, filtering "relevant" papers from the MEDLINE database requires an operational definition of what constitutes "relevant" papers. The information retrieval framework provided an indexing scheme for clinical bioinformatics predictive models that is based on a clinical bioinformatics *Context* ontology (Chapter III and Appendix A). Models or papers that describe models should be annotated along four different dimensions that constitute that model's clinical bioinformatics *Context*: *Disease*, *Population* (biological sample), *Modality* (assay type), and *Purpose* (type of clinical outcome). A paper is relevant if it is amenable to annotation according to this scheme.

The first criterion for relevance is that **a paper should describe a predictive model**. A paper describes a predictive *Model* if the authors are trying to establish a statistical relationship between a set of independent variables and one or more dependent variables (outcomes). Figure 7 shows a conceptual representation of a predictive *Model* for this purpose. An intentionally relaxed and widely inclusive definition of statistical relationships is chosen for our definition of *Model*. This can include simple tests of the difference of probability of certain measurements occurring in two or more categories (outcomes) using parametric tests such as: *t-tests, ANOVA, fisher exact test,* or non-parametric tests such as: *Kruskal –Wallis, Wilcoxon*, or *Mann-Whitney*. This also includes models that measure statistical correlation between the values of independent variables and the values of dependent variables, i.e. explicit statistical prediction models like: *linear or logistic regression, and Kaplan-Meier models.* Sometimes the relationship between independent and dependent variable is explicitly presented via symbolic *mathematical equations* or via machine learning models that may include *artificial neural networks, support vector machines, decision trees,* or *Bayes classifiers.*

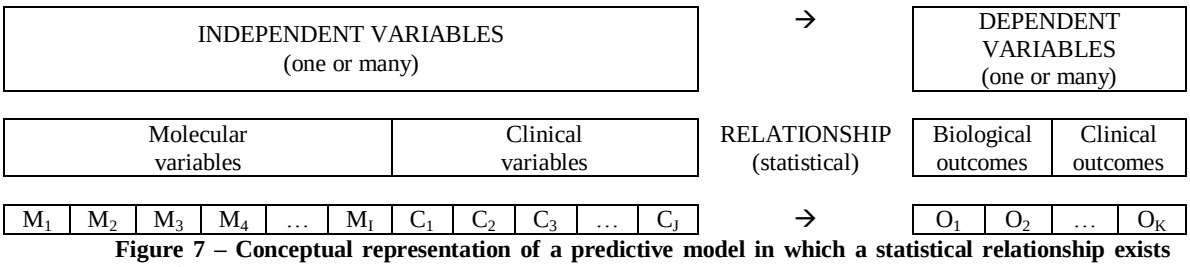


Figure 7 – Conceptual representation of a predictive model in which a statistical relationship exists between a set of independent variables and one or more dependent variables (outcomes).

The independent variables can represent any type of quantifiable observations or experimental measurements. They can correspond to measurements obtained via

molecular biology techniques or assays using biological samples such as local gene expression levels, protein concentrations, or the presence or absence of proteins by using antibodies on biological samples. The independent variables may also represent clinical measurements or patient characteristics such as sex, age, or the presence or absence of disease states (lymph node metastasis, histological subtype, etc.) The *Population* (biological sample) and *Modality* (assay type) components of the *Contexts* are only applicable if the independent variables correspond to molecular biology assays. Therefore the second criterion for relevant papers relates to the independent variables used by the model described in those papers, namely that **the model's independent variables should include molecular features.**

The dependent variables correspond to a model's outcomes of interest. They are also quantifiable observations, based on the type of scientific hypothesis that the authors are investigating. The outcomes of interest can be classified as biological (e.g. cell apoptosis, activation of intra-cellular cascades, cell mobility, presence of a specific protein) or clinical (patient death, presence or absence of disease, response to treatment, treatment toxicity.) The *Purpose* component of the clinical bioinformatics *Context* of model is an assertion about the type of clinical outcome that the model is attempting to correlate with the independent variables. Therefore the third criterion for determining relevant papers is: **the model's dependent variable (outcome) should represent a clinical outcome**.

A pilot review that included partial or complete annotation of over 400 papers was done. This led to the iterative development of an annotation form (Appendix C) and an accompanying set of annotation guidelines that aimed to clarify and operationalize the

criteria above (Appendix B). The guidelines were refined based on feedback from different experts (e.g., meeting with Dr. Pierre Massion and members of his research group). This annotation form was used to annotate all the datasets used in this research. The top part of the form consists of five "yes/no" questions about the given paper. The bottom part of the form consists of multiple boxes where multiple labels can be circled. Each of the boxes corresponds to one dimension of the clinical bioinformatics *Context*. The answers in the top part dictate whether specific boxes in the bottom part should be used. For example if the paper describes a predictive model (question 1 = yes) which uses molecular features as independent variables (question 3 = yes) then the annotator is asked to circle the type(s) of assay used to collect the molecular data and the type(s) of biological sample that was assayed. In other words, the first part of the form contains "gateway" questions that activate the different annotation components in the second part. The 5 questions in the top part of the form correspond to the "filtration" step of the information retrieval framework that is the focus of this chapter. The boxed questions in the bottom part of the form correspond to the "annotation" step of the information retrieval framework and will be further discussed in the next chapter.

By asking the annotators to answer five questions that pertain to a given paper's relevance, the definition of the concept "relevant paper" is essentially decomposed into simpler non-vague atomic definitions. This provides a form of cognitive assistance that can improve the consistency of annotation, because the annotators are asked to make more concrete and focused judgments about the content of the given paper.

Table 2 – The five questions in the first part of the annotation form which were used as target features for the machine learning filters described in this report

Variable Name	Domain	Question
T1	entire paper	Does the article describe at least one predictive model?
T2	independent variables	Is there a model that has more than one independent variable?
T3		Is at least one of the independent variables a molecular measurement?
T4	dependent variables (outcomes)	Is one of the outcomes a clinical outcome?
T5		Is one of the outcomes a biological outcome?

Dataset Construction

Five different datasets were compiled using papers in MEDLINE. The first dataset, named '**Firas-0**', contains 301 articles. It was derived using ad hoc queries and was mostly used to test and refine the annotation guidelines, to determine a preliminary list of journals used in subsequent datasets and to run preliminary machine learning experiments. The remaining three datasets were compiled using combinations of the following PubMed queries:

- A **structural query** is a generic query that specifies the language of the article to be in English and excludes certain types of articles such as: 'review', 'news', 'letter', 'editorial', etc.
- A **date query** specifies a date window from January 2006 until June 2009.
- Three **journal queries** each specifying a mutually exclusive set of journals

- A **bioinformatics journals query** composed of 12 journals that represent the domain of bioinformatics. The 12 journals were provided by Dr. Constantin Aliferis.
- A **lung cancer journals query** composed of 23 journals that represent the domain of lung cancer basic and/or clinical research. The 23 journals were provided by Dr. Pierre Massion.
- A **breast cancer journals query** composed of 7 journals that represent the domain of breast cancer basic and/or clinical research. Dr. Fouad Boulos, a board certified pathologist trained in breast cancer pathology, provided 12 journals. Five of the 12 journals were excluded because they were already provided by Dr. Massion.

Two baseline queries were done to define two mutually exclusive populations of articles in MEDLINE:

1. The **lung cancer + bioinformatics population** was defined as the MEDLINE articles from the 35 journals defined by Drs Massion and Aliferis that fell within the date window of the date query and satisfied the constraints of the structural query. It contained **58,252** articles.
2. The **breast cancer population** was defined as the MEDLINE articles from the 7 journals defined by Dr. Boulos that also fell within the date window of the date query and satisfied the structural query. It contained **5,320** articles.

Four article sets were randomly sampled from those two populations. They were annotated using the forms and guidelines described in the previous sections. The article

sets are summarized in Table 3 and Figure 8. More details are shown in Tables 5 and 6 in the results section.

When generating datasets for machine learning experiments, the features were extracted from the MEDLINE record for each article. The target features were obtained from the annotation of that article by me or by the experts. Each article had five binary target features (classes) corresponding to the "yes/no" value assignment by the annotators to questions T1 – T5. Table 5 in the results section shows the fraction of articles in each article set where these questions were labeled as "yes."

Table 3 – Source, size and function of the three main datasets used for Aim 2.

Article Set Name	Annotator	Size	Baseline Population	Used for Aim (Training / Testing)
Firas-1	Firas	500	lung cancer + bioinformatics	Aim 2a (Train +Test) Aim 2b (Train) Aim 2c (Train)
Firas-2	Firas	200	breast cancer	Aim 2b (Test)
Experts-1	Multiple Experts	340	lung cancer + bioinformatics†	Aim 2c (Test)
Common	Firas + Experts	10	Lung cancer + bioinformatics	Aim 2c (Kappa)

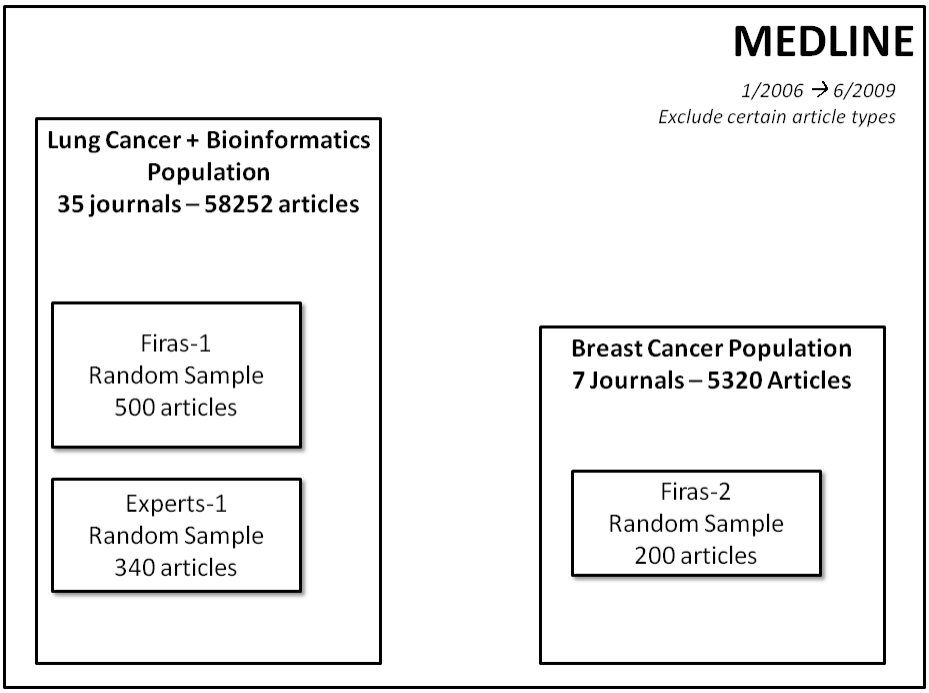


Figure 8: The source and relationship between the three article sets and the article populations. The baseline populations of articles are selected from MEDLINE using a time window between January 2006 and June 2009 and by excluding certain article types like review articles, letters to the editor, etc. The lung cancer + bioinformatics population is selected from a set of 35 journals: 12 bioinformatics and 23 lung cancer. The breast cancer population is selected from 7 lung cancer journals. The populations do not overlap. The Firas-1 and Experts-1 do not overlap and are each selected randomly from the lung cancer + bioinformatics population. The Firas-2 article set is randomly selected from the breast cancer population.

Expert Annotation

Experts were recruited to annotate batches of articles as part of the Experts-1 article set. A $2000 dissertation enhancement grant from the Graduate School at Vanderbilt University was used to recruit the experts. Once the subjects agreed to participate in the study, I met with them and verbally explained the purpose and conceptual framework of the study (in some cases that was more thoroughly discussed at a group lab meeting). The annotation guidelines were provided in print to each expert. The subjects were asked to annotate batches of 30 papers over a period of few weeks. Some experts were able to provide annotation for more than one batch. The first 10

papers of the first batch given to every subject were identical for all participants. This was done to collect data for inter-annotator agreement analysis (see "common" article set above). The subjects were asked to answer all the questions on the annotation forms (i.e. both top and bottom section). The context annotation questions in the bottom section will be used for analysis in Chapter V.

The education level and occupation among experts was diverse and included: pre-doctoral trainees in basic biological or translational science, post-doctoral fellows, physician scientist (faculty), medical students, medical librarians, and epidemiologists. No personal information was collected for the study. The IRB (exemption) study number is 100576. The subjects were compensated $150/batch. One expert declined receiving compensation. One subject (expert #5) asked to be excused from completing the batch for reasons unrelated to the study. That subject's batch was reassigned to expert #2.

Document Representation for Machine Learning

Articles were formatted for learning by text preprocessing and term weighting. Individual terms in the abstract, individual terms in the title, and individual MeSH headings were extracted from MEDLINE records to count their frequency of occurrence within the record. When word stemming was done, multiple forms of the same word were eliminated using Porter stemming algorithm[124] to reduce the dimensionality of the input space.

Terms were weighted using log frequency with redundancy as described by Leopold and Kinderman[125]. First, the number of times a term appeared in a document was transformed into a log frequency. Then it was multiplied by an importance weight (i.e.

redundancy). Redundancy measured how uniformly distributed a term was throughout the entire dataset. A term appearing in all documents is not helpful for classification. A term appearing many times in one article while occurring once in each of the remaining articles is more discriminative[125].

The redundancy value for term *k*, r_k, is:

$$r_k = \log N + \sum_{i=1}^{N} \frac{f(w_k, d_i)}{f(w_k)} \log \frac{f(w_k, d_i)}{f(w_k)}$$

where N is the number of documents in the corpus, $f(w_k, d_i)$ is the number of occurrences of term *k* in document *i*, and $f(w_k)$ is the number of occurrences of term k in the corpus. The final step was L2-normalization to account for different text lengths. The vector of feature weights for a document *i*, x_i, is:

$$x_i = \frac{l_i * r}{\left\| l_i * r \right\|_{L2}}$$

where l_i is a vector of the log frequencies for all terms in document *i*, r is a vector of redundancy values for all terms in the corpus, $l_i * r$ signifies component multiplication, and $\left\| l_i * r \right\|_{L2}$ is the L2-norm of the resultant vector. Each weight was a value between 0 and 1.

Alternative pre-processing approaches of the corpus without term weighting were done using only L1- or L2-normalization. In all cases, the corpus was represented as a matrix where rows corresponded to documents and columns represented terms.

Machine Learning Method and Error Estimation

Support vector machine (SVM) models were used as the learning algorithm. They are a supervised learning method where a kernel function maps the input space to a higher-dimensional feature space, and a hyperplane is calculated to separate the classes of data[126]. The optimal hyperplane is the solution to a constrained quadratic optimization problem. SVM models are usually sparse since the solution depends on the support vectors or points closest to the hyperplane[127]. Most features have zero weights, and the number of support vectors will be much smaller than the number of instances in most cases. This property makes SVMs suitable for representing text which typically involves high-dimensional data. Prior research has demonstrated that they perform well in categorizing text and identifying high-quality articles[66,125].

The SVM models' performance, their ability to discriminate between positive and negative cases in test samples, was measured by area under the receiver operating characteristic curve (AUROC or AUC). The ROC is a plot of the true positive rate (sensitivity) against the false positive rate (1-specificity) at different output thresholds of the decision function calculated by the SVM model for each test case. The AUC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. An AUC of 0.5 describes a random classifier, AUC of ~0.75 a mediocre classifier, AUC of ~0.85 a very good classifier, and AUC > 0.9 an excellent classifier (while an AUC of 1 denotes perfect classification).

For Aim 2a, the SVM models were tested using 5-fold stratified nested cross validation over the Firas-1 dataset. The dataset was randomly split (stratified) into five folds, each containing 100 articles. For every target feature (questions 1-5 in the

annotation form), four folds were used as a training set and the fifth was left out as a test set. This was repeated for all 5 folds as follows: The training and model selection within each training set of 400 articles was done using another internal nested 5 fold (80+320) cross validation to optimize for cost and degree. The set of costs was [0.1, 0.2, 0.4, 0.7, 0.9, 1, 5, 10, 20, 50, 1000], and the set of degrees was [1, 2, 3, 4, 5, 7]. The best performing model was then trained on all 400 documents in the training set and applied to the 100 documents in the test set to obtain an unbiased estimate of the model's performance. The area under ROC for all five test folds is reported in the results section. For Aims 2b and 2c, the models were trained using the Firas-1 dataset by using cross validation to optimize model parameters then using the best parameters to train on the entire Firas-1 dataset. The models were then saved in files. The models were loaded and tested on the independent datasets Firas-2 (Aim 2b) and Experts-1 (Aim 2c) to test their generalizability.

The dataset preparation, feature extraction, and filter training and validation were all done using the Python programming language with the PyML machine learning module. Dr. Yin Aphinyanaphongs has graciously provided a code library that interfaces with PyML and that supports many of the manipulations that were required for this analysis.

Results

Summary of Manual Annotations

A summary of the manual annotation by Firas and by the experts are presented in this section. Table 4 shows the number of articles from each journal in the lung cancer + bioinformatics population. The rows containing bioinformatics journals provided by Dr. Aliferis are shaded. The rest were provided by Dr. Massion. The Firas-1 and Experts-1 article sets were randomly sampled with no overlap from that population. For Experts-1, fourteen batches of 30 articles were given to eight unique experts (See Figure 9). The first batch for every expert contained the 10 articles set aside for the common article set, so only 20 articles were added to Expert-1. For experts who did more than one batch, all 30 articles in the subsequent batches were added. Except for the 10 common articles, there was no overlap between the expert batches. For example, expert #6 did three batches. The first batch contained the 10 common articles + 20 articles that were added to Expert-1. The second and third batches contained 30 articles each. Therefore the total number of articles from expert #6 that were included in the Experts-1 article set was 20 + 30 + 30 = 80.

Table 4 – This table describes the lung cancer + bioinformatics baseline population of articles. The Firas-1 and Expert-1 article sets were randomly sampled from this population. There is no overlap between the two article sets. For each journal, the number of articles in the population and the two articles sets are shown. The journals listed were provided by Drs. Mission (23 journals, lung cancer research) and Aliferis (12 journals, shaded, bioinformatics).

Journals (lung cancer + bioinformatics)	Population	Firas-1	Experts-1
Proc Natl Acad Sci U S A	12205	102	66
PLoS One	6036	46	32
Cancer Res	4675	40	22
Nucleic Acids Res	3940	37	34
Clin Cancer Res	3083	28	14
Int J Cancer	2479	26	19
J Clin Oncol	2448	19	15
Bioinformatics	2318	23	13
BMC Bioinformatics	2227	17	13
Oncogene	2001	14	11
N Engl J Med	1896	19	13
Br J Cancer	1861	16	10
Cancer Epidemiol Biomarkers Prev	1515	6	13
Am J Pathol	1296	11	6
J Clin Invest	1067	12	8
Carcinogenesis	975	7	8
Am J Respir Crit Care Med	973	10	6
Lung Cancer	891	11	6
PLoS Comput Biol	774	8	2
Nat Genet	746	5	4
Mol Cell Proteomics	666	5	4
J Thorac Oncol	641	4	6
PLoS Med	640	6	4
Nat Med	554	5	0
J Pathol	537	7	3
J Comput Biol	331	3	2
Cancer Cell	299	4	2
J Biomed Inform	262	0	0
IEEE/ACM Trans Comput Biol Bioinform	227	2	2
Pac Symp Biocomput	189	1	0
Artif Intell Med	175	2	1
Cancer Prev Res (Phila Pa)	121	1	0
OMICS	88	3	0
Brief Bioinform	61	0	1
Int J Data Min Bioinform	55	0	0
Total	**58252**	**500**	**340**
lung cancer	**47602**	**404**	**272**
bioinformatics	**10646**	**96**	**68**

Figure – 9. The source and composition of the Expert-1 article set. Eight experts were asked to annotate 14 batches of 30 papers. The first 10 papers of the first batch that every expert received were identical and used for analysis of inter-annotator agreement (The Common article set)

The following table (Table 5) shows the number of articles in the breast cancer population and in the Firas-2 article set that were randomly sampled from that population.

Table 5 - This table describes the breast cancer baseline population of articles. The Firas-2 article set was randomly sampled from this population. For each journal, the number of articles in the population and the article sets are shown. The journals listed were provided by Drs. Fouad Boulos (12 journals, breast cancer research, 5 were excluded for overlap with Dr. Massion's journals).

Journals (breast cancer)	Population	Firas-2
Cancer	2302	101
Breast Cancer Res Treat	1117	38
Mod Pathol	651	21
J Natl Cancer Inst	589	19
Breast Cancer Res	389	15
Lancet Oncol	271	6
Breast Cancer	1	0
Total	**5320**	**200**

Table 6 – This table shows the percentage of articles in each article set where the corresponding filter question (T1 – T5) was annotated as "yes." Experts-1 was annotated by multiple experts, and the numbers reflects the sum over all the non-overlapping batches of articles that the the experts annotated. The 10 overlapping articles ("Common" article set) that were annotated by Firas and all the experts are not included in this table.

→ Question on Annotation Form	T1 Has model % of total (N)	T2 Multivariate model % of models (N)	T3 Has molecular features % of models (N)	T4 Outcome is clinical % of models (N)	T5 Outcome is biological % of models (N)
Firas-1 (500)	65% (325)	96% (312)	85% (275)	48% (157)	74% (240)
Firas-2 (200)	87% (174)	98% (171)	52% (91)	93% (159)	40% (58)
Experts-1 (340)	52% (177)	85% (150)	75% (132)	46% (81)	65% (115)

Table 6 shows the percentage of articles in Firas-1, Firas-2, and Experts-1 where the corresponding filter question was answered as "yes". Notice that the percentages reported for question 1 ("T1: Does the article describe at least one predictive model?") report the fraction of all papers that were annotated as "yes" in each dataset. According to the workflow indicated in the annotation form, questions 2-5 (T2-T5) are only answered

if the paper describes a predictive model. The percentages shown for questions T2-T5 report the fraction of papers that describe a predictive model that where annotated as "yes" for these questions.

Notice that the Firas-2 article set (from the breast cancer population) has a high fraction of predictive models and that the majority of those models are models with clinical outcome. Recall that clinical outcome is indicated by a "yes" to question 4 ("T4: Is one of the outcomes a clinical outcome?") More than half (101 of 200) of the articles in Firas-2 were from the journal "Cancer." Many of the articles that were encountered in Firas-2 were typical of epidemiological research i.e. risk or survival analysis using clinical variables in cancer populations.

The fractions reported for the Experts-1 article set are aggregated from all fourteen batches of articles given to experts. There was variability (not shown here) of percentages of answers between the individual batches. In addition to the effect of random sampling, this variability may be due to differences in annotation behavior between the individual experts. Experts with a "conservative" understanding of what constitutes a predictive model will tend to answer "yes" less often on question T1.

Training and Validation of Machine Learning Filters (Aim 2a)

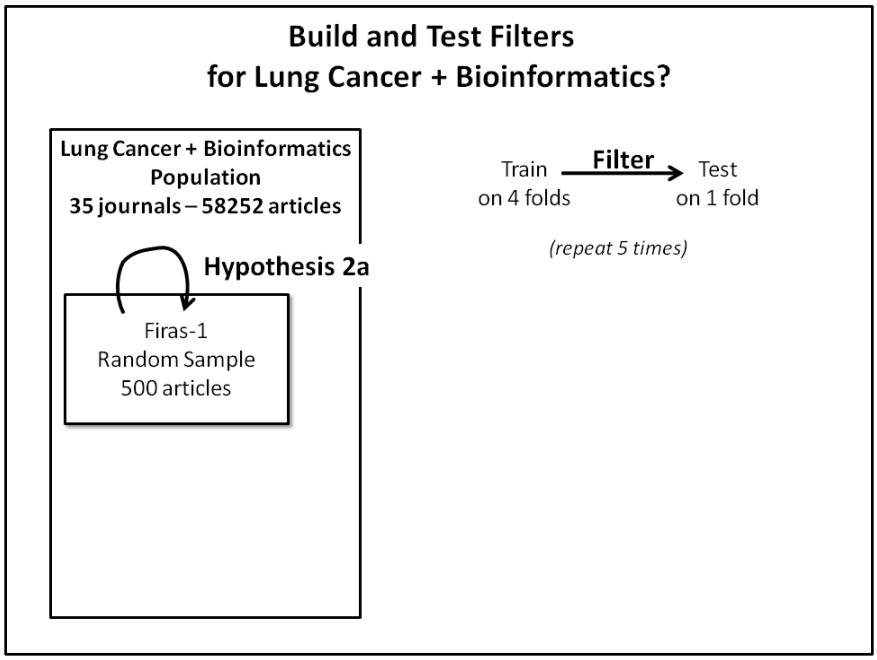
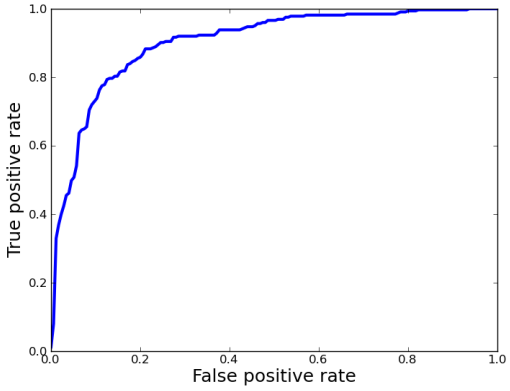


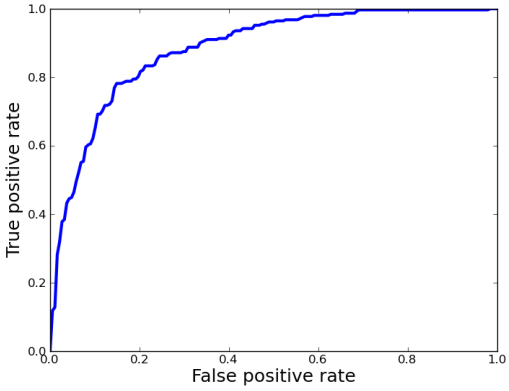
Figure 10 – For the research question in Aim 2a, models were trained and tested via stratified nested 5-fold cross validation using the Firas-1 articles set.

This section reports the result of the first research question: *Can existing or modified feature extraction transformations be used to train machine learning filters that can identify relevant papers from MEDLINE?* The results of the 5-fold cross validation experiment to test whether SVM filters can replicate the annotations by Firas of the Firas-1 article set are shown in Table 7. These results are obtained using the standard feature extraction process (weighting using log-rel frequency with redundancy) described in the methods section. Overall, 19939 features were extracted from the 500 articles in this set. This means that for all 500 articles the sum of unique words in the abstracts, unique words in the title, and unique MeSH headings is 19939.

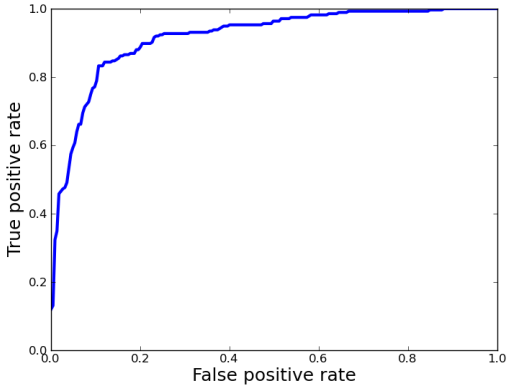
Table – 7: Each cell in this table corresponds to the performance of the SVM filter on each of the filter questions (1-5) for the Firas-1 article set. The average area under ROC (AUC) reported is the average over 5 folds of the stratified nested cross validation experiment using polynomial SVM kernels and feature extraction as described in the methods section.



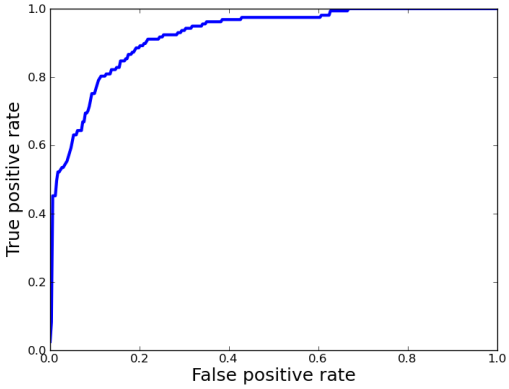
T1: Has a model?
Average AUC = 0.904



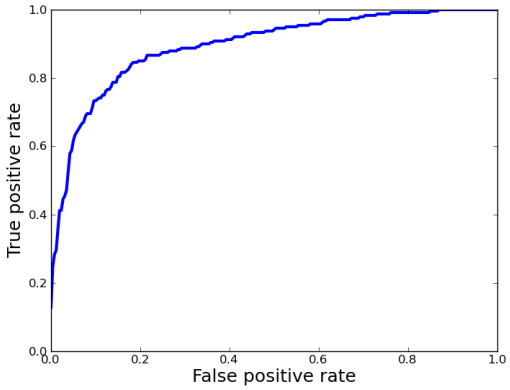
T2: Multivariate Model?
Average AUC = 0.881



T3: Has molecular features?
Average AUC = 0.917



T4: Outcome is clinical?
Average AUC = 0.924



T5: Outcome is biological?
Average AUC = 0.896

Effect of Feature Extraction and Preprocessing

The result of the 5-fold cross validation *using the same folds* but different feature extraction and preprocessing methods before training the SVM models are shown in Table 8. This is for the first question on the annotation form. The first method (log-relative frequency with redundancy vector) is the default method used for the reported results in the previous section.

Table 8 – The effect of different feature extraction and pre-processing methods on the performance of the SVM filter for question 1 on the form. The third column shows the number of features extracted from 500 articles in Firas-1 by including/excluding MeSH terms (column 1) and the use of Porter Stemming to reduce the input space. The fourth column shows the pre-processing step used and the resultant average AUC using the same 5-fold stratified nested cross validation splits.

MeSH	Use Porter Stemming	Number of Features in 500 Articles	Preprocessing	Average AUC for 5 test folds in Firas-1 (Question 1)
Included	N	19939	Log-rel freq with redundancy	0.904
Included	Y	16619	Log-rel freq with redundancy	0.901
Included	N	19939	L2-normalization	0.890
Excluded	N	13337	L2-normalization	0.887
Included	N	19939	L1-normalization	0.881
Excluded	N	13337	L1-normalization	0.892

Analysis of Misclassifications in Aim 2a

Applying an SVM filter to an article means that the SVM filter computes a value called the "decision function" that corresponds to that article. In the case of the filters for

questions 1-5, the output of the decision function can be used to rank the articles based on the predicted likelihood that they will have the answer to the corresponding question as "yes." For each of the test cases within the 5 folds (i.e. 5 x 100 articles per fold) I looked at the ranked list of articles as predicted by the SVM **for the T1 question** and compared that to my annotation value for that question. Table 9 shows this ranking for the top 20 and bottom 20 articles within each fold. The articles (indicated by their PubMed ID) that I classified as not describing a predictive model are shaded red, those that I considered as describing predictive models are shaded green. A lower rank value (top rows) corresponds to a higher decision function value by the SVM filter. Perfect discrimination by the SVM (corresponding to an AUC value of 1) would have occurred if all green cells were segregated at the top from all the red cells in the bottom. Red cells near the top adversely affect the **precision** of this filter (i.e. lower positive predictive value, more false positives in the retrieved set). Green cells near the bottom adversely affect the **recall** of this filter (i.e. lower sensitivity, more false negatives in the retrieved set).

Table 9 – Ranked lists of the 100 cases in each of the test folds based on the value of the decision function for the T1 SVM filter. The top rows are expected to have the value "yes" for T1 by the SVM. Green cells indicate articles that I manually annotated with "yes" for T1. (table best viewed in color)

Rank	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
1	17699846	18003960	18339877	19497995	18172257
2	17545551	17940499	16552436	17372279	17409400
3	17308096	18316570	16823852	17311288	18519766
4	17266042	19289620	19513064	17470860	17906207
5	19190139	17693662	17942905	19147983	17315194
6	18392050	18701494	19088723	18725989	19430494
7	16575786	19151763	18381943	17908961	17052995
8	17578909	16912199	18337602	16707608	19066609
9	17096351	18827607	16899629	19383912	17847021
10	16778098	17135638	19266094	18559598	16505430
11	16639698	19088020	18829558	19138971	18767034
12	16698114	17606972	17483316	17486061	17409416
13	16624828	17158282	18349395	19228733	16818620
14	16709935	16449974	18334633	19431143	17582601
15	19276245	16707424	19384944	17289903	18483267
16	17119046	18648364	18165641	19293795	18765551
17	17093248	18708389	17179993	19541629	17292828
18	17261802	17353899	17045205	19035462	19447898
19	16953234	17483353	16400030	17545531	17763396
20	17110434	18048820	18565887	18483247	19332718
	…	…	…	…	…
81	17932069	18621689	18174223	16772402	19247482
82	18436648	16754871	18304946	16431844	17032674
83	16892060	16672366	17606921	17283341	18508970
84	17412830	18387210	16789820	19399170	17167056
85	19283079	17172439	18987010	19117739	18518950
86	16595561	18808329	16901214	18663013	17581870
87	18562466	19033184	17341495	18089620	19036931
88	19370150	16733546	18203770	16488977	19336412
89	16547201	18424799	17145709	17710141	17572025
90	16832051	18852878	17991681	18344323	16848637
91	19321429	19515936	18211675	18187508	18978014
92	16845086	19369499	18703323	19208138	18042553
93	19010966	18442400	18787685	17277078	18042272
94	17872912	17392332	17425803	16402894	16504085
95	16899490	17059592	17984083	19549335	16789817
96	19535537	19103665	16873487	18697772	18945683
97	17267434	17537824	19063730	18229697	16756676
98	16942624	19158162	16845040	18725927	17691896
99	19008251	19269990	19129210	18387199	17584798
100	16912992	18388142	16817972	16845081	18184684

To understand the filter's limitation I looked at the cells that violated the expected rank order in the top and bottom 20 rows. First I looked at the two red articles at the top part of the list ("false positive"). They are:

PubMed ID	Journal	Title
17940499	Br J Cancer	Weekly epirubicin plus docetaxel as first-line treatment in metastatic breast cancer
16707608	Clin Cancer Res	Effect of cA2 anti-tumor necrosis factor-alpha antibody therapy on hematopoiesis of patients with myelodysplastic syndromes

It can be verified by referring to the annotation guidelines that those two articles should be classified as describing predictive models (T1 = "yes"). They both clearly describe statistical associations between independent variables and outcomes. This can be considered an error on the part of the annotator (Firas).

Then I looked at the 12 green cells at the bottom of the table ("false negative"):

PubMed ID	Journal	Title
16789820	PLoS Comput Biol	Adaptation to different human populations by HIV-1 revealed by codon-based analyses
17872912	Bioinformatics	Graph-based consensus clustering for class discovery from gene expression data
16772402	Nucleic Acids Res	A base pair at the bottom of the anticodon stem is reciprocally preferred for discrimination of cognate tRNAs by Escherichia coli lysyl- and glutaminyl-tRNA synthetases
16431844	Nucleic Acids Res	Oct-2 DNA binding transcription factor: functional consequences of phosphorylation and glycosylation
17710141	PLoS One	Genome dynamics of short oligonucleotides: the example of bacterial DNA uptake enhancing sequences
19549335	BMC Bioinformatics	Filtering genes for cluster and network analysis
18229697	Pac Symp Biocomput	Combining molecular dynamics and machine learning to improve protein function recognition
17032674	Bioinformatics	Large scale data mining approach for gene-specific standardization of microarray gene expression data

18508970	Proc Natl Acad Sci	Synthesis and bioassay of improved mosquito repellents predicted from chemical structure
18518950	BMC Bioinformatics	A simple and robust method for connecting small-molecule drugs using gene-expression signatures
18042553	Bioinformatics	Identification of differentially expressed gene categories in microarray studies using nonparametric multivariate analysis
18042553	BMC Bioinformatics	Identification of DNA-binding proteins using support vector machines and evolutionary profiles

Most of these articles (10 out of 12) were published in the bioinformatics journals. This is a disproportionate representation of the overall number of articles from bioinformatics journals in this article set (96 out of 500, table 4). The annotation guideline lists the following categories as examples of papers that do NOT describe predictive models: statistics papers (including population genetics); bioinformatics methods papers (without reporting clinical or biological experimental results); structural biology (3D structures, binding sites); biotechnology (synthesizing new drug molecules). Some of the papers in this list may fall under these categories yet they were still annotated as "yes" for T1 by the annotator (Firas). These types of article may be indicative of remaining ambiguity in the annotation guideline.

Filter Generalizability to a Different Disease Domain (Aim 2b)

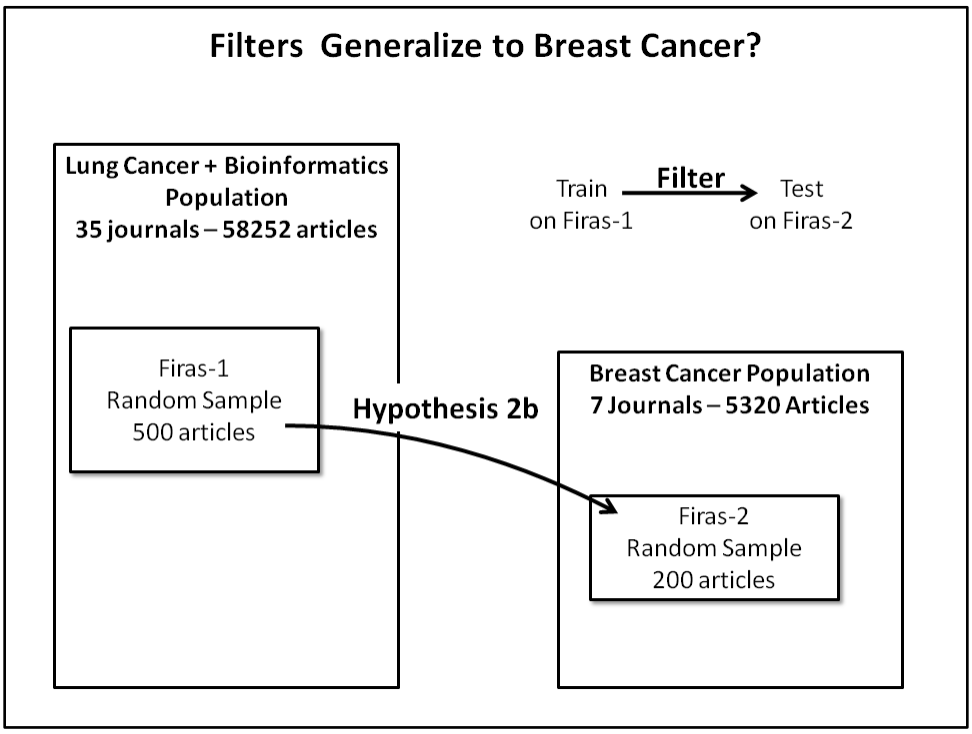


Figure 11 – For the research question in Aim 2b, models were trained (using 5-fold cross validation to optimize degree and cost parameters) using the Firas-1 dataset and tested using the independent Firas-2 dataset.

This section reports the results for the research question: *Can the filters that were trained using the bioinformatics and lung cancer gold standard, and found to have favorable performance identify relevant papers in other domains?* Using 5-fold cross validation and the Firas-1 dataset, SVM models were training using combinations of different cost and degree parameters. The models with optimal parameters were then trained on the entire Firas-1 (lung cancer + bioinformatics) dataset and applied to the independent Firas-2 (breast cancer) dataset. This was repeated for questions 1-5. The AUC was calculated as a measure of the ability of those filters to rank the articles in Firas-2 using the manual annotation as a gold standard. The values of the AUC for each of the T1-T5 questions are shown in table 8. Overall AUC values support the hypothesis

that the SVM filter can rank the articles in a manner similar to the manual annotation. There were 9473 features extracted from the 200 articles in Firas-2. Only 5555 of those features existed in the Firas-1 dataset and thus were recognizable to the SVM filters.

As discussed above, this dataset contains a higher fraction of traditional epidemiological articles describing clinical predictive models and is relatively homogenous. More than half of the articles are from a single journal. Due to the large size and variability of the lung cancer + bioinformatics population relative to the breast cancer population, there may not have been enough power in Firas-1 training samples to build models that can better discriminate articles within this relatively specialized dataset. For example the range of output of the decision function of the SVM for T4 was [-1.06,+1.31] when testing within the Firas-1 cross validation folds. The range of output of the same SVM for T4 in Firas-2 was [-0.64,+0.51].

Table – 10 The results of this experiment to test how well filters can generalize to a different population of MEDLINE articles. The filters were trained on 500 articles (Firas-1 set) in the lung cancer and bioinformatics population. They were then tested on an independent set of 200 articles (Firas-2 set) in the breast cancer population.

→ Question on Annotation Form	T1 Has a model?	T2 Multivariate model?	T3 Has molecular features?	T4 Outcome is clinical?	T5 Outcome is biological?
AUC when testing the filter on Firas-2	0.916	0.896	0.881	0.884	0.917

Filter Generalizability to Annotations by Different Experts (Aim 2c)

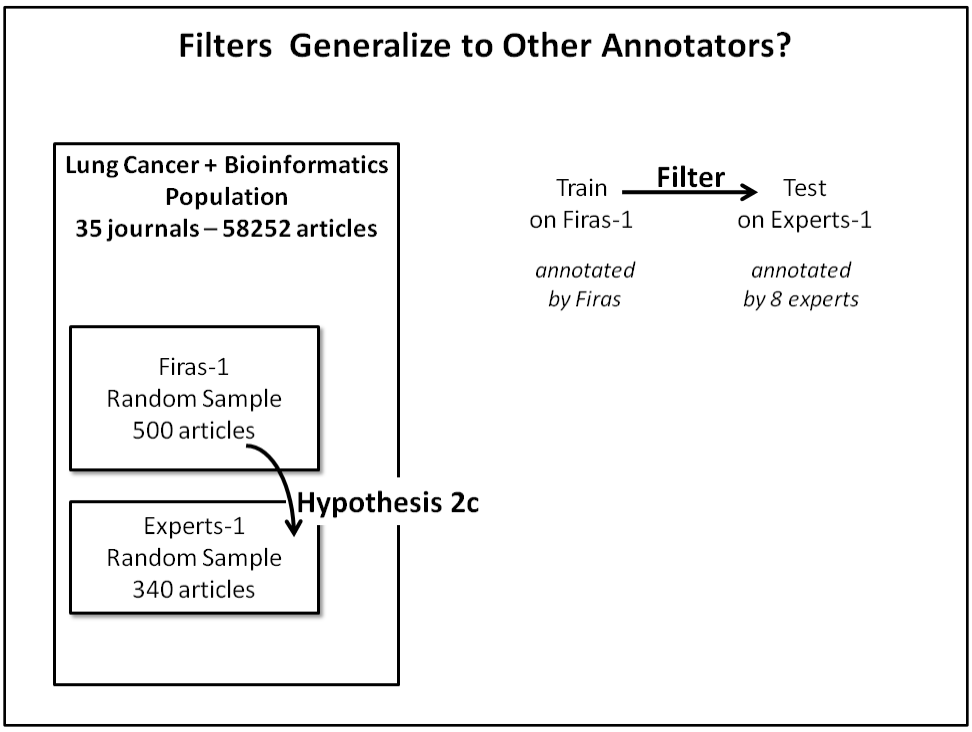


Figure 12 – For the research question in Aim 2c, the same models that were trained for Aim 2b (using the Firas-1 dataset for 5-fold cross validation to optimize degree and cost parameters) were tested using the independent Experts-1 dataset that was annotated by different experts.

This section reports the results for the research question: *can filters trained to identify relevant papers in the domains of bioinformatics and lung cancer using annotation by one expert, identify relevant papers in the same domain as judged by other experts whose annotations were not used to train those filters?* The *same filters* that were trained using the 500 Firas-1 dataset in Aim 2b and described in the previous section were used for this experiment. These filters were tested on an independent set of 340 articles (Experts-1 set) that were annotated by 8 different experts. The experts were given non-overlapping batches of article to annotate according to the annotation form and guidelines as described above. There were 15774 features extracted from the 340 articles in Experts-1, of which 8230 existed in Firas-1 and were thus recognizable to the saved

SVM filters. The results are shown in table 11 for each expert and for the entire pooled Expert-1 dataset. Except for the T4 question, the performance of the SVM filters was less favorable when applied to Experts-1 than when they were applied to Firas-2. The performance of the filters varied for different experts. This will be analyzed in the next section.

Table – 11 The results of the experiment to test how well filters can generalize to articles annotated by experts who did not annotate the training set. The last two rows show the performance of the filters when applied to the entire pool of articles in Experts-1. Removing expert #3 from that pool improves the performance of most filters.

→ Question on Annotation Form	T1 Has a model?	T2 Multivariate model?	T3 Has molecular features?	T4 Outcome is clinical?	T5 Outcome is biological?
Expert #1 (20)	0.944	0.913	0.975	0.817	0.816
Expert #2 (80)	0.885	0.863	0.879	0.977	0.838
Expert #3 (80)*	0.653	0.638	0.623	0.949	0.510
Expert #4 (20)	0.857	0.648	0.827	1.000	0.864
Expert #6 (80)	0.939	0.882	0.936	0.910	0.969
Expert #7 (20)	0.725	0.667	0.485	0.971	0.342
Expert #8 (20)	0.938	0.942	0.935	0.926	0.837
Expert #9 (20)	0.731	0.630	0.727	0.944	0.694
All experts (340)	**0.798**	**0.758**	**0.799**	**0.936**	**0.768**
All experts except #3 (260)	**0.857**	**0.805**	**0.858**	**0.930**	**0.830**

Analysis of Misclassifications and Expert Variability in Aim 2c

The second to last row in table 11 shows the SVM model's predictivity using the pooled dataset, i.e. when treating Experts-1 as one dataset. This assumes that the annotation behavior was the same by the different experts. When applying the SVM filter to each the batches done by separate annotators to rank the articles within that batch, there was an observed variability in the AUC between batches. One stark example is the difference in outcome observed between expert #3 and expert #6 shown in Table 12. Both of these experts volunteered to do three batches of papers. As table 12 shows, when applying the **SVM filter for T1** to rank the articles in their respective batches, there was more segregation of the manual annotation results of expert #6 than of expert #3.

Table 12 - This table highlights an extreme case in the variability of AUC when ranking every batch individually or when pooling articles by expert. The batches in this table are ranked based on the output of the T1 SVM filter. The green cells represent a "yes" annotation by the experts, and the red cells represent a "no" annotation of the article indicated by the PubMed ID.

	Annotator #3			Annotator #6		
Rank	1st Batch	2nd Batch	3rd Batch	1st Batch	2nd Batch	3rd Batch
1	18844223	18682710	17173139	19318480	18204076	17330233
2	18276942	16638863	17210994	17164357	18635524	17653092
3	18491402	16452183	17487844	19291793	17363595	18386818
4	19224839	17053067	17600087	16840742	19336555	19431210
5	17565742	18509179	18167534	16532035	17546598	17372254
6	17409988	17726451	19349548	16769899	17875732	16702377
7	19047288	19383924	19287092	19293260	18641128	19081160
8	19440374	17575155	17893910	17627015	16835325	16619035
9	19002244	18076072	18270592	17360361	16651521	16551849
10	18670317	19243020	17984110	17925449	17121811	19444914
11	17284608	16670771	16767156	17377161	17217525	19359485
12	18648666	17804806	17638882	16501576	16864781	17699815
13	19283069	16537381	17537754	18757739	18216266	18398482
14	18268323	19435903	18499801	19033359	19286565	18346967
15	19264681	18030348	19497884	18369201	17278107	17043219
16	17311100	18231590	18469852	18174226	16407111	17584784
17	18094749	18321995	19533687	18621757	16980979	17409941
18	18427124	16436675	18398474	18524801	16702391	19095792
19	16815972	18836447	17978184	19465379	17675576	19119996
20	16464251	17957241	19380442	18296747	18940870	19156197
21		18523009	17983263		18230720	18779562
22		19365537	16441182		18946033	18032432
23		17940610	16537382		16423899	18632578
24		19228613	18347737		17710132	18716296
25		19541622	18596928		16793924	18794075
26		17151368	18045785		18030325	19179708
27		18927109	17958908		18788908	16500937
28		16707745	19389733		17616981	16545116
29		17166289	17963510		16845103	18784187
30		16844981	17044168		19036790	19075236
AUC (batch)	0.654	0.938	0.655	0.981	0.917	0.958
AUC (pooled)	0.653			0.939		

Upon further examination of the batches provided by expert #3, it is possible to consider that expert's annotation as an outlying case. By inspecting some of the extreme cases of disagreement between the filter and that expert's answers, for the red articles near the top of the table or the green articles near the bottom, it is possible to assume that that expert's adherence to the annotation guidelines deviated from the rest of the

annotators. For example, note the title of article number 17600087 “Let-7 expression defines two differentiation stages of cancer,” and that one of its MeSH headings is “predictive value of tests.” That article was classified by expert #3 as not having a predictive model.

The pooled articles for all of Expert-1 without the annotations of expert #3 are shown in the last row of Table 11. Figure 13 shows the ROC curve for the pooled annotation performance for question 1 with and without expert #3.

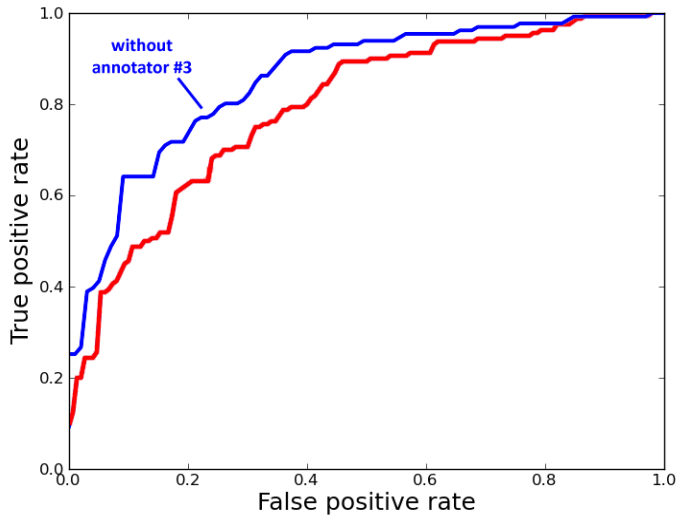


Figure 13 – The performance of the filter for question 1 (“Does the paper describe a predictive model?”) when applied to the pooled Experts-1 dataset with and without Expert #3. (AUC = 0.798 and 0.857 respectively)

Inter-Annotator Agreement

Finally, Table 13 shows the concordance using the Kappa statistic for question 1 (“Does the paper describe at least one predictive model?”) between all the annotators for the 10 articles in the Common article set. Annotator 0 is Firas, annotators 1-9 are the experts. The following are concordance values of note:

- Firas and expert #6 showed high concordance on question 1 annotation. Note also the high AUC for the question 1 SVM filter when tested on expert #6's 80 pooled articles in Table 12.
- Perfect concordance between expert #3 and expert #7 on question 1. Both of those experts have provided annotation batches with conservative annotation of models. The batch provided by expert #7 had only 5 articles out of 30 where question 1 was answered "yes". Expert #3's second batch (Table 12) also had 5 articles out of 30 where question 1 was answered "yes"
- High concordance between experts #1 and #9 also for the question 1. Both experts seem to agree on whether articles included predictive models. Both annotators have formal training in epidemiology.

Table 13 – Concordance values between the different annotators for the T1 question: "does the article describe at least one predictive model?" Annotator "0" is Firas. The experts used for the Expert-1 article set are annotators 1-9

Kappa for T1	0	1	2	3	4	6	7	8	9
0		0.60	0.40	0.00	0.20	0.80	0.00	0.40	0.40
1	0.60		0.07	0.21	0.09	0.44	0.21	0.19	0.74
2	0.40	0.07		0.29	0.29	0.17	0.29	0.09	0.29
3	0.00	0.21	0.29		0.05	-0.07	1.00	0.12	0.38
4	0.20	0.09	0.29	0.05		0.29	0.05	0.62	0.05
6	0.80	0.44	0.17	-0.07	0.29		-0.07	0.55	0.29
7	0.00	0.21	0.29	1.00	0.05	-0.07		0.12	0.38
8	0.40	0.19	0.09	0.12	0.62	0.55	0.12		0.12
9	0.40	0.74	0.29	0.38	0.05	0.29	0.38	0.12	
Mean	0.35	0.32	0.23	0.25	0.20	0.30	0.25	0.27	0.33

Discussion

Summary of Results

The results reported in this chapter are centered on building machine learning filters for finding molecular medicine articles in MEDLINE that can be used to populate a knowledgebase of clinical bioinformatics models. This began with an operational definition of "relevant papers." This definition is part of an annotation guideline document that was based on the semantic analysis described in Chapter III and that was iteratively refined using feedback from experts and pilot annotation of articles and models. As part of this operational definition, human annotators answered 5 questions about the content of given articles that pertain to clinical bioinformatics predictive models. The human annotations of different article sets were used to train SVM-based machine learning filters. Commonly used feature extraction and pre-processing steps were used in the development of these filters. The first article set that I annotated consisted of 500 articles from the domains of bioinformatics and lung cancer. Validation of the machine learning filters via 5-fold cross validation showed very good predictivity on this dataset. The performance of the filters was only slightly reduced when the feature extraction and/or pre-processing steps were modified to exclude MeSH terms, to utilize word stemming, or to forego term redundancy weighting. This first dataset was used to train filters that were saved and applied to other datasets. The second dataset that I annotated consisted of 200 articles in a separate set of journals from the domain of breast cancer. The saved filters also showed very good predictivity on that dataset and therefore generalizability to a different domain. The third dataset consisted of 340 independent

articles that were manually annotated by a group of experts. The saved filters from the first dataset (annotated by me) showed good predictivity using this dataset and therefore generalizability to annotations by other experts. The filters' ability to discriminate relevant articles was variable across the different annotators. Specifically, marked improvement of the performance was observed when one specific annotator was excluded from analysis. Examination of that expert's annotations suggests that he may be an outlier in his annotation behavior from the rest of the experts. Inter-annotator agreement using the Kappa statistic provided a partial explanation of the variability in model performance along different experts.

Structural Limitations of Training Machine Learning Filters

The main structural limitation of the proposed methodology for building the filters is that the information used in the article representation for machine learning is only obtained from the MEDLINE record and not from the full text of the article. The annotators, on the other hand, relied on the full text of the article. Despite this limitation the filters have shown very good predictivity for the manual annotations for these datasets even when discarding the MeSH terms in the MEDLINE record. Human annotators typically add MeSH terms to the MEDLINE record of an article after assessing that article's content.

Under-Representation of Model-Describing Articles from Specific Domains

Analysis of misclassified articles in the article sets that I annotated for Aim 2a (e.g. "green" articles at the bottom of Table 9 and the corresponding text in that

subsection) highlighted the possible limitation of having a relatively small number of specialized types of relevant articles in the training dataset. The majority of the "green" articles at the bottom of Table 9 are from the bioinformatics journal set. Some of those bioinformatics articles were found, upon their re-examination in light of a strict interpretation of annotation guidelines, to be erroneously labeled "positive" for question 1 i.e. labeled as describing *Models* when they actually do not under the guideline's definition (see next subsection). However, other articles like article 18518950 ("A simple and robust method for connecting small-molecule drugs using gene-expression signatures" *BMC Bioinformatics*) were true "positive" articles that were given a low ("negative") decision function by the SVM filter. Such "positive" articles in the bioinformatics journal set may describe *Models* using terms that are not similar to the terms that describe *Models* in the lung cancer journal set (e.g. due to difference in communication style used by researchers in different domains). The majority of articles in the overall population are obtained from the lung cancer journal set (47602 lung cancer articles out of a total of 58252). The prevalence of positive articles from the bioinformatics literature is therefore lower and may not provide sufficient statistical power for training of the SVM filters to learn their characteristics.

A similar under-representation of the "positive" models from a specialized type of articles was also observed in the Firas-2 (breast cancer) dataset. As shown in Table 5, over half of the articles in Firas-2 (101/200) were obtained from the journal "Cancer." It seems the breast cancer dataset includes mostly traditional epidemiological *Models* that relate clinical outcomes to purely clinical independent variables. Recall from Table 6 that the percentage of articles annotated as having molecular features in Firas-1 and Expert-1

(lung cancer + bioinformatics) was higher than the percentage of articles in Firas-2 (breast cancer) - 85%, 75% and 52% respectively. Conversely, there were fewer articles annotated as having a clinical outcome in Firas-1 and Expert-1 than in Firas-2 – 48%, 46%, and 93% respectively. This difference was further illustrated by comparing the output of the SVM filters using a randomly selected set of 10,000 articles from the lung cancer + bioinformatics population and a randomly selected set of 1,000 articles from the breast cancer population (Figures 14-16).

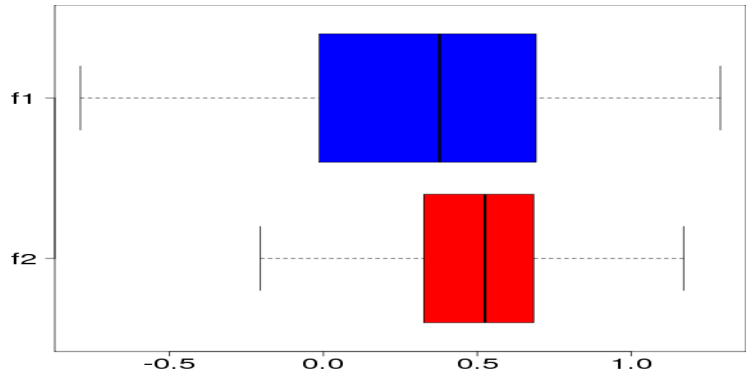


Figure 14 – The distribution of the SVM decision function for question 1 ("Does the paper describe at least one predictive model?") for 10,000 randomly selected articles from the lung cancer + bioinformatics population (top) and for 1,000 randomly selected articles from the breast cancer population (bottom).

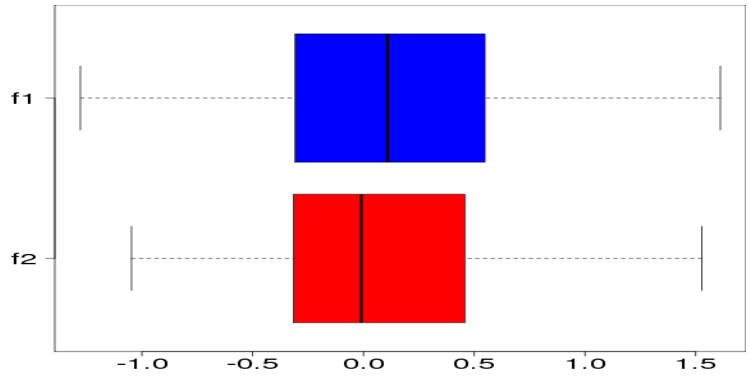


Figure 15 – The distribution of the SVM decision function for question 3 ("Do the model's independent variables contain molecular measurements?") for the same samples used in Figure 14.

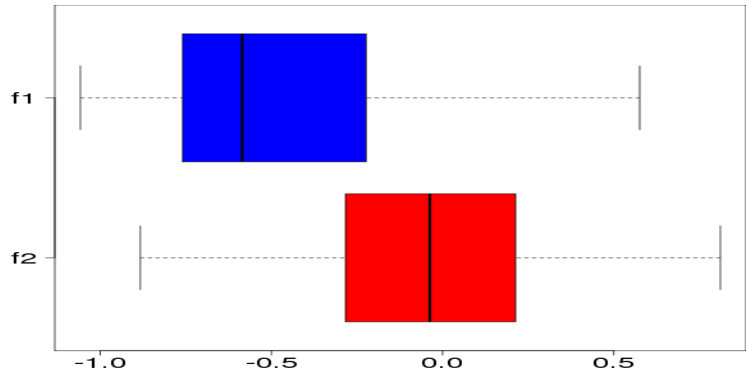


Figure 16 – The distribution of the SVM decision function for question 4 ("The model has a clinical outcome?") for the random samples used in Figure 14.

Guideline Ambiguity and Variability by Expert Annotators

The annotation guidelines provided an operational manual for the annotation of individual articles. As described in the methods section it was iteratively refined and disambiguated using pilot annotation and feedback from domain experts. The misclassified articles (for question 1) used for testing Aim 2a were analyzed in the "Analysis of Misclassification in Aim 2a" subsection. It was shown that upon re-examination of the two false positives ("red" articles near the top of Table 9), both of these negative articles were actually found to describe a predictive model per the annotation guideline, and should therefore be considered true positive classifications by the SVM filter. Furthermore, some of the false negatives (the "green" articles near the bottom of Table 9) should be considered as true negatives according to the annotation guideline. For example, the articles 16431844 ("Oct-2 DNA binding transcription factor: functional consequences of phosphorylation and glycosylation" *Nucleic Acid Res*) is a structural biology paper that does not describe a model as defined in the annotation guideline. Similarly, the results in the subsection "Analysis of Misclassifications and Expert Variability in Aim 2c" found that the annotation behavior of expert #3 may be an

outlier to the behavior of the rest of the annotators. Experts #3 and #7 seem to assign the answer "yes" less frequently within their batches than their counterparts. Assuming that the annotators (including Firas) are consistently and faithfully annotating articles according to their understanding of the annotation guidelines, their annotation behavior will diverge from one another if their cognitive interpretations of the guidelines are not the same. The kappa concordance based on the 10 overlapping articles in the Common dataset that was annotated by all annotators, including Firas, only provides a partial explanation of the inter-annotator agreement. Further analysis of inter-annotator agreement and of potentially diverging cognitive interpretations of the annotation guideline will be deferred to the next chapter after the rest of the annotation guideline and manual annotations by experts are discussed.

Conclusion

In this chapter, common machine learning text classification techniques were applied to the problem of finding MEDLINE articles that are relevant to the information retrieval framework described in this dissertation. These filters were validated using manual annotation and were found to have very good predictivity using the AUC metric. The predictivity was minimally affected when different feature extraction techniques were used (including removing the manually assigned MeSH terms in the MEDLINE record). Also, the filters' predictivity was found to successfully generalize to articles in another disease domain as well as to articles that were annotated by a different set of experts. These filters are promising scalable techniques for the problem of large-scale retrieval of relevant articles to populate the framework's knowledgebase. The results

found in this chapter point to the importance of the development of a clear operational definition of the semantic entities that define clinical bioinformatics predictive models. Specifically, it is important that there exists a clear set of annotation instructions that can be interpreted and applied in consistent manner by human annotators. This topic will be explored further in the next chapter.

CHAPTER V

SCALABLE SEMANTIC ANNOTATION OF MOLECULAR MEDICINE PUBLICATIONS USING MACHINE LEARNING CLASSIFIERS

Introduction

This chapter describes a scalable machine-learning-based approach for semantic annotation of molecular medicine publications. Recall that the information retrieval framework proposes an annotation (and indexing) scheme of models and papers that describe these models. This annotation scheme, which was the result of the semantic analysis of this domain in chapter III, basically annotates clinical bioinformatics *Models* by associating them with a clinical bioinformatics *Context*. The primary source of information about *Models* will be from published MEDLINE articles. The previous chapter described an operational definition of "relevant papers" – MEDLINE papers that describe clinical bioinformatics predictive *Models* – as well as a scalable machine-learning-based approach for finding these relevant papers. This chapter will carry this work forward by refining the definition of clinical bioinformatics *Context* and by investigating the performance of a scalable machine-learning-based approach for extracting from relevant articles the semantic attributes that can be used to annotate the *Models* described wherein. Figure 17 illustrates the work described in this chapter (Aim 3 of this dissertation) within this information retrieval framework.

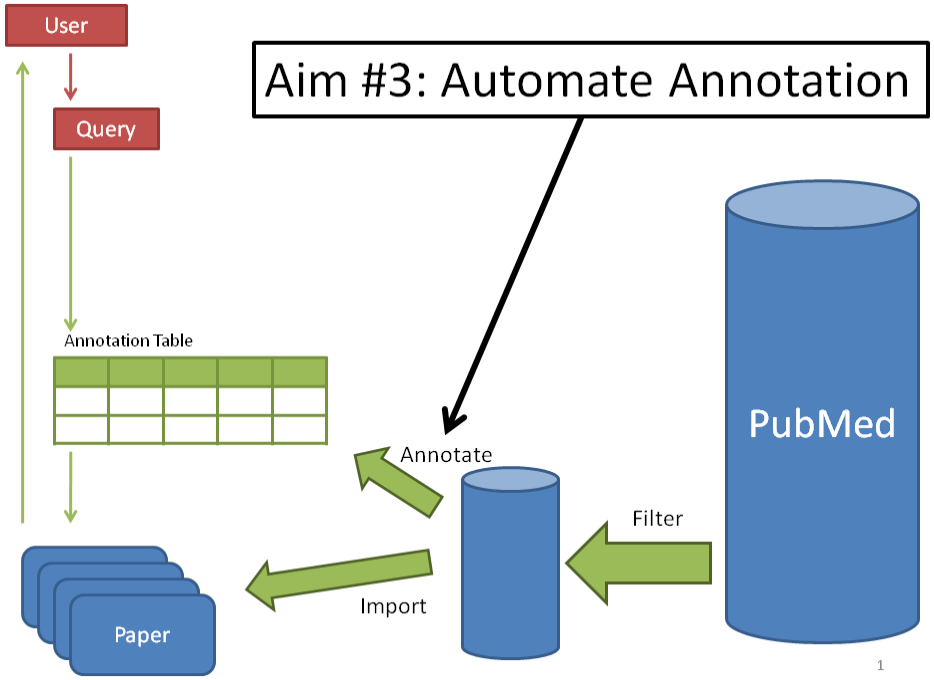


Figure 17 – The third aim of the PhD dissertation is to build and evaluate automated or semi-automated methods for annotating and indexing relevant papers within knowledgebase that will support the information retrieval framework.

Refinement and Operational Definition of the Annotation Scheme

According to the framework described in Chapter III, clinical bioinformatics predictive *Models* and related objects in this domain should be annotated using a clinical bioinformatics *Context*. A *Context* is a tuple that annotates a *Model* (or a *Paper* describing the *Model*) along four dimensions: *Disease*, *Population*, *Modality*, and *Purpose*. Objects will be indexed based on the values of their *Context* annotation. The annotation table shown in Figure 17 shows a conceptual representation of how this indexing scheme supports the overall framework. As mentioned earlier in this dissertation, this indexing scheme was based on the semantic analysis of a specialized and relatively small set of papers that were obtained via a focused ad hoc search of MEDLINE. As more articles were annotated to construct the datasets used for the

research described in the previous chapter and in this chapter, the a priori definition of the attributes that constitute a clinical bioinformatics *Context* were found to be deficient. Specifically, they did not provide enough guidance to annotate many of the new instances of *Models* that were encountered. An operational definition of these attributes is essential for helping the annotator (initially myself) make consistent annotations and for communicating the annotation guidelines to other annotators who are also expected to make consistent annotations. As mentioned in the previous chapter, an annotation guideline document (Appendix B) and an associated annotation form (Appendix C) were created to codify the annotation of models described in MEDLINE articles. This annotation guideline was iteratively refined based on the experience of annotating more articles and based on discussion with domain experts and other annotators. The part of the guideline that defined "relevant papers" – MEDLINE articles that describe a predictive *Model* – was discussed in the previous chapter. This chapter will discuss the remainder of the guideline which provided operational definitions of the *Context* attributes used to annotate the relevant papers based on the *Models* that they described.

Automated or Semi-Automated Annotation Methods

This chapter describes the work done for Aim 3 of this dissertation that is to build and evaluate reproducible scalable automated or semi-automated methods for annotating and indexing papers for the supporting knowledgebase of the information retrieval framework. As discussed earlier, manual annotation of the potentially large number of articles that describe clinical bioinformatics predictive models is a tedious task that can be avoided by using scalable automated or semi-automated methods. The approach that

will be used will rely on machine learning, specifically SVM text classification. The methods will be very similar to those used in the previous chapter. There is, however, a conceptual difference in how the machine learning classifier is used to solve this problem. In the context of the automated "filtration" the classifier will be used to assign a positive or negative label to the document that determines ("filters") its suitability for annotation, essentially acting as a retrieval agent for articles from a larger dataset (MEDLINE). In the context of automated "annotation" the classifier is applied to a set of papers with the implicit assumption that these papers are suitable for annotation. The purpose of an annotation classifier is to compute a binary decision: whether or not the semantic attribute that is associated with this classifier is true for this paper. For example, one of the attributes that can be assigned to a *Paper* that describes a clinical *Model* is whether the *Purpose* (one of the dimensions that constitute that *Model*'s clinical bioinformatics *Context*) of this model is to "predict prognosis associated with a specific treatment." An associated machine learning classifier can be used to compute whether this attribute assignment (and the truth of the associated semantic assertion regarding the purpose of the *Model*) is true or not. This operation actually mirrors the annotation process that is performed by human annotators. Recall that manual annotation of MEDLINE articles is done using an annotation form (Appendix C) in which the annotators circle the attribute values that they believe are true for the given article. Finally recall also that, as described in the previous chapter, using machine learning classifiers for MEDLINE records requires the transformation of free unstructured text into numerical features that can be used by the classifiers to compute a given paper's

classification and that a set of commonly used feature extraction and pre-processing steps can be used to achieve this transformation.

Based on these observations, the first research question under this aim is:

- *Can existing or modified feature extraction transformations be used to train text classifiers that can replicate human semantic annotation of the gold standard?(Aim 3a)*

This experimental approach will be very similar to that which was followed in Aim 2a. Multiple SVM classifiers will be used to correspond to the different attribute assignments in the annotation/indexing scheme. The performance of the different classifiers will be evaluated using 5-fold cross validation using a gold standard dataset of manually annotated articles selected from the domains of bioinformatics and lung cancer (the Firas-1 dataset used in the previous chapter). The same feature extraction methods used in the last chapter will be used.

The attributes describing the clinical bioinformatics *Context* may be more semantically complex than the concepts that were used to determine paper "relevance" in Aim 2. For example, semantic annotation may rely on more specialized and granular biomedical concepts associated with the different types of clinical outcome or the multitude of molecular biology concepts associated with molecular assays described by these papers. Therefore the performance of the annotation classifiers may be enhanced by adding natural language processing (NLP) techniques to the feature extraction transformation. NLP may add informative features to the articles in the dataset by detecting the presence of complex medical concepts within their MEDLINE record. The second research question for this aim is:

- *Will modifying the feature extraction transformations used for training semantic classifiers in Aim 3a to include natural language processing (NLP) techniques alter their performance?(Aim 3b)*

KnowledgeMap, an NLP tool that can extract Unified Medical Language Systems (UMLS) concepts from biomedical text, will be used. This research question will measure the effect of adding the frequency of occurrence of unique UMLS concepts (CUIs) within the MEDLINE record to the set of features used to train and test the machine learning dataset.

The third research question that will be investigated is very similar to that in Aim 2c and is similarly motivated by the fact that the machine learning classifiers used in Aims 3a/b are trained and tested using labels that were assigned by the same human annotator. Therefore, to test the generalizability of these classifiers to annotations assigned by different annotators, the following research question will be assessed:

- *Can text classifiers trained for semantic annotation of relevant papers in the domains of bioinformatics and lung cancer using annotation by one expert, replicate the semantic annotation of independent papers in the same domain by other experts?(Aim 3c)*

Methods

Annotation Guideline

The final versions of the annotation guideline (Appendix B) and annotation form (Appendix C) were the result of multiple iterations of refinement. During pilot annotations of articles, I codified the annotation decisions that I was making. When encountering papers describing predictive models whose attributes were not clearly defined using previously codified guidelines, the guideline was updated to reflect the new instances. The annotation guideline and form were also modified based on feedback from experts, obtained for example when presenting this project to Dr. Massion's lab group.

The main structural deviation from the original annotation scheme relates to the annotation of *Papers* that describe more than one predictive *Model*. It was originally envisioned that a *Paper* that describes more than one predictive *Model* will be annotated using a set of all the clinical bioinformatics *Contexts* describing these individual models. Recall that a *Context* of a *Model* is an ordered tuple of annotations describing that *Model*'s *Disease*, *Population*, *Modality* and *Purpose*. In the current guideline, the attributes from all the *Contexts* of all the *Models* described in a given paper are indicated on the annotation form (i.e. without specifying an ordered tuple relationship). This modification leads to loss of information about the individual *Models* within such (multi-model) papers because the set of applicable *Contexts* (when explicitly annotated) is smaller than the Cartesian product of all attributes circled on the form. This modification was made because it allowed for significant practical improvement in the time of manual

annotations and because annotators who were approached to conduct initial pilot annotations showed wide variability in individual tuple assignment.

The following are other notable modifications from the original annotation scheme to the annotation scheme used for this study. For a full description of the annotation attributes, please see appendix B:

1. The form provided freedom to the annotators to specify the applicable *Disease* using free text (to account for the large number and granularity of disease states that are expected from the articles in this dataset) as opposed to the original canonical list of diseases.
2. The *Population* is now explicitly referred to as "Biologic Sample." The annotators are asked to identify the source of the biological sample that was used for the molecular assay. In addition to "Human," "Cell Line," and "Animal," the new value "Pathogen" was added to refer to molecular predictive models that rely on molecular data obtained from pathogens such as viral genome or protein coat information.
3. The *Modality* is now explicitly referred to as "Type of Assay." Feedback from the experts overwhelmingly indicated that the original values of "Genetic," "Genomic," and "Proteomic" were ambiguous concepts. The categories were modified to refer to the type of biological molecule measured by the molecular assay and were replaced with a new set: DNA, RNA, or Protein. For example: models that rely on molecular measurements obtained using Northern Blot, Southern Blot, or Western Blot assays would be annotated as targeting "RNA," "DNA," and "Protein" respectively. Methylation assays and other assays that

measure epigenetic regulation were considered under "DNA". Assays that can detect post-translational modifications of proteins such as Eastern Blot were considered as "Protein."

4. The original values for *Purpose* (type of clinical outcome) were changed from "Diagnosis," "Prognosis with no treatment," "Prognosis with one treatment arm," and "Prognosis with more than one treatment arm" to the following values: "Diagnosis," "Risk Assessment," "Prognosis Treatment Unspecified" and "Prognosis: Treatment Specified" as described in Appendix B.

Table 14 – The 11 machine learning annotation classifiers that were trained and validated for Aim 3. They correspond to structured attributes within the semantic annotation scheme (i.e. the questions at the bottom of the annotation form in Appendix C)

Classifier Name	Context Dimension	Attribute Present (True/False)
BS1	Population / "Biologic Sample"	Human
BS2		Animal
BS3		Cell Line
BS4		Pathogen
A1	Modality / "Type of Assay"	DNA
A2		RNA
A3		Protein
CP1	Purpose / "Clinical Purpose"	Diagnosis
CP2		Risk Assessment
CP3		Prognosis: treatment unspecified
CP4		Prognosis: treatment specified

Machine Learning Annotation Classifiers

Machine learning classifiers were used to mirror the manual annotation process. Specifically, 11 machine learning classifiers were constructed and evaluated throughout this chapter. These classifiers correspond to the structured components of the annotation scheme at the bottom of the manual annotation form (Appendix C). The semantic attributes and the associated machine learning classifiers are shown in Table 14.

Dataset Construction

The same article sets Firas-1 and Experts-1 that were used in Chapter IV were used for Aim 3 experiments. The procedure of obtaining these article sets is described in chapter IV. These two non-overlapping article sets were obtained by random sampling from the lung cancer + bioinformatics population. Recall from the last chapter (Table 6) that based on manual annotation, the fractions of articles in these article sets that described a predictive *Model* were 65% and 52% for Firas-1 and Experts-1 respectively. Furthermore, not all of the articles that did describe a predictive model relied on molecular features or described a clinical outcome (85% and 75% of models had molecular features; 48% and 46% of the models had a clinical outcome; Table 6). The sparseness of the article sets with respect to "relevant papers" that are amenable for full annotation may limit these article sets' ability to validate the machine learning annotation classifiers.

To enhance the validation of the machine learning annotation classifiers, an "enriched" article set, Experts-2, was collected and annotated by experts. This article set was deliberately sampled in a way to increase the likelihood that it will contain "relevant

articles". It was constructed as follows (see also Figure 18): I used the machine learning filter that corresponds to question 1 on the form ("Does the article describe at least one predictive model?") and that was validated in the previous chapter. I applied this filter to rank a random set of 1000 articles in the lung cancer + bioinformatics population (non-overlapping with neither Firas-1 nor Experts-1). The top 300 articles were selected and divided into batches. As an informal validation, I manually inspected 60 out of the top 300 articles and found all 60 articles to actually describe a predictive model. Six out of the 8 experts who annotated Experts-1 agreed to further annotate 220 articles in this enriched (Experts-2) article set. Table 15 contains the source, size and usage of the different article sets used in this chapter.

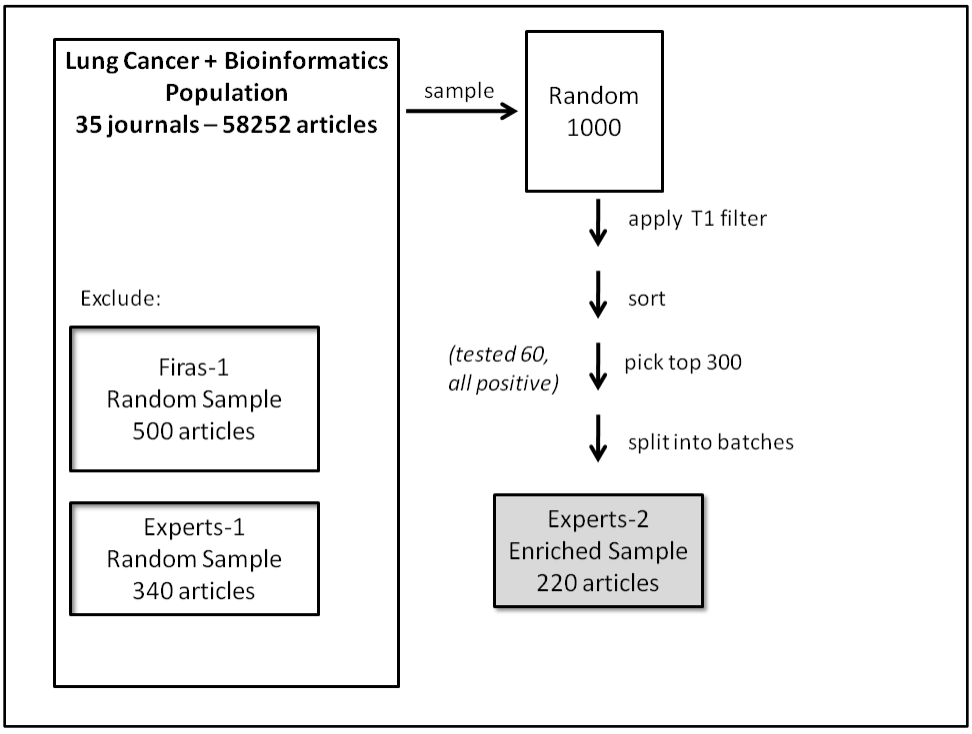


Figure 18 – The procedure used to collect an "enriched" validation dataset Experts-2. After excluding Firas-1 and Experts-1 from the lung cancer + bioinformatics population, a random sample of 1000 articles was obtained. The filter for question 1 ("Does the article describe at least one predictive model?") that was validated in the previous chapter was then applied to the 1000 articles. The top 300 ranking articles per that filter's decision function were selected and split into batches.

Table 15 – Source, size and function of the three main datasets used for Aim 3. Firas-1 and Experts-1 were the same ones used in Aim 2. Experts-2 was constructed specifically for Aim 3. The Common dataset includes 10/15 articles (depending on annotator) that were annotated by Firas and all experts and is primarily for evaluating inter-annotator agreement.

Article Set Name	Annotator	Size	Baseline Population	Used for Aim (Training / Testing)
Firas-1	Firas	500	lung cancer + bioinformatics	Aim 3a (Train+Test) Aim 3b (Train+Test) Aim 3c (Train)
Experts-1	Multiple Experts (8)	340	lung cancer + bioinformatics[†]	Aim 3c (Test)
Experts-2	Multiple Experts (6)	220	lung cancer + bioinformatics	Aim 3c (Test)
Common	Firas + Experts	10+5	lung cancer + bioinformatics	(Test inter-anntotor agreement)

Expert Annotation

The annotation of Experts-1 article set was described in the last chapter. The same experts were approached and asked to annotate more batches of articles for the Experts-2 article set. The annotation of this article set was funded via a $2000 voucher (VR#1017) from the Vanderbilt Institute for Clinical and Translational Research (VICTR). The subjects were informed that this was an "enriched" article set, meaning that it contained a higher fraction of papers describing predictive *Models*. The expected fraction of articles describing predictive *Models* (100% according to *my* estimation) was *not* revealed to them to avoid interfering with their judgment of whether a given paper describes a predictive model or not. The subjects were asked to annotate batches of 25 papers over a period of few weeks at a higher compensation rate per batch than for Experts-1

($200/batch). This is because papers that describe predictive models require significantly more time and effort than those that do not. Some experts were able to provide annotation for two batches. Similar to the procedure followed for Experts-1, the first 5 papers of the first batch given to every subject were identical for all participants. These common papers and their annotations were added to the "Common" article set for analysis of inter-annotator agreement. The IRB (exemption) study number is 100576 (same as that used during annotation of Experts-1).

Document Representation for Machine Learning Including NLP Output

The same procedures used in the last chapter for formatting articles for learning by machine learning text classifiers were used in this chapter. The methods section of the last chapter has a detailed discussion of these transformations which include permutations of the following methods: Porter Stemming (yes/no), MeSH term inclusion (yes/no), and one of the following normalization and weighting methods (Log-relative frequency with redundancy weighting / L1-normalization / L2-normalization). Additional feature construction using UMLS CUIs detected in the title or abstract of MEDLINE records was done using KnowledgeMap[128,129] (KM), an NLP-based concept extraction tool developed at Vanderbilt by Dr Josh Denny and others. Part of the output of KM is a list of UMLS Concepts (CUIs) sorted by decreasing order of occurance within the document (see Figure 19). The unique CUIs were treated as terms (features) and their frequency was added to the appropriate columns in the dataset.

Antibody to human leukocyte antigen triggers endothelial exocytosis.
Proc Natl Acad Sci U S A. 2007 Jan 23;104(4):1301-6.

```
Total concepts: 108
Unique concepts: 56

3241|Antibody (substance)|13
19721|HLA antigen|8
42971|von Willebrand factor products|4
86418|Humans|4
21031|Fab Immunoglobulins|3
21368|Inflammation|3
14257|Endothelium (body structure)|3
205177|Active (qualifier value)|3
40300|Tissues|2
225336|Endothelial Cells|2
134835|Lymphocyte antigen CD62|2
1444748|
181078|
1283071|Release|2
25914|House mice|1
1522240|Process (observable entity)|1
439857|Dependence|1
1704419|Effective|1
1534709|Split|1
1280500|Effect (qualifier value)|1
699748|Pathogenesis|1
678227|Causing|1
175677|Injury|1
3320|Antigen|1
993609|Granules Drug Form|1
413393|Causes of injury and poisoning|1
443302|Several|1
26802|Murids|1
449719|Part|1
547040|Slight|1
444706|Measured|1
10957|Damage (morphologic abnormality)|1
178602|Dosages|1
40732|transplantation|1
345468|Transplanted organ rejection|1
```

Frequency

Concept Name

Concept ID (CUI)

Figure 19 – Part of the output that results from applying KnowledgeMap to the the title and abstract obtained from a MEDLINE record (in this case for the paper with PMID = 17229850). KM sorts the unique concepts detected in the record by decreasing order of their frequency of occurrence within this record.

Machine Learning Training and Validation

The same learning algorithm that was used in the previous chapter, SVM models, was also used for Aim 3 in this chapter. The methods section in the previous chapter describes the training of SVM models and the method used (AUC) to measure their performance. Recall that in the previous chapter, 5 SVM classifiers were used. They correspond to the 5 "filtration" questions described in Table 2. For Aim 2a, the predictivity of the SVM models was tested using 5-fold stratified nested cross validation using the Firas-1 dataset. For Aims 2b and 2c, the models were trained using the Firas-1 dataset by using cross validation to optimize model parameters then using the best parameters to train on the entire Firas-1 dataset. The models were then saved in files. The models were loaded and tested on the independent datasets Firas-2 (Aim 2b) and Experts-1 (Aim 2c) to test their generalizability.

In this chapter (Aim 3), a very similar training and validation approach was followed. Eleven SVM classifiers were used. Their target features (predicted class) correspond to the 11 structured components of the annotation scheme described in table 14. For each of the 11 SVM annotation classifiers, 5-fold cross validation was used to measure their predictivity for the research questions in Aims 3a/3b using the Firas-1 dataset and *the same cross validation folds used for Aim 2a.* This allows for consistent comparison of the relative predicitivity of all 11 SVM annotation classifiers (as well as the 5 SVM filters). For Aim 3c, the 11 SVM classifiers where trained using the Firas-1 dataset (also using cross validation to optimize for the model parameters) and the resultant models were saved to disk. Their generalizability to annotations by other experts was tested by loading each of the saved classifiers and applying that classifier to a

combined dataset containing annotations from Experts-1 and Experts-2. Additionally some of the saved classifiers were applied to an "enriched only" dataset i.e. Experts-2 without Experts-1.

Large Scale Application of the SVM Classifiers

As a practical test of the scalability of using the machine learning classifiers that have been used so far, a large article set was compiled from both the lung cancer + bioinformatics and the breast cancer populations using the following procedure. The following sets of articles were excluded from the lung cancer + bioinformatics population in MEDLINE (58,252 articles): Firas-1 (500 articles), Experts-1 (340 articles), and Experts-2 (1000 articles originally sampled before enrichment). Ten thousand articles were sampled from the remaining articles in the population. Similarly, the articles in Firas-2 (200 articles) were excluded from the breast cancer population (5,320 articles) and 1000 articles were randomly sampled from the remaining articles in the population. The resulting pool of 11,000 articles was prepared for machine learning using the same methods described above. The SVM classifiers that were trained for Aims 2 (5 SVM "filters" classifier) and 3 (11 SVM "annotator" classifiers) were loaded from disk and applied to all the articles in this pool. The output of these classifiers consists of a computed value by each SVM classifier for each article in this pooled dataset. The outcome of this procedure is described in the results section below.

Results

Re-Annotation of Firas-1

In the previous chapter during the analysis of the results for Aim 2a, some of the misclassifications by the SVM filters during the 5-fold cross validation using Firas-1 were explained by misannotation and not by misclassification by the SVM filter. For example, it was verified by inspecting the papers and the annotation guideline that some of the "false positives" (highly scored negatively annotated papers) should have been "true positives" (see previous chapter). These results were obtained in August/September 2010. The Firas-1 article set was annotated between April and July 2010. The last major change to the annotation scheme as well as the last update of the annotation guideline occurred in June 2010 shortly before I began recruiting the experts. I attempted to re-visit previously annotated articles in Firas-1 whenever I updated the annotation guideline or scheme; however, the results from last chapter seem to indicate that Firas-1 was not annotated using a fixed and consistent annotation guideline. Furthermore, the first 200 articles were annotated without using the printed form. (The results were entered directly into a spreadsheet.)

Therefore, before resuming the analysis for Aim 3 and since the Firas-1 article set is integral to Aim 3, I wanted to revisit and re-annotate Firas-1 by strictly and consistently using the final versions of the guideline and annotation form. I re-annotated the articles in the same chronological order in which they were originally annotated. After completing the re-annotation, I examined all the major changes that occurred between the pre-September 2011 annotations and the new annotations. I defined a "major

change" as a change in annotation that occurred for questions 1 ("The paper describes a predictive model?"), 3 ("The model has molecular features?"), or 4 ("The model describes a clinical outcome?"), because these questions affect the rest of the annotation form. Table 16 shows the major changes that occurred during the re-annotation process. I re-ran the 5-fold cross validation experiments done for Aim 2a using the same 5-folds to compare the effect of the re-annotation on SVM predictivity. As expected, this re-annotation improved the predictive ability of the retrained SVM classifiers (Table 17).

Table 16 – The major changes in annotation that occurred after re-annotation of the Firas-1 article set using the final versions of annotation guideline and form. Note that most changes in annotation occurred in papers that were annotated early during the first round of annotation before the annotation guideline was finalized.

Papers (in chronological order of original annotation)	Number of Major Changes in Annotation
1 – 100	20
101 – 200	13
201 – 300	12
301 – 400	7
401 – 500	4

Table 17 – The effect of the re-annotation of Firas-1 using the last version of the annotation guideline and form on the N-fold cross validation performance of associated SVM "filter" classifiers

Question on the Annotation Form	Average 5-Fold AUC Pre-September 2010 Annotations	Average 5-fold AUC Post-September 2010 Re-Annotation
T1: Has a model?	0.904	0.933
T2: Multivariate model?	0.881	0.899
T3: Has molecular features?	0.917	0.941
T4: Outcome is clinical?	0.924	0.944
T5: Outcome is biological?	0.896	0.918

Summary of Manual Annotations

This section provides a summary of the outcome of manual annotations for each of the article sets used in Aim 3. Table 18 shows the journal origin of the articles in these datasets. Firas-1 and Experts-1 were randomly sampled from the lung cancer + bioinformatics population and previously used in Aim 2. Experts-2 is the "enriched" dataset that was sampled from the same population using the procedure described in the Methods section. Notice that *only one article in Experts-2 was found in the bioinformatics journals* provided by Dr. Aliferis. Recall that the filter for question 1 was used to "enrich" this article set by selecting papers that describe predictive models. The fact that using this filter yielded a small number of articles from bioinformatics journals is consistent with the discussion in the previous chapter regarding the effect on filter bias of under-representation of articles from specialized domains in the training set. Six of the 8 experts who annotated the batches in Experts-1 agreed to annotate more batches in Experts-2. Figure 20 shows the experts who participated in the annotation of Experts-2 and the allocation of batches.

Table 18 – This table describes the 3 article sets used for Aim 3. There is no overlap between these article sets. All article sets were sampled from the lung cancer + bioinformatics baseline population. The Firas-1 and Expert-1 article sets where used for Aim 2 and were randomly sampled from this population. Experts-2 is an "enriched" set obtained via the procedure described in the Methods section. The numbers refer to the number of articles found in the population or the three article sets from each particular journal. The journals listed were provided by Drs. Massion (23 journals, lung cancer research) and Aliferis (12 journals, shaded, bioinformatics).

Journals (lung cancer + bioinformatics)	Population	Firas-1	Experts-1	Experts-2
Proc Natl Acad Sci U S A	12205	102	66	13
PLoS One	6036	46	32	10
Cancer Res	4675	40	22	42
Nucleic Acids Res	3940	37	34	1
Clin Cancer Res	3083	28	14	26
Int J Cancer	2479	26	19	19
J Clin Oncol	2448	19	15	19
Bioinformatics	2318	23	13	0
BMC Bioinformatics	2227	17	13	0
Oncogene	2001	14	11	15
N Engl J Med	1896	19	13	5
Br J Cancer	1861	16	10	11
Cancer Epidemiol Biomarkers Prev	1515	6	13	12
Am J Pathol	1296	11	6	4
J Clin Invest	1067	12	8	3
Carcinogenesis	975	7	8	8
Am J Respir Crit Care Med	973	10	6	9
Lung Cancer	891	11	6	7
PLoS Comput Biol	774	8	2	0
Nat Genet	746	5	4	1
Mol Cell Proteomics	666	5	4	3
J Thorac Oncol	641	4	6	2
PLoS Med	640	6	4	1
Nat Med	554	5	0	2
J Pathol	537	7	3	6
J Comput Biol	331	3	2	0
Cancer Cell	299	4	2	1
J Biomed Inform	262	0	0	0
IEEE/ACM Trans Comput Biol Bioinform	227	2	2	0
Pac Symp Biocomput	189	1	0	0
Artif Intell Med	175	2	1	0
Cancer Prev Res (Phila Pa)	121	1	0	0
OMICS	88	3	0	0
Brief Bioinform	61	0	1	0
Int J Data Min Bioinform	55	0	0	0
Total	**58252**	**500**	**340**	**220**
lung cancer	**47602**	**404**	**272**	**219**
bioinformatics	**10646**	**96**	**68**	**1**

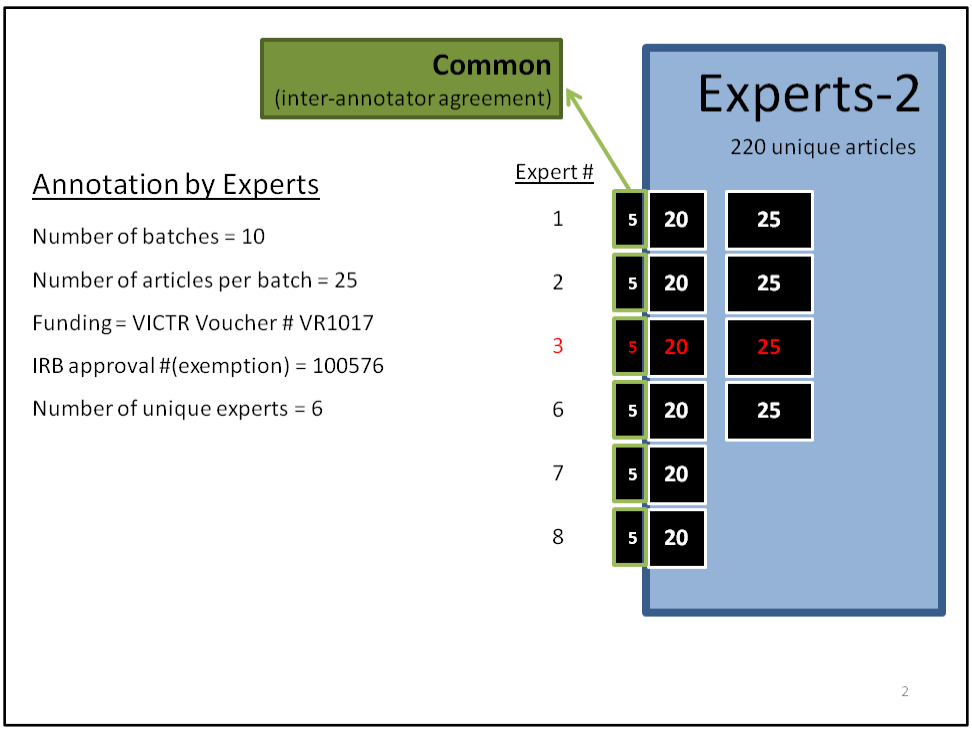


Figure 20 – The source and composition of the Expert-2 article set. Of the original eight experts that annotated Experts-1, 6 agreed to annotate 10 batches of 25 papers. The first 5 papers of the first batch that every expert received were identical and used for analysis of inter-annotator agreement (Added to the Common article set)

The outcomes of the manual annotation for the three article sets are shown in Table 19. All the structured components of the annotation form are summarized in that table. They include the 5 "filter" questions and the 11 structured semantic attributes. The semantic attributes are grouped according to the *Context* dimension that they represent: *Population* ("Biologic Sample," 4 different attribute values), *Modality* ("Type of Assay," 3 different attribute values) and *Purpose* ("Clinical Purpose," 4 different attribute values). The 5 SVM classifiers described in previous chapter and the 11 SVM classifiers described in this chapter correspond to each of those components of the annotation form.

As shown in the previous chapter (see Table 11 and associated discussion), the annotations by experts #3 and #7 were mostly discordant from the classification of the

trained SVM filters. Tables 20 and 21 below summarize their annotation behavior for questions 1, 3 and 4 in comparison to other experts. Their annotation behavior seems to diverge from the annotation behavior of the rest of the experts. These experts seem to label a smaller number of papers as describing predictive *Models*; furthermore, a higher fraction of the *Models* that they identify include clinical outcomes than the *Models* labeled by other annotators. *Assuming their behavior is consistent and faithful to their own personal understanding of predictive Models*, the data in Tables 20 and 21 are consistent with the assumption that these experts consider "having a clinical outcome" a necessary condition for defining a predictive *Model*.

Table 19 – The outcome of manual annotation within each of the article sets used for aim 3. For the "filtration" questions T1-T5, the numbers indicate the fraction (and absolute number) of articles for which the question was answered as "yes." For the rest of questions these numbers indicate the number of papers for which these semantic attributes were circled (indicating the truth of their assignment to this paper). The hierarchical indentation shown in this table mirrors the workflow of the annotation form. For example questions T2-T5 are only answered if T1 was answered "yes." The percentages indicate the fraction of the papers from which that question/attribute is applicable. For example the fractions reported for the Biological Source are based on papers for which T3 was answered "yes". The numbers shown for Firas-1 are those obtained after re-annotation of that dataset.

Question/Semantic Attribute in the Annotation Form	Firas-1 (500)	Experts-1 (340)	Experts-2 (220)
T1: Has a model?	67% (336)	52% (176)	75% (166)
T2: Multivariate model?	95% (320)	85% (149)	89% (148)
T3: Has molecular features?	84% (281)	75% (131)	81% (135)
Biological Source			
BS1: Human	36% (101)	53% (69)	61% (82)
BS2: Animal	40% (111)	31% (41)	23% (31)
BS3: Cell line	45% (126)	38% (50)	39% (53)
BS4: Pathogen	10% (28)	8% (10)	3% (4)
Type of Assay			
A1: DNA	53% (148)	36% (47)	41% (56)
A2: RNA	53% (149)	32% (42)	45% (61)
A3: Protein	74% (208)	69% (90)	66% (89)
T4: Outcome is clinical?	44% (148)	46% (81)	62% (103)
Clinical Purpose			
CP1: Diagnosis	16% (23)	9% (7)	5% (5)
CP2: Risk Assessment	28% (42)	43% (35)	34% (35)
CP3: Prognosis: tx unspecified	32% (47)	16% (13)	32% (33)
CP4: Prognosis: tx specified	35% (52)	42% (34)	41% (42)
T5: Outcome is biological?	74% (248)	65% (114)	65% (108)

Table 20 – The annotation outcome within the Experts-1 article set broken down for experts #3, #7, and all other experts combined.

	Experts-1 Article Set		
Question/Semantic Attribute in the Annotation Form	**Expert #3 (80)**	**Expert #7 (20)**	**All others (240)**
T1: Has a model?	36% (29)	25% (5)	65% (142)
T3: Has molecular features?	62% (18)	80% (4)	77% (109)
T4: Outcome is clinical?	72% (21)	80% (4)	39% (56)

Table 21 – The annotation outcome within the Experts-2 article set broken down for experts #3, #7, and all other experts combined.

	Experts-2 Article Set		
Question/Semantic Attribute in the Annotation Form	**Expert #3 (45)**	**Expert #7 (20)**	**All others (155)**
T1: Has a model?	31% (14)	65% (13)	90% (139)
T3: Has molecular features?	43% (6)	77% (10)	86% (119)
T4: Outcome is clinical?	100% (14)	100% (13)	55% (76)

Training and Validation of Machine Learning Annotation Classifiers (Aim 3a)

The Firas-1 article set was used to answer the research question in Aim 3a: *Can existing or modified feature extraction transformations be used to train text classifiers that can replicate human semantic annotation of the gold standard?* SVM classifiers were developed in a way that mirrors the manual annotation process for the structured components of the annotation scheme as described earlier. The predictivity of these classifiers was tested via 5-fold cross validation using the Firas-1 dataset (Figure 21). The average 5-fold AUCs for the 11 classifiers *using the same folds for all experiments* are shown in Table 22. All the classifiers show excellent predictivity except the classifier associated with the DNA Assay Type attribute (0.800). This can be explained by some of

the semantic ambiguity associated with the definition of this attribute. This will be discussed later in this chapter.

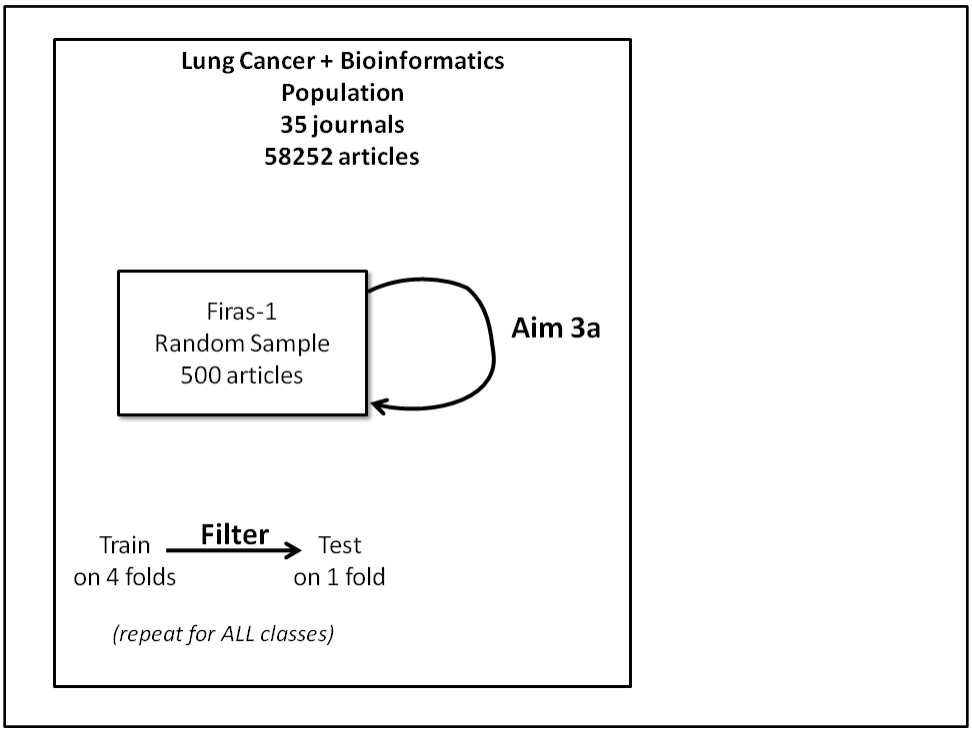


Figure 21 – For the research question in Aim 3a, 11 SVM models were trained and tested via stratified nested 5-fold cross validation using the Firas-1 articles set.

Table 22 – The results of the experiments for testing the research question in Aim 3a. The same folds were used for all 11 "annotation" classifiers as well as the 5 (T1-T5) "filteration" classifiers.

Question/Semantic Attribute in the Annotation Form Associated with the Classifier	Average 5-fold AUC in Firas-1
T1: Has a model?	0.933
T2: Multivariate model?	0.899
T3: Has molecular features?	0.941
Biological Source	
BS1: Human	0.898
BS2: Animal	0.936
BS3: Cell line	0.943
BS4: Pathogen	0.874
Type of Assay	
A1: DNA	0.800
A2: RNA	0.920
A3: Protein	0.920
T4: Outcome is clinical?	0.944
Clinical Purpose	
CP1: Diagnosis	0.903
CP2: Risk Assessment	0.923
CP3: Prognosis: tx unspecified	0.875
CP4: Prognosis: tx specified	0.932
T5: Outcome is biological?	0.918

Effect of Using NLP for Feature Extraction (Aim 3b)

The next research question that I investigated was: *Will modifying the feature extraction transformations used for training semantic classifiers in Aim 3a to include natural language processing (NLP) techniques alter their performance?* The feature extraction transformations were modified to incorporate the output of KnowledgeMap as described in the methods section. Specifically, the performances using 5-fold cross validation experiment of the SVM classifiers that result from 4 different feature extraction methods (shown in Table 23) were compared. The same folds were used for all experiments to be able compare their relative performance. The results are shown in Table 24. There was no significant change in performance associated with using the different transformations in all but one SVM classifier (BS4: "Biological Sample = pathogen").

Table 23 – This table shows the four different feature extraction transformation methods that were used for aim 3b. The first method is the default method used throughout this chapter. The other three methods differ by the MEDLINE information that are utilized in feature extraction: The second method relies only on the terms in the title and abstract, the third method relies in addition on MeSH terms, and the fourth method relies in addition on UMLS CUIs extracted by KM.

Symbol	MeSH Terms	UMLS Concept IDs	# of Features in 500 Articles	Preprocessing
LRF: MeSH(+)KM(-) (default)	+	-	19950	Log-rel freq with redundancy
L2F: MeSH(-)KM(-)	-	-	13337	Raw frequency L2-normalized
L2F: MeSH(+)KM(-)	+	-	19950	Raw frequency L2-normalized
L2F: MeSH(+)KM(+)	+	+	24629	Raw frequency L2-normalized

Table 24 – This table shows the relative predictive performance of the four feature extraction transformations described in Table 23 for all 11+5 SVM classifiers.

	Average 5-fold AUC in Firas-1			
Question/Semantic Attribute in the Annotation Form Associated with the Classifier	**LRF MeSH(+) KM(-)**	**L2F MeSH(-) KM(-)**	**L2F MeSH(+) KM(-)**	**L2F MeSH(+) KM(+)**
T1: Has a model?	0.933	0.907	0.918	0.916
T2: Multivariate model?	0.899	0.872	0.875	0.877
T3: Has molecular features?	0.941	0.900	0.912	0.918
Biological Source				
BS1: Human	0.898	0.873	0.886	0.889
BS2: Animal	0.936	0.932	0.952	0.953
BS3: Cell line	0.943	0.922	0.928	0.923
BS4: Pathogen	0.874	0.694	0.746	0.750
Type of Assay				
A1: DNA	0.800	0.766	0.776	0.790
A2: RNA	0.920	0.891	0.908	0.906
A3: Protein	0.920	0.907	0.911	0.914
T4: Outcome is clinical?	0.944	0.921	0.927	0.932
Clinical Purpose				
CP1: Diagnosis	0.903	0.871	0.875	0.874
CP2: Risk Assessment	0.923	0.915	0.928	0.944
CP3: Prognosis: tx unspecified	0.875	0.854	0.849	0.865
CP4: Prognosis: tx specified	0.932	0.883	0.901	0.903
T5: Outcome is biological?	0.918	0.895	0.897	0.895

The main drop in performance for BS4 was not associated with inclusion/exclusion of MeSH or UMLS CUIs. It was related to the pre-processing steps used. Note that there are only 28 cases (~5 cases per fold) in the Firas-1 dataset of papers describing predictive *Models* that used pathogens as the biological source (Table 19). The

use of feature weighting in the pre-processing step (LRF with Redundancy) may have reduced the over-fitting associated the low prevalence of this class in this dataset.

Testing Classifier Generalization to Annotations by Different Experts (Aim 3c)

The third research question under Aim 3c is: *Can text classifiers trained for semantic annotation of relevant papers in the domains of bioinformatics and lung cancer using annotation by one expert, replicate the semantic annotation of independent papers in the same domain by other experts?* The purpose of the experiments described in this section is to test the generalizability of the classifiers to different annotators. Specifically, I measured the predictive performance of SVM classifiers trained using the Firas-1 dataset and tested using the Expert-1 and Expert-2 datasets.

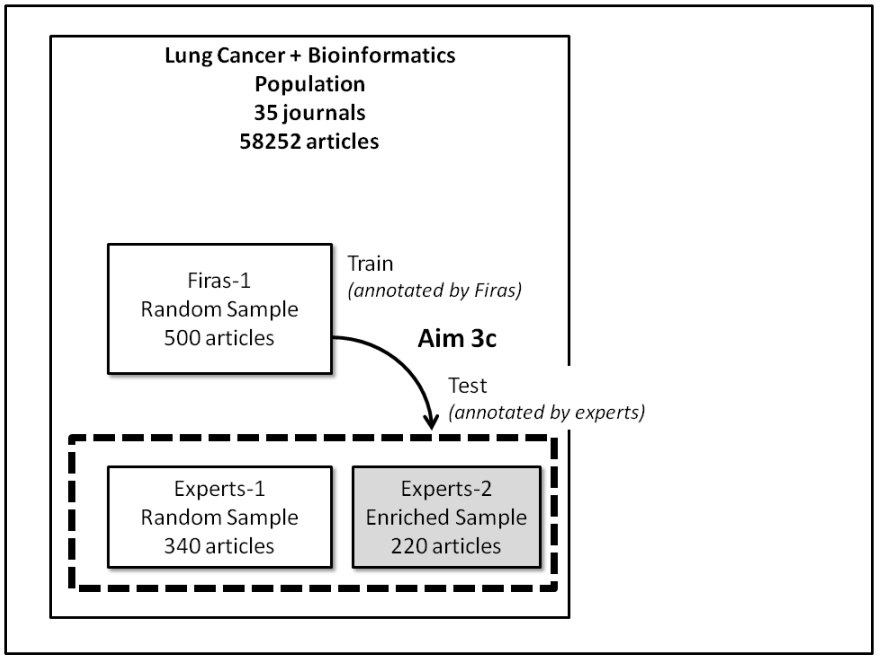


Figure 22 – The experimental setup used for the research question in Aim 3c. The SVM classifiers were trained using the Firas-1 dataset annotated by Firas and tested on an independent pooled dataset of 560 articles annotated by 8 other annotators.

In the first set of experiments (Figure 22), the SVM classifiers were tested using a dataset pooled from both Experts-1 and Experts-2. The results are shown in Table 25. The analysis was done twice for all 11+5 SVM classifiers, once using all experts and once using all experts excluding experts #3 and #7. The main impact of removing experts was mostly seen in the "filter" classifiers, which is also consistent with the assumption that these experts rely on a definition of predictive models that necessarily requires the *Model* having a clinical outcome. Overall the SVM classifiers show very good predictivity on datasets annotated by independent experts. (With only around 5% reduction in AUC from the cross validation results using Firas-1.)

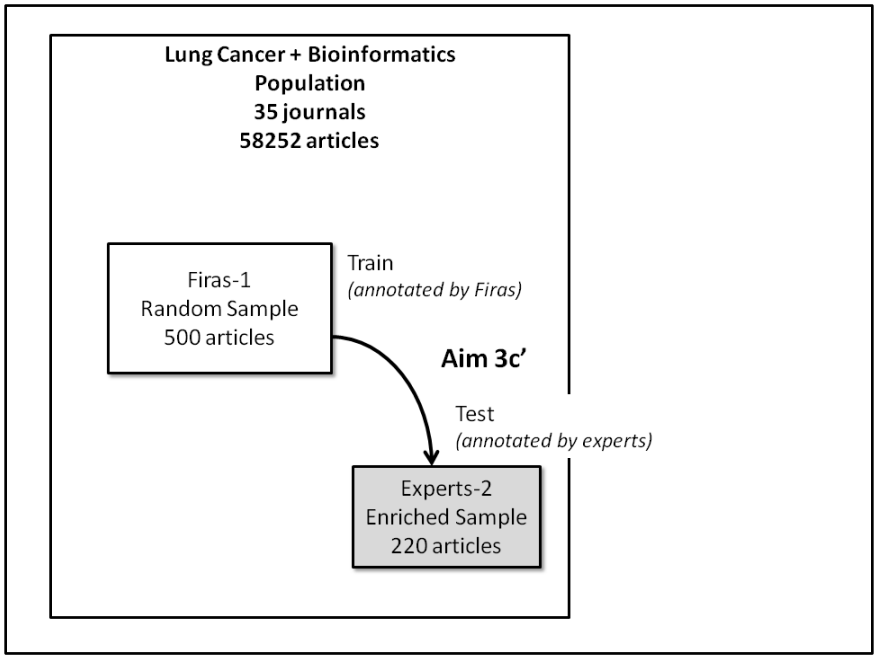


Figure 23 – Another experimental setup used for the research question in Aim 3c. The SVM classifiers were trained using the Firas-1 dataset annotated by Firas and tested on the independent Experts-2 dataset that was "enriched" to include mostly papers describing predictive models and annotated by other expers.

The second set of experiments (Figure 23) used only the enriched Experts-2 dataset to test the SVM classifiers. The purpose here is to evaluate the discriminating

ability of the classifiers in a more specialized dataset where non-relevant (non-model describing) papers are not present. The results of these experiments for all 11+5 SVM classifiers are shown in Table 26. In this setting, the SVM classifiers had good predictive performance that was diminished for some classifiers from that of the pooled (and larger) test datasets. The effect associated with removing experts #3 and #7 was again related to their annotation of mostly models with clinical outcome. Note that the classifier associated with CP3 ("Clinical Purpose" = "Prognosis Tx. Unspecified") had lower performance. This attribute will be discussed later in the chapter.

Table 25 – Predictive performance of SVM classifiers trained using Firas-1 and applied to a pooled dataset of Experts-1 and Experts-2 as shown in Figure 22.

	AUC Using Experts-1,2 as Test Dataset	
Question/Semantic Attribute in the Annotation Form Associated with the Classifier	**All Experts (560)**	**All Experts except #3,#7 (395)**
T1: Has a model?	0.762	0.866
T2: Multivariate model?	0.740	0.833
T3: Has molecular features?	0.781	0.865
Biological Source		
BS1: Human	0.857	0.856
BS2: Animal	0.876	0.878
BS3: Cell line	0.862	0.893
BS4: Pathogen	0.811	0.783
Type of Assay		
A1: DNA	0.758	0.749
A2: RNA	0.821	0.853
A3: Protein	0.846	0.866
T4: Outcome is clinical?	0.919	0.917
Clinical Purpose		
CP1: Diagnosis	0.872	0.880
CP2: Risk Assessment	0.933	0.925
CP3: Prognosis: tx unspecified	0.852	0.851
CP4: Prognosis: tx specified	0.894	0.897
T5: Outcome is biological?	0.761	0.831

Table 26 – Predictive performance of SVM classifiers trained using Firas-1 and applied to the "enriched" Experts-2 dataset as shown in Figure 23.

	AUC Using Experts-2 as Test Dataset	
Question/Semantic Attribute in the Annotation Form Associated with the Classifier	**All Experts (220)**	**All Experts except #3,#7 (155)**
T1: Has a model?	N/A	N/A
T2: Multivariate model?	0.581	0.745
T3: Has molecular features?	0.649	0.775
Biological Source		
BS1: Human	0.848	0.854
BS2: Animal	0.903	0.915
BS3: Cell line	0.789	0.847
BS4: Pathogen	0.811	0.806
Type of Assay		
A1: DNA	0.721	0.717
A2: RNA	0.766	0.806
A3: Protein	0.737	0.827
T4: Outcome is clinical?	0.885	0.886
Clinical Purpose		
CP1: Diagnosis	0.886	0.906
CP2: Risk Assessment	0.899	0.857
CP3: Prognosis: tx unspecified	0.771	0.774
CP4: Prognosis: tx specified	0.898	0.925
T5: Outcome is biological?	0.708	0.807

Large Scale Application of Machine Learning Classifiers

As described in the methods section and illustrated in Figure 24, I tested the practical scalability of using the SVM classifiers that have been discussed in this chapter. Using a Linux virtual private server with 256 MB of RAM to run the Python scripts, this task was relatively easy. The download and feature extraction step for all 11,000 articles and preparation for machine learning was executed in around 30 minutes. Running all 16 SVM classifiers was complete in less than 10 minutes. The output of the SVM classifier for some of the filter questions and semantic attributes are shown in Figures 25-31 and will be discussed in the next section.

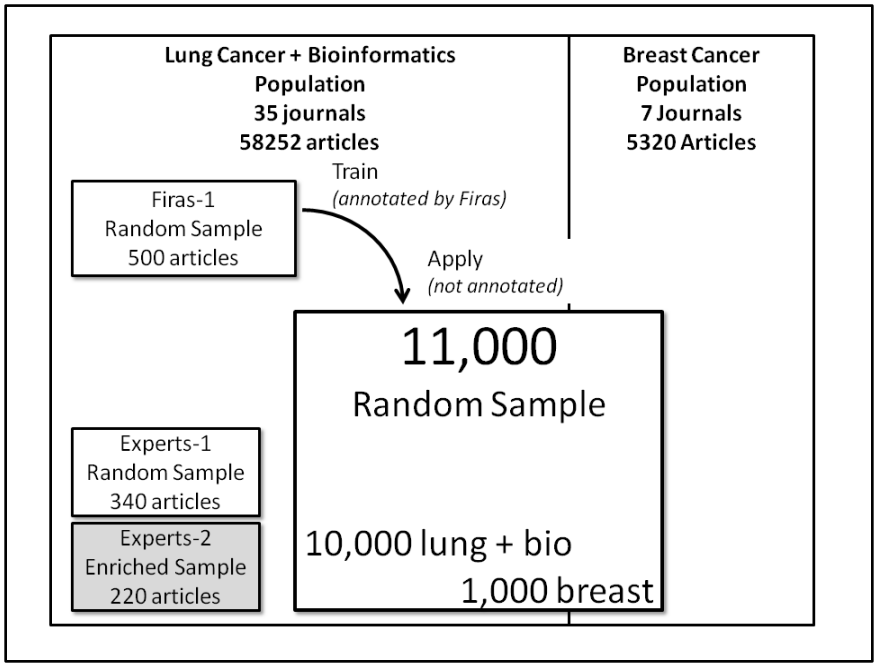


Figure 24 – The 11+5 SVM classifiers that were trained using Firas-1 dataset and that were described in this chapter were applied to a large independent dataset composed of 11,000 articles randomly sampled from the lung cancer + bioinformatics and the breast cancer populations.

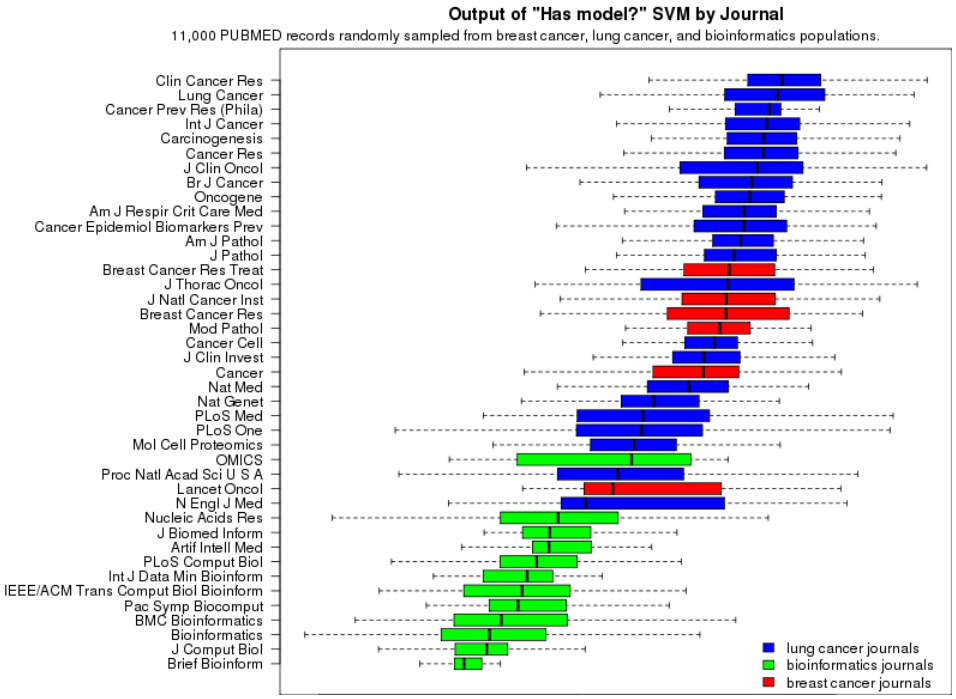


Figure 25

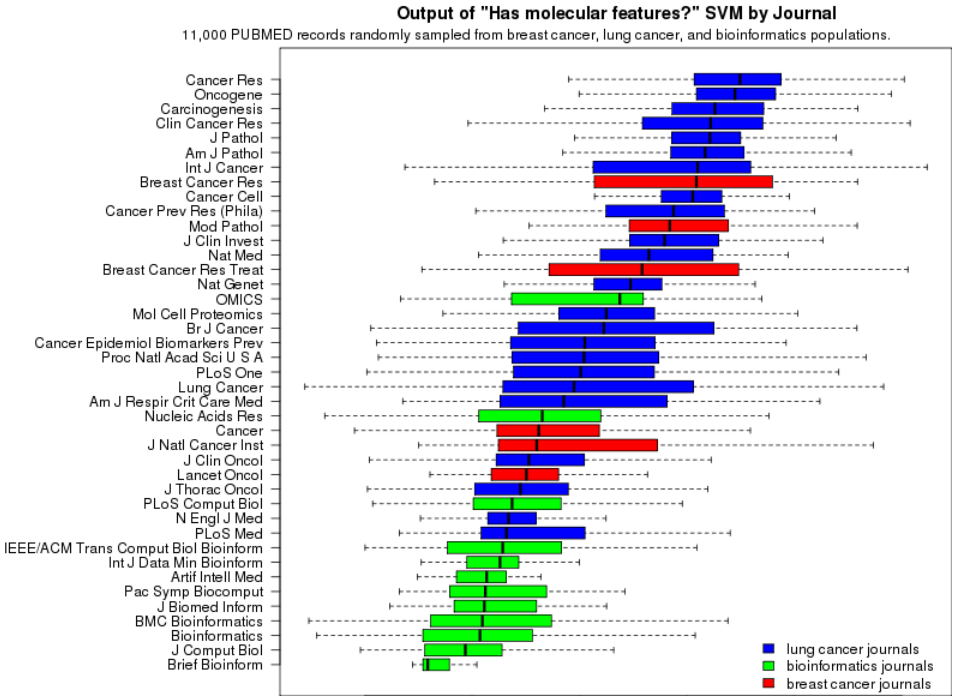


Figure 26

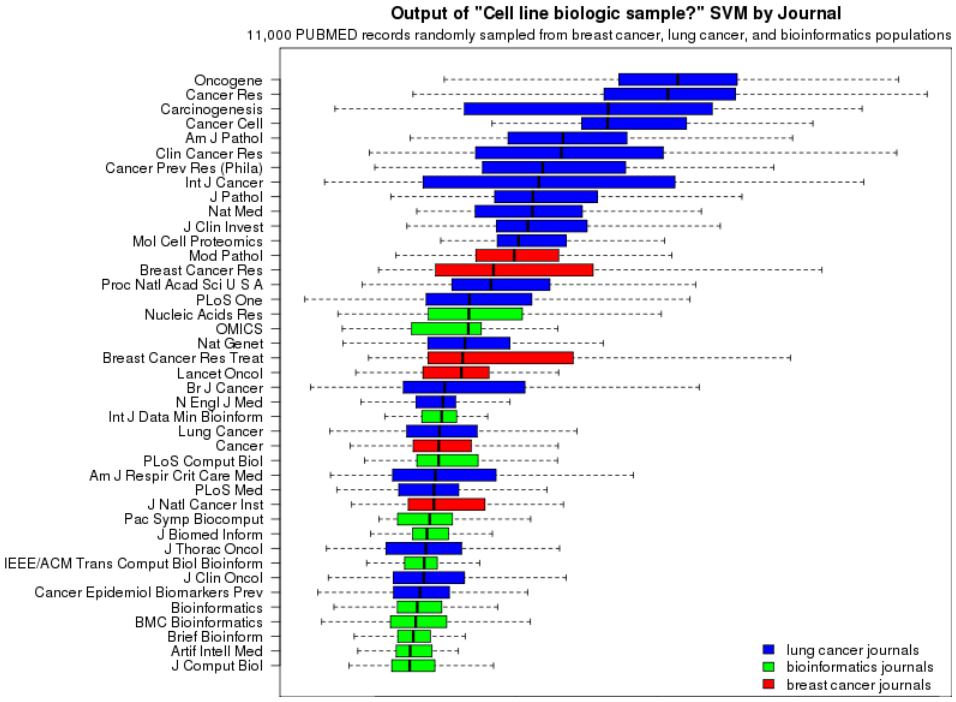


Figure 27

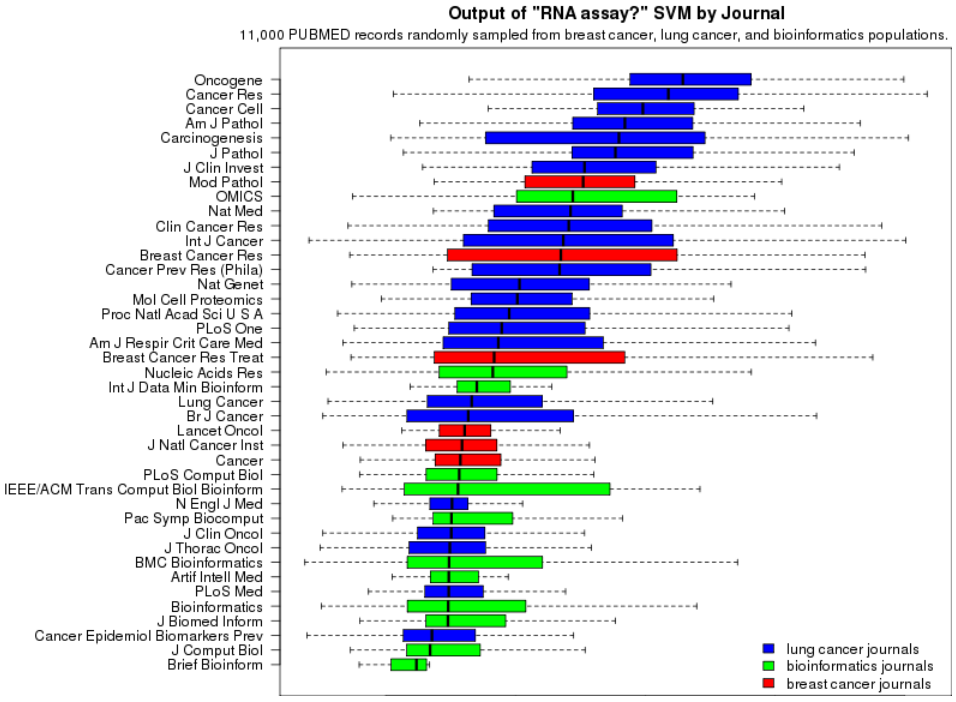


Figure 28

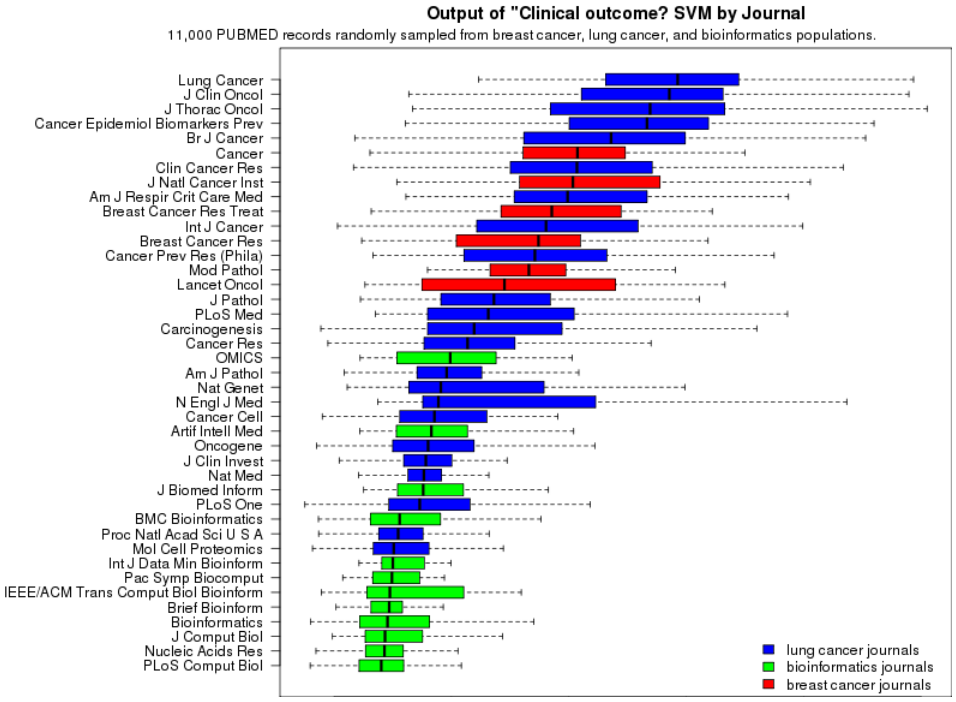


Figure 29

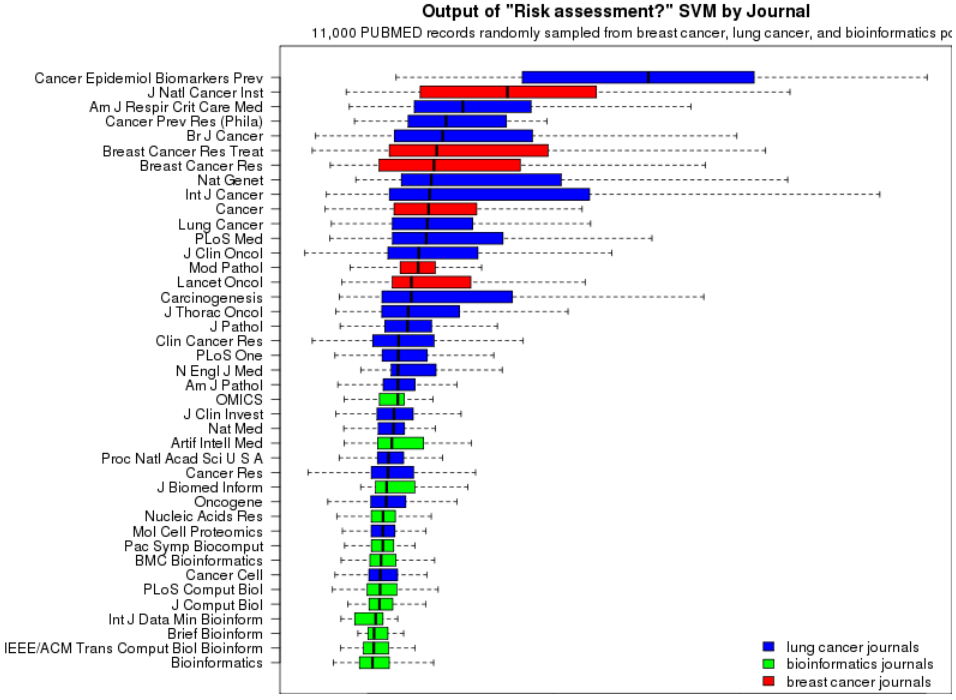


Figure 30

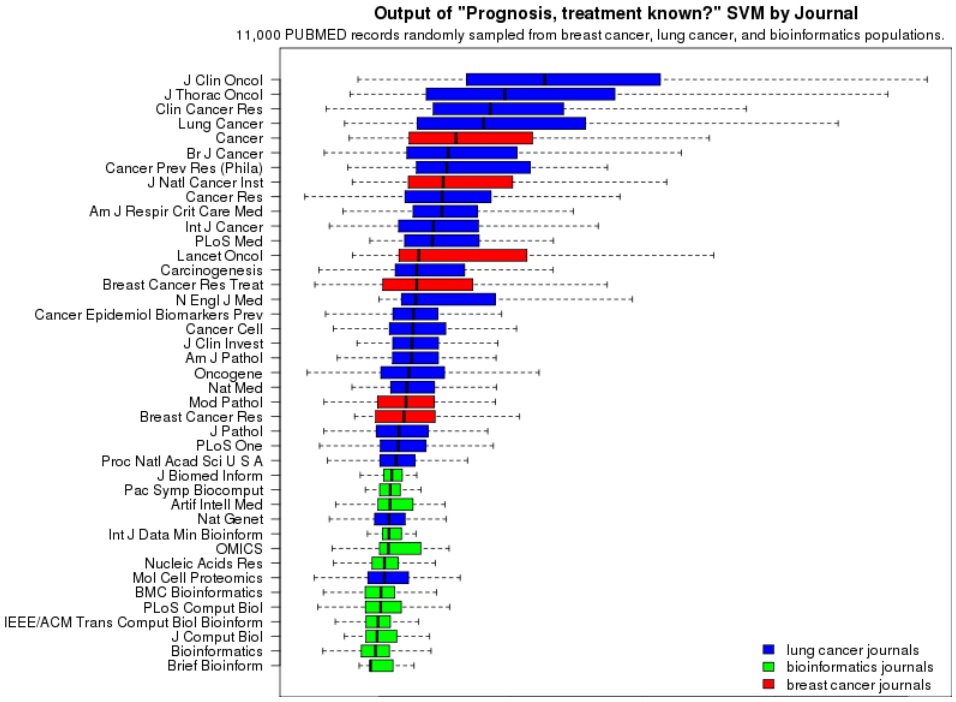


Figure 31

Discussion

Summary of Results

In this chapter I described the building and evaluation of a machine learning based approach for the semantic annotation of MEDLINE articles as befits our information retrieval framework. First, I described how the semantic annotation scheme presented in earlier chapters was refined to arrive to an operational definition of the relevant concepts. These definitions were codified in an annotation guideline document and a related paper annotation form. The resulting annotation process allowed human annotators to assign semantic attributes to MEDLINE articles that characterize the clinical bioinformatics predictive models that these articles describe. The annotation

guideline should lend itself to consistent annotation behavior by humans. This was highlighted when I re-annotated a set of 500 articles and found inconsistencies with my previous annotations. Most of those inconsistencies occurred in articles that were originally annotated prior to the final version of the annotation guideline document.

The ability to automatically replicate manual annotation was investigated by building 11 SVM classifiers that correspond to the structured semantic attributes annotated by humans using the paper form. Using an article set of 500 papers that I manually annotated and 5-fold cross validation to measure average AUC, the majority of the SVM classifiers showed very good or excellent predictive performance. Reliance on an NLP tool (KnowledgeMap) to enhance feature extraction for these SVM classifiers by extracting UMLS Concepts from the MEDLINE record did not alter their predictive performance. The SVM classifiers' ability to predict annotation behavior by an independent set of experts was investigated by applying them to two test sets annotated by 8 unique experts. One of the subsets consisted of 340 randomly selected articles from the background study population. The other dataset was constructed using an "enrichment" procedure that relied on a validated machine learning filter to find 220 articles specifically describing predictive *Models* as previously defined (Chapter IV). Overall, the SVM classifiers showed very good predictive performance for the combined dataset. That performance was slightly diminished when the classifiers were only applied to the "enriched" dataset. The limitations of these machine learning classifiers, the variability in expert annotation behavior, and possible residual semantic ambiguity in the annotation scheme will be discussed below in the context of these experiments. Finally, I wanted to gauge the practical scalability of the SVM classifiers that have been developed

so far. I applied these classifiers to 11,000 MEDLINE articles. The results were presented in this chapter and will be discussed below.

Limitations of the Machine Learning Classifiers

As discussed in the previous chapter, a main structural limitation of the approach used for building the machine learning classifiers is due to relying solely on the contents of the MEDLINE record for feature extraction. This approach omits information about the content of the articles that is otherwise available to the human annotators. This limitation was also observed in the context of semantic annotation as described in this chapter. In many of the articles that I encountered during my annotation, the abstract did not provide definite information about some of the semantic attributes. I had to resort in many instances to the full text of the article and examine the experimental methods section to determine the full list of the types of assays that were used. Sometimes, and especially for certain journals, that information was only mentioned in supplemental material beyond the main body of the article. In addition, sometimes the MeSH terms provided misleading information about the experimental method. In many articles, MeSH terms like "cell line" were found the MEDLINE record. Upon examination of the article, I would find that the experimental manipulations were done on animal xenograft models (i.e. biological source = animal not cell line). Sometimes the MeSH term "gene expression profiling" would refer to articles in which mRNA was not measured and the information about gene expression was obtained via immuneohistochemistry (a "protein" type of assay). This approach for feature extraction was used because it was vastly more practical to obtain the publically available MEDLINE information (via PubMed e-

utlities) than to programmatically access the full text of the articles from all journal sources. Despite this limitation, the SVM classifiers showed very good predictivity for many of semantic attributes. *This shows that it is possible for the SVM classifiers to duplicate the human annotation which was informed by the content of the full article by learning from patterns that were only present in the MEDLINE record.*

Finally, there were two other findings reported in this chapter that may be due to the limitations of the machine learning method that was used. They are probably not related to the SVM algorithm per se but to the experimental design and the composition of the training and testing datasets. The first finding is that the "enrichment" procedure used to construct the Experts-2 dataset only produced a single article from the bioinformatics journals (table 18). This was discussed in the previous chapter and seems to be consistent with the conclusion that was drawn about the low prevalence of "relevant" articles from the specialized bioinformatics journals in the overall lung cancer + bioinformatics population and the effect of that prior probability on the training and performance of the filter. The second finding is the relatively poorer predictive performance of the SVM classifier associated with the "BS4: biological source = pathogen" semantic attribute when tested using the expert datasets. Recall from table 19 that the prevalence of this attribute was low in all of the article sets (28/500, 10/340, and 4/220 in Firas-1, Experts-1, and Experts-2 respectively). The use of redundancy weighting may have suppressed the over-fitting of this classifier during cross validation. (Note the deterioration of cross-validation performance of this classifier when redundancy weighting pre-processing was removed in Table 24.)

An interesting observation is that the "CP1: clinical purpose = diagnosis" attribute had a similar low prevalence in all datasets (23/500, 7/340, and 5/220 in Firas-1, Experts-1, and Experts-2 respectively); however very good predictive performance of the associated SVM classifier was found in all the experiments in this chapter. I speculate that the robustness of this classifier may be due to: (1) less semantic ambiguity surrounding what constitutes a "diagnosis" type of clinical outcome leading to a consistent annotation by human experts of this attribute; and (2) the presence of highly discriminating terms in the dataset with respect to this class such as "diagnosis" or "differential diagnosis."

Variability in Expert Annotation Behavior

Due to the lack of objective criteria for determining the semantic annotations, the judgment of the human experts was treated as the gold standard that was used for training and validation of the SVM classifiers. The variability of expert annotation behavior, specifically the outlying annotation behavior of experts #3 and #7, was highlighted in this chapter (see Tables 19-21) as well as the previous chapter. Assuming good faith effort by the annotators and assuming that the annotators are consistently and faithfully annotating articles according to their understanding of the annotation guidelines, the variability in annotation behavior can be explained by the difference in their cognitive interpretations of the guidelines. As shown in the results section, the behavior of these 2 experts is consistent with the hypothesis that their understanding of the working definition of predictive *Model* requires that the *Model* be associated with a clinical outcome. While variability exists between all experts and at different granularity levels, the variability

between these 2 experts and the rest of the experts was the most profound in that it seems to be related to the definition of predictive *Models* themselves; furthermore, it was readily detected by simple comparison of the composition of their annotations to those of the other experts.

Regardless of the cause of variability in annotation behavior, the procedure that I used to collect expert annotations did not provide for a means to monitor or correct obvious outlying behavior *during* data collection. This was an inherent limitation of this experimental design that stemmed from my dual roles as (1) developer of the system/author of guidelines and (2) evaluator of the system. To maintain consistency I did not want to modify the guideline document after the beginning of expert recruitment for the analysis dataset. During the period when experts were actively annotating their articles, I refrained from discussing the annotation of specific articles with them (beyond general explanation of the guidelines if they asked me for clarifications) to avoid introducing bias. Analysis of the causes of variability in the datasets used for this study and the investigation of possible approaches to prevent or correct inter-annotator variability without introducing bias for one specific annotation behavior is an opportunity for future work.

Residual Guideline Ambiguity

The purpose of the annotation guidelines document is to help the human experts make consistent and deterministic semantic annotations of the articles to be able to construct high quality training and testing datasets. As discussed above, differences in the cognitive interpretation of the guidelines may lead to inconsistency in annotation

behavior which in turn may diminish the ability to train predictive and generalizable SVM classifiers. The following discussion of two semantic attributes highlights the possibility that some semantic ambiguity remains in the final version of the annotation guidelines.

The SVM classifier associated with the "**A1: Type of Assay = DNA**" semantic attribute showed consistently weaker predictive performance in the experiments for Aims 3a, 3b, and 3c. There seems to be less internal consistency in the training dataset (Firas-1) of this attribute's annotation than that of the other attributes. An informal survey of my notes taken during the annotation of the Firas-1 dataset reveals the following examples of the residual semantic ambiguity in the definition of this attribute.

- Sometimes Flow Cytometric Analaysis (FCA) is performed to measure the total DNA content of cells for cell cycle analysis (e.g. to detect apoptosis as an outcome or to select cells in a certain stage in the cell cycle). Even though technically, the assay is measuring the DNA molecule, there is ambiguity if such papers can be truly assigned a positive label for the A1 semantic attribute. The measurement of the DNA is not for genetic purpose such as when PCR or hybridization is used, but as a surrogate for a cellular state. In the case of detecting apoptosis as an *outcome*, the DNA measurement is actually utilized as a *dependent variable* not an independent variable. For example see article 17486061 ("Frequent loss of expression of the pro-apoptotic protein Bim in renal cell carcinoma: evidence for contribution to apoptosis resistance." *Omics*)
- In many studies the genome of model animals or cell lines is altered to study the effect of a gene on an outcome such as a downstream effect in a molecular

pathway or a specific phenotype such as disease response or cell motility. Conceptually, the experimental manipulation is at the gene "DNA" level. In practice the actual measurement of DNA may not be strictly performed. For example, sometimes mice are used which have been altered using germ-line knockout of the specific gene. In some studies, the genome DNA was not measured, but the genotype was confirmed via gene expression analysis (e.g. confirming mRNA expression using Northern Blot or RT-PCR) or protein measurement (e.g. by using immunohistochemistry to measure the presence of the gene product protein in the tissue). Alternatively the genome may not be altered, but the cells are transfected using cDNA plasmids to induce the expression of certain gene products. Gene regulation networks are sometimes studied by inserting certain reporter genes (e.g. Luciferase or CAT, see 16916793 and 17012283) into the genome and swapping them with genes whose activity is investigated. In other words DNA measurement is not consistently associated with genotype measurement: Sometimes DNA is not measured at all, and other molecules are used as surrogates of genotype; on the other hand, sometimes DNA (instead of mRNA) molecules are used/measured during the experimental manipulation or measurement of gene expression states.

The SVM classifier associated with the "**CP3: Clinical Purpose = Prognosis: Treatment Unspecified**" semantic attribute showed relatively diminished predictive performance when applied to the independent expert dataset. Informal discussion with experts *after* their annotation was complete indicated that some of the experts differed in their interpretation of that concept from me and from the other experts. Recall that that

semantic attribute refers to *Models* whose purpose is to predict clinical outcome *irrespective of the type of treatment given*. On the other hand the "CP4: Clinical Purpose = Prognosis Treatment Specified" attribute refers to *Models* whose purpose is to predict differential outcomes for specific treatment such as the study of outcomes during randomized controlled drug trials or the prediction of treatment response based on genomic data. In most of the papers that describe the former (CP3) cases, the treatment that the patients received is actually specified but it is not the variable that is controlled or experimentally manipulated. For example, a paper predicting the metastasis in cancer patients (based on characteristics that are independent of treatment such as molecular subtype or the clinical stage) would typically list the standard chemotherapy treatment that the patients in their study population received. Some of the annotators told me that they would consider the reporting of the treatment in the article as an indication that this should be annotated as using the CP4 ("treatment specified") as opposed to the CP3 ("treatment unspecified") attribute.

Large Scale Application of the Classifiers

In this chapter, I described the application of the 5 "filtration" and 11 "annotation" SVM classifiers to a large number of articles. The purpose of this procedure was to verify the practical scalability of using the SVM classifiers. The results that were reported are only descriptive of the outcome of this procedure and cannot be used to draw conclusions beyond what can be concluded from an observational study. The main outcome from this procedure is that it will be practical to scale the application of the SVM filters to a large fraction of the articles in MEDLINE. The majority of the

computational time was consumed by downloading the MEDLINE records, feature extraction and preparation of the dataset for machine learning. The application of the already trained SVM classifiers was trivial.

The output of this procedure is a computed decision function for each article by each of these 16 classifiers. The results for Aims 2 and 3 have shown that the decision function can be used to discriminate articles based on the semantic annotation that is associated with that SVM classifier. The range and mean of the decision function for MEDLINE articles in each journal are shown in figures 25-31. The following are some observations about this output: The filter for question 1 "T1: The paper describes at least one predictive model?" seems to assign lower values on average to articles from the bioinformatics journals. Some journals like *NEJM* and *PLoS One* had a wider range of output for this filter. The journals with the highest mean values for "T3: Has Molecular features?" include *Cancer Res, Oncogene, Carcinogenesis* whereas journals with highest mean values for "T4: Outcome is clinical?" include *Lung Cancer, J Clin Oncol, J Thorac Oncol*. Note the difference in range of these two filters for the *NEJM*. *Cancer Epidemiol Biomarkers Prev* had a noticeably higher mean and range than all the other journals for the semantic attribute filter "CP2: Clinical Purpose = Risk Assessment."

One important thing to note is that the output of these classifiers is not completely independent. For example the semantic interpretation of a positive label for "T3: Has molecular features?" is that T1 = True AND T3 = TRUE. This is based on the workflow for the annotation form (Appendix C). Conversely, if that label is false then it that can include two different types of papers: (1) papers that do not describe a *Model* and (2) papers that do describe a *Model* but that *Model* does not have molecular features.

Similarly, the 11 semantic attributes are dependent on two levels of filters that precede them. This explains why most of the (CP) semantic attributes in the figures have ranges that extend to the left edge of the figures. Assuming the Firas-1 dataset composition (500 randomly sampled articles) is a good estimate of the composition of the 11,000 randomly sampled articles, then 70% of articles (calculated using Firas-1 annotations in Table 19) do not describe a predictive Model OR describe a Model that is not clinical. Therefore the CP1-4 classifiers should provide a low decision function for at least 70% of the articles.

Conclusion

The work in this chapter was an extension to the work described in the previous chapter. While the previous chapter investigated the use of scalable SVM classifiers for identifying relevant articles, this chapter applied similar SVM classifiers to semantically annotate the relevant articles. The classifiers have shown very good predictive performance when validated using manually annotated articles. The use of the KnowledgeMap natural language processing tool to extract UMLS biomedical concepts from the MEDLINE record did not seem to alter the performance of the classifiers within our experimental design. The good predictive performance of the classifiers was found to generalize to annotations made by an independent set of experts. The annotations, by man and machine, were based on annotation guidelines that were developed for these experiments based on the previous semantic analysis of this domain. Some of the variability in expert annotation behavior and in classifier performance can be explained by semantic ambiguity that remains in some of the concepts described in the annotation

guidelines. Finally, this chapter demonstrated the practical scalability of the SVM classifiers to a large numbers of articles in the MEDLINE database.

CHAPTER VI

DISCUSSION

Summary of Results

The work presented in this dissertation describes a framework for retrieving and organizing published information about clinical bioinformatics predictive models, models that can predict clinical outcomes based on the results of molecular biology assays or techniques. This information retrieval framework has to overcome two challenges: (1) the semantic complexity and (2) the large volume and fast pace of published information in this domain. The limitations and challenges of existing tools were discussed in Chapter II.

The first aim of this dissertation was to conduct a semantic analysis of this domain and use the results of that analysis to inform the design of the information retrieval framework. In chapter III, a focused in-depth analysis of a small number of well known publications in this domain led to the definition of an ontology of predictive *Models* and of related objects. To answer the envisioned queries for these *Models*, an indexing scheme was developed that relied on the annotation of *Models* (and of *Papers* describing these *Models*) according to a clinical bioinformatics *Context*. A *Context* can characterize *Models* along four dimensions: *Disease*, *Population* (biological source of molecular data), *Modality* (molecular assay type), and *Purpose* (the type of clinical outcome).

The information used by the proposed framework will be obtained from published articles in this domain. The second aim of this dissertation was to train and test machine learning SVM filters that can automatically retrieve relevant articles from MEDLINE. These filters were trained and validated using a corpus of manually annotated MEDLINE articles. To ensure the consistency and quality of this manually annotated dataset, a procedure for determining "relevant" articles was developed based on an operational definition of bioinformatics predictive *Models*. The SVM filters showed excellent predictive performance when evaluated using a dataset of manually annotated articles selected from the domains of lung cancer and bioinformatics. Furthermore, their predictive performance was found to extend to articles that were sampled from another domain (breast cancer) and to independent articles that were annotated by a separate group of expert annotators.

The third aim of this dissertation was to train and test machine learning SVM classifiers that can automatically annotate relevant articles using the indexing scheme that was proposed earlier. The definition of clinical bioinformatics *Context* was found to be deficient and did not provide adequate guidance for consistent annotation of newly encountered articles beyond the focused set that was used in the first aim. An annotation guideline and a paper annotation form were iteratively refined using experience gained from the annotation of new articles and feedback from experts. The guideline and form were utilized to construct a corpus of manually annotated articles. SVM classifiers were developed to mirror the structured semantic attributes on this form. These classifiers were trained and tested using the manually annotated datasets. Machine learning experiments showed very good predictive performance by these filters when validated using the

dataset sampled from lung cancer and bioinformatics domains. This performance was also found to generalize to annotations provided by independent experts. The effect of using natural language processing techniques to enhance feature extraction (by identifying UMLS biomedical concepts) did not improve the performance of these classifiers.

Analysis of the experiments conducted under both Aim 2 and Aim 3 found that the variability in the predictive performance of the SVM classifiers can be partially explained by the remaining semantic ambiguity of some of the concepts in the annotation guideline. Finally, the scalability of the SVM classifiers was practically verified by applying them, with relative ease, to a large set of randomly selected articles in MEDLINE.

Limitations and Lessons Learned

Knowledge Representation and Annotation

The semantic complexity of the domain of clinical bioinformatics was one of the main challenges facing the development of the information retrieval framework described in this dissertation. A multitude of biomedical concepts are relevant to the meaningful organization and retrieval of the predictive *Models* (e.g. molecular assay techniques and related biology, epidemiological concepts regarding clinical outcomes, etc.) The information elements required for solving this problem were not clearly defined at the beginning of this work. The choices made during the knowledge representation phase played a big role in shaping the practical components of this framework such as the

annotation form or the classes used for the machine learning classifiers. In retrospect, this work seems to encompass two different approaches for knowledge representation. The first approach, followed during the early phase (Chapter III), relied on a focused and in-depth analysis of a small set of illustrative examples as well as a priori specifications of the envisioned queries for this framework. This produced a formal ontology of objects and relationships in this domain. The original intent was to extract these objects from published articles and to use sophisticated techniques such as description logic based knowledgebases to support complex semantic queries. Practical realities frustrated this effort. The second approach relied on iterative and piecemeal refinement of the relevant semantic attributes and indexing scheme based on patterns that emerged as I annotated more articles. Furthermore, feedback from experts about the utility or ambiguousness of certain definitions also helped refine the semantic annotations (and remind me of the teleological nature of this process). This approach resulted in the relatively simple and flat annotation scheme. I was then able to practically construct the annotated corpus that was used to train and validate the machine learning classifiers that mirrored the annotation scheme. I regret not using this grounded approach earlier in my PhD work. Recall the grounded theory approach that was used by Chapman et al[59] to design an annotation scheme for extracting clinical conditions from emergency department records. They started with a general theory statement and followed an iterative approach to refine the annotation schema.

While the annotation scheme that was eventually used allowed many practical accomplishments to occur, it did suffer from some limitations. Some of its limitations that relate to Aims 2 and 3 were extensively discussed in Chapters IV and V. Another

limitation of this annotation scheme is that the tradeoff for simplicity resulted in less representation of some of the concepts that were originally envisioned for this framework. *Model* characteristics such as the strength of evidence (e.g. type of validation) or the type of *Algorithm* that is used by the predictive *Model* (e.g. logistic regression, artificial neural network, etc) are not represented. Pointers to related objects, such as the *Dataset* used to create the *Models* are also not represented or indexed.

Machine Learning Classifiers

The limitations of the machine learning classifiers that relate to Aim 2 and Aim 3 were discussed in their respective chapters. They include the structural limitations of using only the MEDLINE record for feature extraction, and the over-fitting that results from some of the classes having low prevalence in the datasets (such as articles where the biological source of molecular information was "pathogen.") An additional limitation that I would like to highlight here is that the SVM classifiers for the "annotation" of semantic attributes are not independent from the SVM classifiers for the "filter" questions. The entire datasets (including articles where the pre-requisite "filter" questions were false) were used for training the semantic annotators. This will have implication on how the sequential application of the SVM "filters" (Aim 2) followed by the SVM "annotators" (Aim 3) will be used in practice.

Future Work and Open Questions

Machine Learning and Automated Annotation

The natural continuation of this work is to apply the classifiers that have already been trained to retrieve and organize papers that describe predictive *Models*. For example, this approach can be readily used to reduce the search space if one is interested in a comprehensive literature search for molecular signatures for a disease. Beyond a direct and ad hoc application of the filters to search for articles, these classifiers can be used to construct the back-end of a query based information retrieval tool. As discussed in the previous section and when discussing the outcome of the large scale application of the classifiers, the outputs of the different classifiers are not independent. An open question is how to transform the decision functions computed by each of the classifiers about a given article into data that can be leveraged in response to specific queries. Note that these classifiers are entirely coupled to the MEDLINE database. There are many existing resources that allow searching the space of MEDLINE articles (e.g. PubMed and other tools that enhance PubMed) or of related databases (GEO, GenBank, Protein, etc.) The annotations that can be derived from the SVM classifiers can be used as an additional layer of information that will compliment and leverage the other existing resources.

The following are some of the other open research topics relating to the strict machine learning aspect of this work:

1. Analysis of informative features and the effect of feature selection on the performance of the different SVM filters or semantic annotation classifiers.
2. The addition of new sources for features extraction such as the list of chemicals or the journal name from the MEDLINE record.

Building High Quality Annotated Corpora

The manual annotation that has been undertaken can be expanded and improved beyond the experimental set-up that was intended to evaluate the SVM-based approach. There is room to revise and improve the annotation guideline and use it to annotate more articles from a variety of MEDLINE populations. The expanded annotated datasets can be used to directly populate a backend database for a query based information retrieval tool. It can also be used to re-train the classifiers with higher power. For example, I have already collected an additional article set (Firas-3, not discussed in this dissertation) that was obtained via the same "enrichment" method as Experts-2. This dataset can be added to the existing training dataset used for the SVM classifiers to increase the prevalence of papers that describe *Models*.

Further research can be done using the collected manual annotations. A group of experts can review these annotations and analyze the sources of discordance. The resulting insights can be used to improve the guidelines or the general process that was used to create these guidelines.

Interesting research can be done to investigate different tools and methodologies for building high quality annotation corpora. These tools can be used to build annotated bibliography articles, but the methods can very well generalize to different types of text corpora such as text content of medical records. These methodologies can occur along three fronts:

1. Building new or adopting existing annotation workbench tools. For example, I have already implemented a simple multi-user annotation website (not described in this dissertation) that was used by Drs. Aliferis, Boulos, and Fu during an early

pilot phase of article annotation. This can be expanded to allow simple database manipulations of multi-user annotations (storing and exporting annotations, generating summary statistics or statics about concordance, use to support experimental manipulations, etc.)

2. Investigating the methods and processes for generation, codification, and communication of different annotation schemes and guidelines such as using wikis for the collaborative authoring of annotation guidelines.
3. Investigating the use of different forms of incentives for the creation of high quality annotation datasets. This can include incentives based on games that reward consistency, adherence to guidelines or concordance with other players.

Conclusion

The main goal of this work was to develop a framework for retrieval and organization of clinical bioinformatics predictive models. The framework relies on a specialized annotation and indexing scheme that was developed using semantic analysis of this domain. Scalable machine learning classifiers were successfully trained to replicate human experts' ability to retrieve relevant MEDLINE articles and to annotate these articles using the specialized annotation scheme. The experiments that were performed highlighted the importance of using clear annotation guidelines that provide unambiguous operational definitions for semantic annotations.

APPENDIX A

PUBLISHED SUPPLEMENTARY MATERIAL TO CHAPTER III

Note

This appendix is a verbatim replication of the "Appendix" section of the published article: Wehbe FH, Brown SH, Massion PP, Gadd CS, Masys DR, Aliferis CF. A novel information retrieval model for high-throughput molecular medicine modalities. *Cancer Inform. 2009 Feb 9;8:1-17.*

Context Indexing and Automation

As mentioned earlier, an object's *Context* is represented by a tuple that specifies *Disease*, *Population*, *Purpose*, and *Modality*. Whenever an object is described in a *Paper* that object is indexed by the *Context* with which it is described in that *Paper*. An object, e.g. *Dataset*, can be indexed by many *Contexts* because more than one *Paper* can reference the same object and in multiple contexts. For example, a "neural network" *Algorithm*, can be described in the following *Context* in one *Paper* (<DLBCL, Human Patients, Prognosis with Treatment, Proteomics>) i.e. neural network predictive *Models* were developed to predict prognosis in DLBCL using proteomic data. It can then be described in a different *Context* in another Paper. A *Paper* can be indexed by all the *Contexts* that apply to the objects in that *Paper*; however, individual objects described in a *Paper* are not necessarily described by all the *Contexts* that are mentioned in that *Paper*. For example, a *Paper* that evaluates a certain *Algorithm* using multiple *Datasets*

drawn from multiple diseases can be indexed by *Context* tuples that reflect all the diseases, but each individual *Dataset* can only be indexed using tuples that reflects its specific disease.

We use a canonical set of terms to specify the individual elements of a *Context* tuple. Initially we are only covering Neoplasms, and we will adopt the following nomenclature for *Disease*: Breast Neoplasms, Lung Neoplasms, Colorectal Neoplasms, Prostatic Neoplasms, and so on to cover all neoplasms in the domain of clinical bioinformatics. *Population* refers to one of three types: Human Patients (*Datasets* created by assays on tissues taken from patients, this can include normal tissue taken as control), Cancer Cell Line, and Animal Model. *Purpose* refers to the type of clinical outcome, we have determined four categories of clinical outcomes: (1) Diagnosis, i.e. using a computational Model to assign a diagnostic label based on molecular profile, an example in this category is the well known AML/ALL classification Dataset by Golub et al. (Golub et al. 1999); (2) Prognosis with no treatment, (3) Prognosis with one treatment arm, e.g. 5 year survival or metastasis prediction for patients on standard treatment; and (4) Prognosis with more than one treatment arm. The latter refers to situations where molecular computational models predict whether patients benefit from certain treatments, e.g. hormone therapy susceptibility based on molecular pathway activations. It also includes situations where the biological effect of certain chemicals, e.g. when tested on cancer cell lines, is measured. Finally, we determined three categories for *Modality*: (1) Genetic, refers to high throughput modalities that assess inherited genetic characteristics, e.g. SNPs and haplotypes; (2) Genomic, refers to high throughput modalities that assess functional genomic characteristics of disease or disease- related tissues, e.g. gene

expression microarrays, array CGH; and (3) Proteomic, e.g. high throughput modalities like Mass Spectrometry and Gel Proteomics.

There are a plethora of reference ontologies[130] and other formalisms that can represent *Context* elements with high granularity, e.g. SNOMED-CT for *Disease* and *Purpose*. A very expressive annotation of *Context* elements using complex ontologies with extensive subsumption hierarchies has many benefits. However it is labor intensive and with current and foreseeable technology relies heavily on human operators. As explained, our aim is to accelerate the indexing and annotation of *Papers* using automated or semi-automated means.

Classes, Objects and Relationships

We chose to represent the different object types, their relationships, as well as other entities in the clinical bioinformatics domain using Description Logic. Using Protégé's OWL plug-in (Knublauch, Musen and Rector, 2004), we developed an ontology (Discovery Systems Laboratory, 2008) that uses OWL axioms to define classes (concepts) of clinical bioinformatics entities and their respective properties (attributes). We chose OWL because the supporting tools are readily available, because we can use it to represent the domain unambiguously, and because we can use it to share our representation. We note that our aim is not to build extensive DL-based knowledgebases or to develop reference ontologies.

The main classes are *Papers*, *Datasets*, *Algorithms*, and *Models*. *Datasets* can have simple properties such as dataset dimensionality and sample size or complex ones such as related diseases and population characteristic. *Algorithms* are annotated with

properties to reflect the different methodologies e.g. “supervised” vs. “unsupervised learning”. Decision *Models* are annotated by the specific outcomes that they predict.

The semantics of relationships between classes in clinical bioinformatics is captured through *relationship classes*. For example, a *Paper* “proposes” or “invents” a specific *Algorithm*, “evaluates” that *Algorithm* using a *Dataset*, or simply “applies” that *Algorithm* on a given *Dataset*. So in addition to *classes of objects*, the ontology specifies *classes of relationships* between classes. Most relationships are binary, although there are some that are of higher arity. Relationships in our ontology are represented as classes and not properties (or “roles” in DL jargon). Our reasons for that include: (1) uniformity in representing all relationships, a significant fraction of which is not binary and thus cannot be represented by a DL-role, and (2) the need for rich annotation of the relationships themselves. For example, the relationship *Validate_Internal* (when a model is validated within a study) requires further annotations such as the type of validation performed (independent prospective sample? N-fold cross validation? Leave One Out cross validation?) Modeling relationships using classes instead of roles will add complexity to reasoning; however, for the foreseeable applications, we envision that a relational database with indexed relationship tuple tables will be adequate (for implementation and reasoning) for typical queries. Please see section on inference and implementation. Using classes to model relationships may also make reuse of this ontology more cumbersome, and is a limitation of this ontology. The four retrievable classes along with a subset of relationship classes are shown in Figure 6.

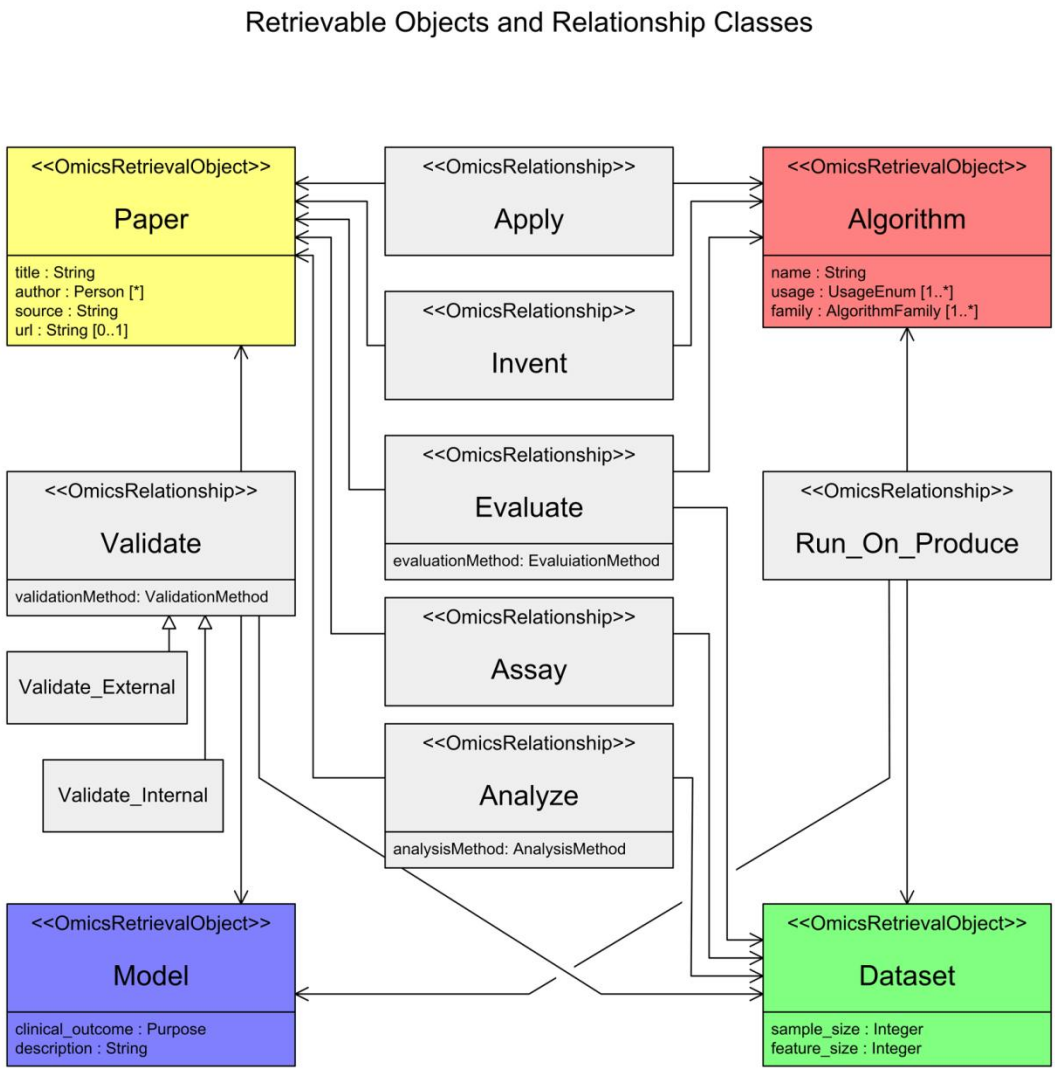


Figure 6. A UML diagram showing the four retrievable classes (subclasses of the abstract OmicsRetrievalObject class), some relationship classes (subclasses of the abstract OmicsRelationship class), and their associations. Some relevant properties of the retrievable classes are shown here as well. Apply, Invent, Assay, and Analyze are binary relationship classes, whereas the rest are ternary. The knowledgebase will contain instances of the retrieval and the relationship classes (as well as others not shown here, such as Context-related classes). For example, a given paper p (instance of Paper) may describe how a given model m (instance of Model) was validated using a dataset d (instance of Dataset). An instance v of the Validate relationship will be created referencing the objects p, m, and d. If d was the same data- set that was used to produce m, then v will belong to the Validate_Internal. class. Validate_Internal and Validate_External are subclasses of the ternary relationship, Validate. As such, they inherit its properties but offer more specialized properties such as specifying whether the validation method described by the Validate_Internal instance was done on independent samples within the related Dataset or not.

Research and discovery within the domain of clinical bioinformatics can be conceptualized as an overarching process that consists of: (a) collection of high-throughput molecular profiling data through molecular assays, (b) analysis of such data using specialized techniques, and (c) generation and validation of respective decision Models. These processes can be represented via a set of axioms that constrain relationships between classes in our ontology. Such constraints represent implicit domain knowledge such as: "In a *Paper*, one or more *Datasets* are assayed," or "An *Algorithm* is applied on a *Dataset* to produce a *Model*". Some of those constraints can be inferred from the UML diagram in Figure 6.

Currently, relationships between objects are manually annotated. Annotated relationships will be used to support the third step in the query process (semantic organization and display). These relationship instances are indexed and will be used to construct edges between the objects returned by the query and to drive the visual organization of results.

Support for Evidence Annotation and Filtering

As mentioned earlier, decision *Models* vary in the degree of validity and of generalizability outside of the population from which they were formulated. This variability results from the different methods with which the investigators validate their models and from the different experimental designs.

The performance of decision *Models* is usually evaluated on independent samples within the study *Dataset*, or on *Datasets* collected from different studies altogether. The former case is represented through the *"Validate_Internal"* relationship, and the latter

through the *"Validate_ External"* relationship. Both are subclasses of the *Validate* ternary relationship class (Fig. 6). Note that internal validations are sometimes done on non-independent samples. This is a bad practice that likely leads to over-fitting of the resultant decision Models, and is therefore an important attribute to highlight when displaying results. The *Validate_Internal* relationship is annotated as being done on either non-independent or independent samples.

The class *ValidationMethod* is a property of the *Validate* relationship class. Instances of this class correspond to specific validation methods such "Leave-One-Out Cross Validation," "N-Fold Cross Validation," etc. Statistical (Aphinyanaphongs et al. 2005; Wilczynski et al. 2005) classification methods have been used successfully before to classify the nature of evidence based on document content. We plan to automatically identify the *ValidationMethod* classes based on *Paper* contents.

Brief Discussion of Inference and Implementation

This paper addresses representational requirements of the information retrieval task at hand and the expressiveness of the model and underlying formalism. However we will briefly discuss inference and implementation of this model. In the first phase of our work, the papers were collected and organized manually. As we added more objects, and as the model was formulated we found that a simple relational model was enough to store and execute our simple queries. The objects were stored in their own tables, the relationships between the objects were stored in join tables, "Context" tuples were stored in a separate table, etc. It can be easily shown that matching the pattern of a "Context" query can be done via simple SQL queries that are dynamically generated. With the

correct choice of index keys, the retrieval process has been very efficient and we expect it to scale efficiently for simple queries. We used a simple (PHP-based) web framework with a browser interface and a MySQL database backend to build an application for storing and retrieving representations of our objects and their relationships. We have not yet implemented graph extraction and visualization. Graph extraction should be a trivial problem (identifying objects a certain depth from a model of interest, filtering out/in objects with specific properties, etc.) Graph visualization can be done via any of available graph-layout software (e.g. Graphviz). Graph elements can be passed to a web browser for rendering using a mark-up standard like SVG.

Semantically, we modeled the relevant objects of the domain, their relationships and the domain knowledge using OWL-DL axioms. This OWL file is available for download as indicated earlier. This leaves the door open for future storage and retrieval of the objects using DL-based databases and query languages; however, we do not see a need in the near future for DL-based inference and implementation. We think that using OWL to model the domain will facilitate semantic integration of this framework with other resources in the future. We envision implementing this framework as a web service that will be compatible with standard web services technology.

The inference task that we find most challenging is the automated identification of relevant papers from the literature and the automated annotation of the objects (for now only papers) by the correct "Context" tuples. Again, using automated or semi- automated methods is essential for building a comprehensive and up-to-date knowledgebase. This has motivated our drive towards simple representation formalism. Our current work is focused on building machine learning filters for identifying and annotating domain

papers using text categorization, and on investigating different approaches for tuple extraction. The purpose, and subsequent evaluation, of this effort is done along two lines. The evaluation of information retrieval recall and precision is done using a human-annotated corpus of papers that serves as a gold standard (currently exists for two domains, Lung Cancer and Breast Cancer with more annotations by domain experts underway). The individual papers are labeled for many things such as whether they describe the domain of clinical bioinformatics, whether they correspond to single gene vs. high throughput experiments, as well as all the *Context* tuple assignments that apply to each specific paper. The second dimension of evaluation relates to the adequacy of these automated techniques as means for building the knowledgebase required for this purpose, and how users interact with the resultant system.

APPENDIX B

ANNOTATION GUIDELINES

Note

This appendix contains the content of the Annotation Guidelines that was given to the expert annotators. It has been re-formatted to conform to the rest of the dissertation style

Thank you for agreeing to participate in this study. This document accompanies the set of article annotation forms. It will provide detailed explanation of the questions on the annotation form along with general guidelines and instructions to help you with your annotation. Please do not hesitate to contact me if you have any questions.

Email: firas.wehbe@vanderbilt.edu

Phone: (615) 936-3016

This document contains three sections:

1. General instructions
2. Guidelines for specific questions
3. Pointers that might help you go faster through the annotation process

I attached an "Examples" document. It has detailed explanations of predictive models and many examples of models and of types of outcomes. You do not need to refer to it to be able to go through the forms; I am only providing it as a detailed reference.

General Instructions

Your task is to annotate a set of 30 articles. Annotation is defined as your written response, based on your knowledge of molecular medicine research, to the questions on the paper form for each article. Please complete all 30 forms in this batch. When you are finished annotating this batch, please contact me to let me know and I will arrange to pick the paper forms from you and to deliver to you your compensation for participation. If you wish to do more annotations, then please let me know and I will prepare another batch of 30 papers that I will also arrange to deliver to you along with your compensation for the first batch. My aim is to collect as many expert annotations as the available funds allow. Having multiple batches from the same expert will be ideal for the design of this study. It is my personal experience that one's annotation will become faster and more efficient with practice.

This set of 30 articles is randomly sampled from 35 journals between January 1, 2006 and June 30, 2009. Some types of articles such as "Review Articles," "News," "Letters to the Editor," "Comments," or "Editorials" were removed from the original group of articles. The aim is to annotate articles that describe original clinical or basic science research. There are no "right" or "wrong" answers. I want to test whether my automatic annotation system can mirror your judgments about the questions in the form. When answering the questions, please use all resources that are available to you including the PubMed record and MeSH terms, the full text of the article or any other source of knowledge you think you need (Wikipedia, google searches, your course notes, etc.)

You may not need to answer all questions for every article. Questions 1 through 5 are "yes or no" questions. Your answer to these questions will determine how to proceed

and whether you will need to answer the four questions at the bottom of the page. The grey arrows are intended to guide the flow of your annotation based on your answers to these questions. These are the possible scenarios:

- If you circle "no" to **question #1**, then you do not need to go any further. <u>You are done annotating this article</u>.
- If you circle "yes" to **question #1** then you will need to answer all of the following questions: **Question #2, #3, #4, #5**, and **DISEASES**.
    - If you circle "yes" to question #3 then you will need to answer **BIOLOGIC SAMPLE** and **TYPE OF ASSAY**
    - If you circle "yes" to question #4 then you will need to answer **CLINICAL PURPOSE**

Specific Questions

Question #1

"Does the article describe at least one predictive model?"

For this study, an article describes a **predictive model** if the authors are trying to establish a **statistical relationship** between a set of **independent variables** and one or more **outcomes**.

Independent Variables

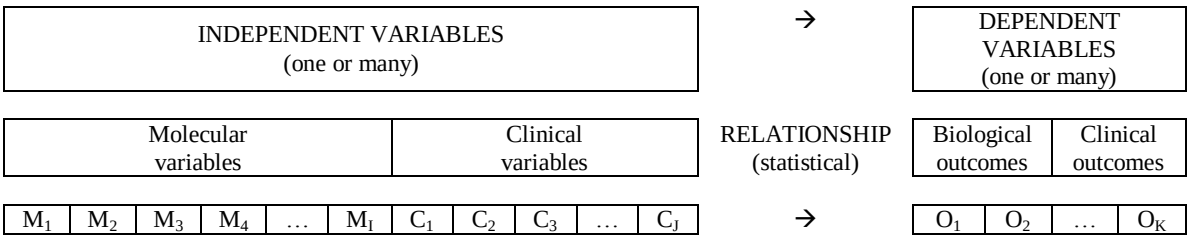
The **independent variables** can represent any type of quantifiable observations or experimental measurements. For example, the authors may be measuring local gene

expression levels, protein concentrations, or the presence or absence of proteins by using antibodies on biological samples. Sometimes the independent variables are clinical measurements or patient characteristics, such as blood pressure, sex, age, or variables that indicate the presence or absence of disease states (lymph node metastasis, histological subtype, etc.)

Dependent Variables or Outcome

The outcomes of interest are also quantifiable observations. They are based on the scientific hypothesis that the authors are investigating. If the authors are trying to establish that the outcomes of interest are related in a mathematical or statistical way to the independent variables (depend on them) then the paper describes a predictive model and the outcome variables are called **dependent variables.** The outcomes of interest can be classified as biological (e.g. cell apoptosis, activation of intra-cellular cascades, cell mobility, presence of a specific protein, etc) or clinical (patient death, presence or absence of disease, response to treatment, treatment toxicity, etc.)

A predictive model can be conceptually represented as follows:



Types of Relationships:

Statistical relationships can be assessed by tests for the probability of certain measurements occurring in two or more categories: Look for parametric tests such as *t-tests, ANOVA*, *fisher exact test*; or non-parametric tests such as *Kruskal –Wallis*, *Wilcoxon*, or *Mann-Whitney*.

Look also for models that measure statistical correlation between the values of independent variables and the values of dependent variables: *linear regression*, *multivariate linear or logistic regression.*

If the outcome of interest is the probability of occurrence of a given event, such as death or having metastasis, look for *Kaplan-Meier* or similar *survival functions*.

Sometimes the relationship between independent and dependent variable is presented via *mathematical equations* or via complicated so-called machine learning models. Examples of machine learning models include *artificial neural networks, support vector machines, decision trees*, or *Bayes classifiers*. The evaluation of such machine learning predictive models is typically reported using *sensitivity, specificity and ROC curve (AUC)*.

The following questions all assume that the paper describes one or more predictive models as described above, and that you have determined the independent and dependent variables for these models. Questions #2 and #3 relate to the independent variables. Questions #4 and #5 relate to the dependent variables (outcomes).

Question #2

"Is there a model that has more than one independent variable?"

If the article describes **more than one independent variable** then please circle question #2 "yes." In other words is the model a single or multivariate model?

CAVEAT: Sometimes it may seem from the abstract and title that the paper describes one single variable association with the dependent variable(s) when that is in fact not the case. For example, a paper may describe how the expression of a certain biomarker protein may be a predictor of breast cancer outcome. If you look at the full text of the paper (see last section on tips on how to quickly browse the full text of articles online) you will see that other independent variables were measured in the study and controlled in the final analysis as potential confounders. This is because there are many variables that are typically known to affect clinical outcome (e.g. clinical staging, histological type, and Estrogen Receptor status in breast cancer) and good statistical analysis always accounts for known confounders. In this case, this study – despite reporting one protein as an independent risk factor – is still considered as describing a multivariate prediction model.

Question #3

"Is at least one of the independent variables a molecular measurement?"

This question is self explanatory and easy once you have determined the independent variables. If any of the independent variables is the result of a molecular

assay, then please circle question #3 as "yes." The following table lists examples of molecular data as well as those that are not considered molecular data.

Independent Variables	
Molecular Variables (obtained from Molecular Assays)	Non-molecular Variables
• ELISA • Immunohistochemistry (including Tissue Micro Array) • Immunofluorescence • Flow Cytometry • Western, Northern, Southern, and Eastern blot • Any hybridization with probes • Gel Electrophoresis • DNA, cDNA and Oligonucleotide Microarrays • PCR, RT-PCR, qRT-PCR • Array CGH, FISH • Sequencing • SNP and Haplotype Chips • Mass Spectrometry • Methylation Assays Presence or absence of certain cell membrane receptors (usually determined by one of the methods above)	Clinical parameters such as: • Demographics: Age, sex, race • Histological and Clinical Staging of Tumors • Disease Diagnosis or Disease States (e.g. presence or absence of metastasis) • Blood pressure Clinical lab values such as: • Hematology (e.g. CBC) • Histology with regular H&E staining • Blood Chemistry (Electrolytes, Creatinine) • Iron and heavy metals • Liver function tests (ALT, AST, Bilirubin, LDH) • Urinalysis • Blood glucose • Cortisol and steroid tests Radiological Tests • X-Ray, CT • MRI, fMRI • PET, SPECT

Question #4

"Is one of the outcomes a clinical outcome?"

Is at least one the dependent variables (outcome) a measurement of a clinical outcome? ***Ask yourself if the outcome of the experiment or study mentioned is directly applicable to clinical care today***. For example, if the study measures the effect of certain genes on outcomes like apoptosis or DNA repair mechanism, then those outcomes are NOT clinical outcomes. Even though we know that damage to DNA repair is oncogenic

and the genes in question may prove useful in personalized medicine in the future, this study is not making a direct evidence link to a present day clinical outcome. This is similar to using a drug that has been proven to lower blood cholesterol level. Without a randomized controlled study that directly measures the association between cardiovascular outcomes (heart attack, stroke) and using this drug, we cannot say that that drug improves clinical outcomes. We can only say that that drug lowers blood cholesterol. The table "Identifying Clinical Outcomes" in the examples document has different scenarios of clinical outcomes in the second column.

Question #5

"Is one of the outcomes a biological outcome?"

Is at least one of the dependent variables (outcomes) a measurement of a non-clinical outcome? In other words does the article describe a predictive model that has at least one outcome that is a measurement of biological behavior that is not directly applicable to clinical care? ***Questions #4 and #5 are not mutually exclusive.*** If an article describes a study that measures molecular data and tries to associate that data with a direct clinical outcome AND to an outcome that provides a testable hypothesis about the biology of the disease then the answer "yes" applies to both questions #4 and #5. Examples of biological outcomes:

- Studies of intra-cellular effects or gene pathways: Activation or de-activation of a cascade of intracellular signals based on the value of independent molecular

variables such as the relationship between certain receptor proteins in cancer cells and the change in gene expression of other intracellular proteins

- The effect (activation, inhibition, or modulation) of a specific substances or drugs on intra cellular molecular pathways
- Cell biology outcomes such rate of cell growth or cell mobility studies.
- Studying basic human physiology: For example articles describing a clinical study whereby human subject characteristics (age, body mass, diet) are measured and are used to predict physiologic response to blood glucose challenge.
- Preclinical or Phase 0 clinical trials that study pharmacokinetics or pharmacodynamics of drug metabolism.

The table "Identifying Clinical Outcomes" in the examples document has more scenarios of non-clinical (biological) outcomes in the third column.

DISEASE

"Write all the diseases that are studied in this article"

If you answered "yes" to question #1, please write all the diseases for which the predictive model is applicable. Be as specific as possible. If the article describes a study on lung cancer patients and the article clearly states that all patients were included from a Non-Small Cell Lung Cancer (NSCLC) population, then please indicate NSCLC in the DISEASE box. If the article describes a study done on breast cancer cell line with no further specification, then please write "Breast Cancer." If the study is a basic biology research study and you cannot find the biological source of the cell samples or the cells

are obtained from normal non-diseased tissue then you can leave this box empty. Cell lines are usually referenced using a unique alpha-numeric identifier. If the specific origin of the cell line is not clearly indicated, I have found that a quick google search will usually return the origin. There are well known cell lines like HeLa (Cervical Cancer) and HEK-293 (Human Embryonic Kidney). If the response to question #4 (clinical outcome) is "yes" then it is almost certain that the disease will be specified in the paper.

BIOLOGIC SAMPLE

"What types of biological samples were used?"

Please circle all the answers that apply. The answers to this question may not always be evident from the abstract or title of the paper. This information is typically found in the "methods" section of the full article. This question is straight forward. If molecular assays were performed on tissue obtained from human patients (or healthy controls) then circle "Human." If the molecular assays were performed on tissue obtained from animals (including xenograft of human cells into mice models) then please circle "Animal." The molecular assays were applied to cell lines then please circle "Cell Line." If the molecular assays are targeting proteins and genes that are usually present in pathogens (viruses, bacteria, parasites, fungi) then please circle "Pathogen." Sometimes viral genome may be detected in human tissue without evidence of viral particles. It will usually be clear from the text of the article that the investigators were looking for or came upon viral genomic material. In this case, and even if the molecular assays were applied to human tissue, please circle "Pathogen" in addition to "Human."

TYPE OF ASSAY

"What biological molecule is measured by the molecular assay?"

If some of the independent variables were molecular variables, please classify the assays used to measure those variables into three categories. If the article describes more than one type of molecular assay, please circle all that apply. The classification of molecular assays for this annotation is roughly based on the type of molecule that is being targeted: DNA, RNA, or Protein. For example: Northern Blot → "RNA," Southern Blot → "DNA," Western Blot → "Protein." Please classify Methylation assays and other assays that measure epigenetic regulation under "DNA". Assays that can detect post-translational modifications of proteins such as Eastern Blot can be classified as "Protein."

CLINICAL PURPOSE

"What are the types of clinical outcomes

(dependent variables) that are described in the article?"

Please answer this question if you have answered "yes" to question #4 (clinical outcomes). There may be more than one clinical outcome discussed in the paper, so please circle all the classifications that apply. The middle column of the "Identifying Clinical Outcomes" table in the examples section has different scenarios where the dependent variables of the predictive model represent clinical outcomes. Here are the definitions of the different types of clinical outcomes:

Diagnosis: The purpose of a diagnostic predictive model is to detect or confirm the presence of a specific disease. One scenario is when models are used for *screening asymptomatic patients* e.g. using molecular cytology analysis of sputum in patients with smoking history for early detection of lung cancer. Another scenario is when models are used to identify *molecular subtypes* of a given disease or to help with *differential diagnosis* e.g. gene expression profiling of diffuse large B-Cell lymphoma (DLBCL) when a determination cannot be made by regular histology. Another scenario is when models can help in the *identification of the primary tumor* e.g. when there is a metastatic tumor of unknown origin and the molecular tests are used to find the organ of origin of the cancer lesion.

Risk Assessment: If the purpose of the model is to predict or quantify the risk that healthy patients will get a specific disease or clinical outcome. One example is using SNP arrays to identify whether people with certain haplotypes are at *higher risk to develop specific diseases*. Similar examples include *genetic testing* for high risk genes for cancers like BRCA. This may include models that are not based on molecular data. Obvious examples are epidemiological studies that *look for environmental risk factors* for lung cancer (smoking, pollution, occupation).

Prognosis, treatment unspecified: This category includes models whose purpose is to predict clinical outcome *irrespective of the type of treatment given* such as risk scores, or molecular models that try to predict the aggressiveness or metastatic potential of certain tumors. Models that predict risk of relapse after treatment should be included in this category.

Prognosis, treatment specified: This category includes models whose purpose is to study specific treatment outcome, sometimes in the presence of other modulating independent variables. This category would include: *phase II or III drug clinical trials*; studies that try to assess individual response to treatment or individual risk of drug adverse effects based on genetic testing (*personalized medicine / pharmacogenomics*); studies that aim to *select candidates for specific treatments* based on molecular tests e.g. selecting patients for adjuvant hormone therapy for breast cancer based on estrogen-receptor status of the cancer tissue.

Pointers and Tips

Finding Papers

The fastest way to find a paper online is to type the following URL substituting the ##### characters with the PubMed ID on your form.

http://www.ncbi.nlm.nih.gov/pubmed/#####

The page that will come up will display the PubMed record with the title and abstract. If you click on the "Publication Types, MeSH Terms, Substances, Grant Support" link you will find additional information about the paper that will help you in your annotation (MeSH terms, molecular tests, drug names). If you are accessing PubMed from Vanderbilt campus there will be links in the top right corner to the full text of the article.

You can have access to the full text from off campus if you visit (http://www.mc.vanderbilt.edu/diglib) and authenticate via your vunetid and epassword.

Quick Scanning of Online Article

When reading the full text of the article online, I found that it is easier to read the 'html' or 'full text' version of an article instead of the 'pdf' version. The pdf version is optimized for printing and it is hard to quickly scan if you are reading on a screen.

I found that most of the information that I needed for annotation can be found in the "methods" section and by scanning the figures in the "results" section. The molecular assays and biological samples are usually clearly listed in the methods section. There is usually a "statistical analysis" paragraph in the methods section that I found useful for determining the information about the independent variables and types of outcomes. There is usually one or more "patient characteristics" table that clearly shows the independent variables that are used for the study (and sometimes dependent variables as category columns).

The fastest way for me to scan an online article is by pressing "CTRL+F" (or "Cmd+F") which works in all browsers. When you press CTRL+F a small dialogue appears where you can type words or parts of words and the browser will take you quickly to the part of the page that matches what you typed. If you press enter you can skip through the document to every part where your search term matches. Some of the terms that I have used out of habit are:

- "blot", "immune", "fluor", "assay", "chip", "protein", "gene", "genom" , "pcr", "sequenc" for molecular variable scanning
- "prognosis", "survival", "treatment", "drug", "therapy" when I am looking for information about clinical outcomes

- “cell line”, “sample”, “culture”, “assay” when I am looking for the biological source

Examples

Identifying Predictive Models

Study Designs

To help you determine whether certain variables are dependent or independent, this section will highlight different types of study design. If you try to determine the kind of experimental design used in a paper, it may be clearer for you to identify the outcomes and/or the independent variables described in an article.

Studies with **clinical outcomes** generally fall into two categories based on the how they are structured to test the hypothesis:

1. In *cohort studies* or *randomized controlled trials* the patients are separated into two or more groups based on their characteristics or risks, i.e. **based on their independent variables**. In this case, the independent variables are assigned or controlled as if to conduct an experimental manipulation. Then the outcomes (dependent variables) are measured and the difference in outcome between the experimental groups is tested. For example, assume that the authors want to study the mortality associated with two drugs. Patients can be assigned to different treatment groups (treatment would be an independent variable) while other independent variables such as age, smoking, sex and blood pressure are also controlled between the two groups (because they are possible confounders). The outcome of interest (death) is then measured over a period

of time and the difference in the probability of death is measured between the two groups (typically via a survival function). The strength of statistical difference in outcome (dependent variable) in such studies is typically reported using *relative risk (RR)* or *hazards ratio*.

2. In case control studies, patients are grouped **based on the outcome (dependent variables)**, and differences in independent variables are measured and compared between the groups. For example, lung cancer patients are assigned to one group, and controlled healthy patients are assigned to another group. Independent variables, such as smoking, exposure to asbestos, and SNP mutations are measured. In the analysis, the difference in measurement of these risk factors is analyzed between the two groups (e.g. via multivariate regression). The strength of statistical difference between the independent risk factors is typically reported using *odds ratio (OR)*.

Studies with **biological outcomes** may also fall into two categories that mirror the study design mention above:

1. Example 1 – **assign to categories based on independent variables**: The investigators want to test the effect of the absence of a given gene on tumor growth and vascularization. They construct a xenograft mouse model and compare a wild-type group vs a gene knock-out or silenced gene (via siRNA) group. They compare tumor size after subjecting the mice to a given treatment. The dependent variable is the tumor size and the independent variable is the gene mutation status.
2. Example 2 – **assign to categories based on outcome**: The investigators measure the response of two cell line cultures that respond differently to a given chemotherapeutic agent. They conduct gene expression oligonucleotide microarray analysis of both cell

lines to measure the expression level of thousands of genes. They use statistical methods to look for differentially expressed genes between the two samples. The thousands of gene expression signals on the microarray chip are the independent variables, and the response to treatment is the outcome.

Examples of Articles That Do NOT Describe Predictive Models

Here are some examples of types of papers that generally do NOT describe predictive models and therefore you should circle "no" for the first question and move to the next article.

- **Descriptive Studies**: Some papers report on the health or genetic profile of an entire population but do not test any measurable differences for that population. Examples of such papers are ancestry studies that analyze human migration and genomic marker frequencies over geographic location or within ethnic groups. Other examples are papers that report prevalence data (e.g. of childhood cancers, or of specific gene mutations) from national registries without any experimental manipulation or statistical testing. A paper that presents the prevalence of smoking in a given population per se should not be included. ***[However a paper that presents a survey of smoking within a given population and that uses smoking status to predict other patient characteristics such as disease occurrence or low birth weight DOES include a predictive model and question #1 should be circled as "yes."]***
- **Bioinformatics Methods Papers**: If the purpose of the article is to describe a *new methodology for analysis* or *new experimental platforms* (e.g. new microarray chips) without reporting any clinical or biological experimental results then the paper does

not describe a predictive model. You may encounter bioinformatics papers that describe the development of new sequence alignment techniques, or new techniques for measuring gene expression signals from microarray chips. Some computational biology papers describe new algorithms for building machine learning models. If these papers are purely interested in the mathematical proof or theoretical limitations of such machine learning algorithms then question #1 should be circled "no." ***[However, sometimes bioinformatics investigators benchmark their machine learning algorithms using real experimental data. In this case, results are typically reported using sensitivity, specificity and ROC curves. These papers are considered as describing predictive models and question #1 should be circled "yes."]***

- **Statistics Methods Papers:** If you encounter a statistics or epidemiology paper that describes the *mathematical proof* behind a new statistics test, then that paper does not describe a predictive model. Similarly papers that describe new statistical analysis methods (based on mathematical proofs or on abstract computer simulations) in the field of statistical genetics (linkage disequilibrium, population genetics, etc) are also papers that do not fall in the predictive model category.
- **Biology Methods Papers:**
    - **Biochemistry/Structural Biology/ Biophysics/Chemistry:** Papers that describe 3D models of proteins or other molecular structures using crystallography or computer simulations. Biophysics papers that study cellular membrane stability, or electric voltage potential across a membrane. Papers that describe new mass spectrometry techniques. Papers that describe enzyme-substrate dynamics using

computer simulation. Papers that describe basic research into microRNA molecular structure by analyzing binding sites and structural motifs.

- **Biotechnology:** Papers that focus on the new mechanisms of vector construction or restriction fragment enzymes. Papers describing pharmacological/chemical techniques for discovering or synthesizing new drug molecules (without any specific drug or disease action).
- **Neurophysiology/Neuroanatomy:** Papers that analyze how the nervous system works without mentioning application to diagnosis or treatment of diseases (e.g. papers that report new brain or spinal cord connections; papers that simulate or analyze human cognition).
- **Systems Biology:** Papers describing genomes and gene circuitry of synthesized or model organisms such as yeast, bacteria, or viruses. For example, there are papers that simulate complex regulatory mechanisms (gene circuits for regulation of cell cycles and nutrient consumption) in such organisms using model computer simulations. If such models are not validated based on independent measurement and statistical correlation then these papers are not relevant.
- **Developmental Biology:** Papers that describe embryological development (notochord, germ layers, cellular differentiation, etc) should generally not be considered if there is no reporting of statistical analysis of experimental measurements, e.g. if the paper is qualitatively describing a stage of embryonic differentiation or providing pure descriptive statistics of developmental diseases. ***[However if the paper describes an experiment where genetic measurements***

(independent variables) are statistically correlated with developmental events (outcomes) then question #1 should be circled "yes."]

- **Radiology Methods Papers**: Papers describing new biomedical engineering techniques for image analysis, image reconstruction or signal processing. Papers describing new imaging modalities (e.g. new SPECT, fMRI techniques). ***[However if a papers statistically evaluates the ability of new radiology techniques (e.g. automatic detection of calcifications on mammograms) to predict clinical outcomes (e.g. screening for breast cancer) then it DOES include a predictive model and question #1 should be circled "yes."]***
- **Papers Describing Resources, New Research Centers, Research Cohorts, or Consortia**: For example, some papers report a new database for protein sequences, genetic diseases, or whole genome databases for model organisms like drosophila. Sometimes there are papers that describe the formation of a new research network, consortium or give descriptive statistics of new clinical cohorts (without providing any statistical hypothesis testing). Papers describing the establishment of new disease registries with some summary statistics.
- **Synthesis of Research and Prospective Papers**: Papers that present new guidelines that are proposed by professional societies such as new guidelines for diagnosing and treating asthma. Such papers are typically based on research that is published elsewhere and that is not directly presented in these papers. Papers that review or synthesize results from multiple other papers but do not describe the actual models or the statistical validation of the models. Prospective papers by seasoned researchers about the need for new research directions.

- **Case Reports**: Papers that report and describe cases of new, interesting, or very rare diseases without any statistical analysis. Such papers typically report a very small number of cases with a qualitative description of the disease symptoms and progression.
- **Errata, Comments, Letters, Editorials, Reviews:** I tried removing such papers from the main set. Some of them may have escaped my filters. Just answer the question #1 as "no."
- **Non-biological Papers:** Some of the journals like PLOS One or Nature may include astronomy, chemistry, or other non-biomedical disciplines. Just circle "no" for the first question.

Identifying Clinical Outcomes

Source of Biologic Sample	Clinical	Not Clinical
Human	Alteration in blood measurements that lead to a diagnosis. • For example correlating independent variables to high blood glucose (diabetes diagnosis) Detection and diagnosis of disease: • screening for cancer • confirming neurological or psychological conditions for example by imaging Help in making a differential diagnosis: • the model helps identify the type or origin of a cancer/leukemia based on molecular assays Assessment of risk to have disease: • predictive model for lung cancer risk from smoking • lifetime risk of getting a certain condition if you have an inherited	Alteration in blood measurements that have no direct clinical significance: • some protein that has no clinical significance or that significance is suspected but not yet confirmed The outcome is a risk factor, but no direct link to a known disease is established yet: • Identifying independent reasons that can predict whether someone will have elevated cholesterol later in life. Independent reasons not yet directly tied to the bad outcomes of high cholesterol Alteration in histologic characteristics also of unknown significance: • Molecular assays find traces of a virus DNA in cancer tissues.

Source of Biologic Sample	Clinical	Not Clinical
	mutation Prognosis / disease outcome: • Mortality/Survival • Disease-free survival • Risk of metastasis or exacerbation of disease Response to specific treatment: • Stage I clinical trial: drug toxicity • Stage II and III clinical trials, showing measured clinical improvement from certain drug regimens compared to others • Risk of toxicity, such as chemotherapy adverse effects tied to genetic markers (personalized medicine) • Finding that a mutation in the cancer tissue makes the patient resistant to a given chemotherapy agent • Tailoring the chemotherapy regimen based on molecular assay	This finding has not yet been linked to adverse outcome Physiologic studies for basic research: • A clinical study that measures the glucose metabolism of different healthy people based on clinical characteristics and/or molecular assays Drug metabolism investigation: • A clinical trial on healthy volunteers (e.g. Phase I) that measures pharmacokinetics and pharmacodynamics of a given drug Basic research in genetics • Looking for association between different haplotypes and non-clinical outcomes such as metabolism or physiologic variability.
Animal Model	When analyzing the type of outcome in animal models, apply the same reasoning used for human outcomes above. See extra cases that do not apply →	When analyzing the type of outcome in animal models, apply the same reasoning used for human outcomes above. There are types of outcomes that should be considered non-clinical. Specifically, when the outcome (dependent variable) in the study protocol cannot be applied to humans for obvious reasons: • Diagnosis: Inducing a disease and then testing whether a new diagnostic test can be used to detect the disease • Risk Assessment: Testing carcinogenicity of substances by giving the animals very large doses of the substance that is being studied. • Prognosis: Time-series studies where a few mice are sacrificed every day to study disease progression, e.g. by measuring tumor size • Response to Treatment: Treatment is administered in a way that never applies to humans

Source of Biologic Sample	Clinical	Not Clinical
		e.g. inducing brain metastasis and then administering therapy directly into animal brains
Cell Line	In vitro studies that are part of preclinical investigation: • Diagnosis: The cell line is derived from human disease AND the biomarker assay can be directly tested in in vivo pre clinical studies relating to same disease • Prognosis: The cell line is derived from human disease AND the experiment is measuring different markers that predict aggressiveness or remission for same disease • Treatment response: The cell line is derived from human disease AND the substance that is or will be used to treat the same disease in humans is applied during the cell line experiment to study differential outcomes	Almost all other cases involving cell lines. Here are some examples: • Basic research: investigation into intra cellular pathways and associations between different genes and/or proteins groups. • Cancer cell motility studies • Cancer cell metabolism studies such as studying the rate that different substances are metabolized by cancer cells • Drug discovery: applying a battery of substances to find possible reactions • Studies of cell potency (stem cells) and cell differentiation based on molecular profiling
Pathogen	Infectious Disease: • Diagnosis: confirming the diagnosis of infection by detecting the presence of the pathogen e.g. by PCR • Prognosis: Molecular assays that measure the virulence of the infectious disease pathogen and can be used to assess the risk to the host e.g. molecular subtypes • Treatment response: Molecular assays that can predict the drug resistance profile of the pathogen Pathogen is known to cause an increased risk of neoplasm or known to alter the outcome of existing neoplasm.	Genome sequencing, phylogenetics. Basic investigation into mechanism of disease. Molecular disease epidemiology: Descriptive statistics showing the prevalence of different molecular viral subtypes across geographic regions without correlation with specific outcome. Confirming presence or absence of pathogen in tissue when the presence or absence does not affect clinical outcome or has no known clinical significance yet.

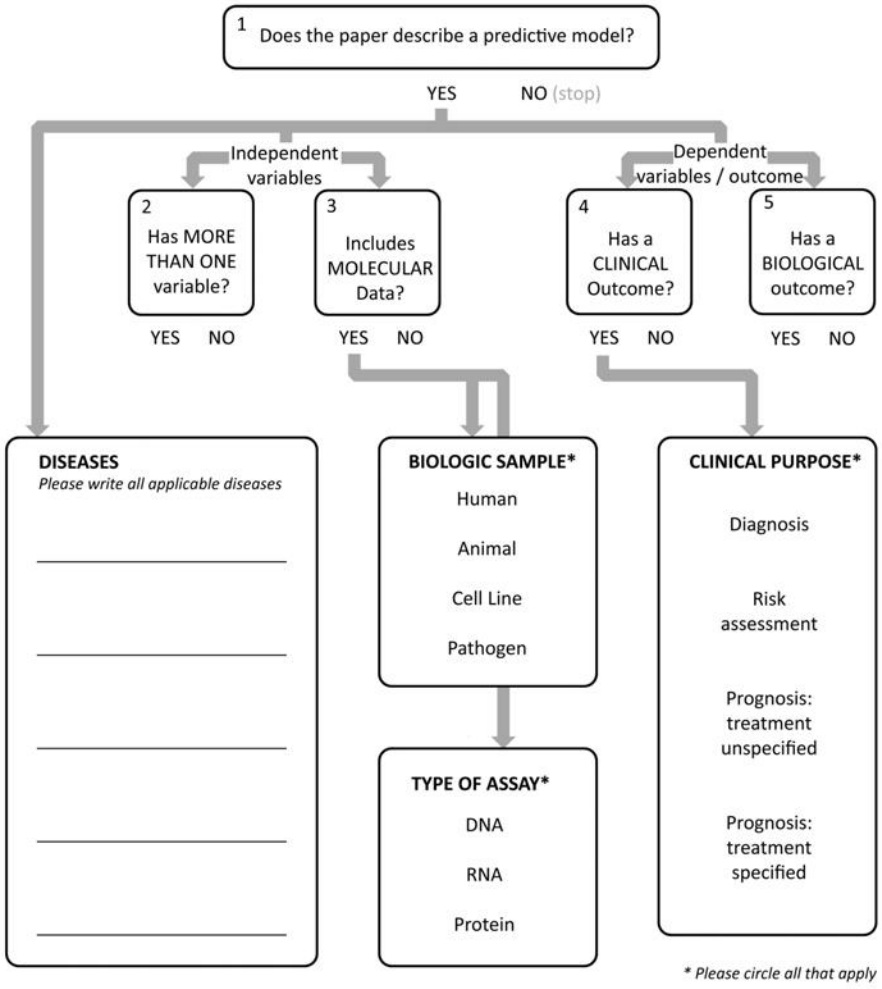
APPENDIX C

SAMPLE ANNOTATION FORM

PubMed ID: **17878312** Annotator: **S**

Phospholipase Cepsilon is a nexus for Rho and Rap-mediated G protein-coupled receptor-induced astrocyte proliferation.

Proc Natl Acad Sci U S A 2007 Sep 25;104(39):15543-8



REFERENCES

1. Sobie EA, Guatimosim S, Song L-S, Lederer WJ. The challenge of molecular medicine: complexity versus Occam's razor. *J. Clin. Invest*. 2003;111(6):801-803. Available at: Accessed December 3, 2009.

2. Quackenbush J. Microarray analysis and tumor classification. *N. Engl. J. Med*. 2006;354(23):2463-2472. Available at: Accessed December 3, 2009.

3. Ntzani EE, Ioannidis JPA. Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. *Lancet*. 2003;362(9394):1439-1444. Available at: Accessed December 3, 2009.

4. Ross JS, Ginsburg GS. The integration of molecular diagnostics with therapeutics. Implications for drug development and pathology practice. *Am. J. Clin. Pathol*. 2003;119(1):26-36. Available at: Accessed December 3, 2009.

5. van \uc0\u8217{}t Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*. 2002;415(6871):530-536. Available at: Accessed December 3, 2009.

6. Simon R. Roadmap for developing and validating therapeutically relevant genomic classifiers. *J. Clin. Oncol*. 2005;23(29):7332-7341. Available at: Accessed December 3, 2009.

7. Office of the Commissioner. FDA Clears Breast Cancer Specific Molecular Prognostic Test. Available at: http://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/2007/ucm108836.htm. Accessed December 3, 2009.

8. Glas AM, Floore A, Delahaye LJMJ, et al. Converting a breast cancer microarray signature into a high-throughput diagnostic test. *BMC Genomics*. 2006;7:278. Available at: Accessed December 3, 2009.

9. Paik S, Shak S, Tang G, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N. Engl. J. Med*. 2004;351(27):2817-2826. Available at: Accessed December 3, 2009.

10. Center for Devices and Radiological Health. Draft Guidance for Industry, Clinical Laboratories, and FDA Staff - In Vitro Diagnostic Multivariate Index Assays. Available at: http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm079148.htm. Accessed December 3, 2009.

11. Center for Drug Evaluation and Research. Table of Valid Genomic Biomarkers in the Context of Approved Drug Labels. Available at:

http://www.fda.gov/Drugs/ScienceResearch/ResearchAreas/Pharmacogenetics/ucm083378.htm. Accessed December 3, 2009.

12. Marsh S, McLeod HL. Pharmacogenomics: from bedside to clinical practice. *Hum. Mol. Genet.* 2006;15 Spec No 1:R89-93. Available at: Accessed December 3, 2009.

13. Ross JS, Schenkein DP, Kashala O, et al. Pharmacogenomics. *Adv Anat Pathol*. 2004;11(4):211-220. Available at: Accessed December 3, 2009.

14. Innocenti F, Undevia SD, Iyer L, et al. Genetic variants in the UDP-glucuronosyltransferase 1A1 gene predict the risk of severe neutropenia of irinotecan. *J. Clin. Oncol*. 2004;22(8):1382-1388. Available at: Accessed December 3, 2009.

15. Aliferis CF, Statnikov A, Tsamardinos I, et al. Factors influencing the statistical power of complex data analysis protocols for molecular signature development from microarray data. *PLoS ONE*. 2009;4(3):e4922. Available at: Accessed December 3, 2009.

16. Statnikov A, Li C, Aliferis CF. A statistical reappraisal of the findings of an esophageal cancer genome-wide association study. *Cancer Res*. 2008;68(8):3074-3075; author reply 3075. Available at: Accessed December 3, 2009.

17. Altman RB, Flockhart DA, Sherry ST, et al. Indexing pharmacogenetic knowledge on the World Wide Web. *Pharmacogenetics*. 2003;13(1):3-5. Available at: Accessed December 3, 2009.

18. Oliver DE, Rubin DL, Stuart JM, et al. Ontology development for a pharmacogenetics knowledge base. *Pac Symp Biocomput*. 2002:65-76. Available at: Accessed December 3, 2009.

19. Hewett M, Oliver DE, Rubin DL, et al. PharmGKB: the Pharmacogenetics Knowledge Base. *Nucleic Acids Res*. 2002;30(1):163-165. Available at: Accessed December 3, 2009.

20. Rhodes DR, Yu J, Shanker K, et al. ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia*. 2004;6(1):1-6. Available at: Accessed December 3, 2009.

21. Rhodes DR, Kalyana-Sundaram S, Mahavisno V, et al. Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia*. 2007;9(2):166-180. Available at: Accessed December 3, 2009.

22. Barrett T, Troup DB, Wilhite SE, et al. NCBI GEO: mining tens of millions of expression profiles--database and tools update. *Nucleic Acids Res*. 2007;35(Database issue):D760-765. Available at: Accessed December 3, 2009.

23. Barrett T, Troup DB, Wilhite SE, et al. NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res*. 2009;37(Database issue):D885-890. Available at: Accessed December 3, 2009.

24. Brazma A, Hingamp P, Quackenbush J, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet*. 2001;29(4):365-371. Available at: Accessed December 3, 2009.

25. Ball CA, Brazma A, Causton H, et al. Submission of microarray data to public repositories. *PLoS Biol*. 2004;2(9):E317. Available at: Accessed December 3, 2009.

26. Zhao Y, Simon R. BRB-ArrayTools Data Archive for Human Cancer Gene Expression: A Unique and Efficient Data Sharing Resource. *Cancer Inform*. 2008;6:9-15. Available at: Accessed December 3, 2009.

27. National Cancer Institute. REMBRANDT - Repository for Molecular Brain Neoplasia Data. Available at: https://caintegrator.nci.nih.gov/rembrandt/. Accessed December 3, 2009.

28. Culhane AC, Schwarzl T, Sultana R, et al. GeneSigDB--a curated database of gene expression signatures. *Nucleic Acids Res*. 2009. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19934259. Accessed December 8, 2009.

29. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. *Methods Inf Med*. 1998;37(4-5):394-403. Available at: Accessed December 3, 2009.

30. Shah NH, Jonquet C, Chiang AP, et al. Ontology-driven indexing of public datasets for translational bioinformatics. *BMC Bioinformatics*. 2009;10 Suppl 2:S1. Available at: Accessed December 3, 2009.

31. International Health Terminology Standards Development Organisation. IHTSDO: SNOMED CT. Available at: http://www.ihtsdo.org/snomed-ct/. Accessed December 3, 2009.

32. Sioutos N, de Coronado S, Haber MW, et al. NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. *J Biomed Inform*. 2007;40(1):30-43. Available at: Accessed December 3, 2009.

33. Shah NH, Bhatia N, Jonquet C, et al. Comparison of concept recognizers for building the Open Biomedical Annotator. *BMC Bioinformatics*. 2009;10 Suppl 9:S14. Available at: Accessed December 3, 2009.

34. Rubin DL, Lewis SE, Mungall CJ, et al. National Center for Biomedical Ontology: advancing biomedicine through structured organization of scientific knowledge. *OMICS*. 2006;10(2):185-198. Available at: Accessed December 3, 2009.

35. Aitken JS, Webber BL, Bard JBL. Part-of relations in anatomy ontologies: a proposal for RDFS and OWL formalisations. *Pac Symp Biocomput*. 2004:166-177. Available at: Accessed December 3, 2009.

36. Smith B, Ceusters W, Klagges B, et al. Relations in biomedical ontologies. *Genome Biol*. 2005;6(5):R46. Available at: Accessed December 3, 2009.

37. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet*. 2000;25(1):25-29. Available at: Accessed December 3, 2009.

38. Hoffman M, Arnoldi C, Chuang I. The clinical bioinformatics ontology: a curated semantic network utilizing RefSeq information. *Pac Symp Biocomput*. 2005:139-150. Available at: Accessed December 3, 2009.

39. Manduchi E, Grant GR, He H, et al. RAD and the RAD Study-Annotator: an approach to collection, organization and exchange of all relevant information for high-throughput gene expression studies. *Bioinformatics*. 2004;20(4):452-459. Available at: Accessed December 3, 2009.

40. Jones A, Hunt E, Wastling JM, Pizarro A, Stoeckert CJ. An object model and database for functional genomics. *Bioinformatics*. 2004;20(10):1583-1590. Available at: Accessed December 3, 2009.

41. Jones AR, Pizarro A, Spellman P, Miller M. FuGE: Functional Genomics Experiment Object Model. *OMICS*. 2006;10(2):179-184. Available at: Accessed December 3, 2009.

42. Alonso-Calvo R, Maojo V, Billhardt H, et al. An agent- and ontology-based system for integrating public gene, protein, and disease databases. *J Biomed Inform*. 2007;40(1):17-29. Available at: Accessed December 3, 2009.

43. Pérez-Rey D, Maojo V, García-Remesal M, et al. ONTOFUSION: ontology-based integration of genomic and clinical databases. *Comput. Biol. Med*. 2006;36(7-8):712-730. Available at: Accessed December 3, 2009.

44. Baader F, Calvanese D, McGuinness D, Daniele N, Patel-Schneider P eds. *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press; 2003. Available at: .

45. McGuinness D, van Harmelen F. OWL Web Ontology Language Overview - W3C Recommendation. Available at: http://www.w3.org/TR/owl-features/. Accessed December 3, 2009.

46. Cakmak A, Ozsoyoglu G. Discovering gene annotations in biomedical text databases. *BMC Bioinformatics*. 2008;9:143. Available at: Accessed December 3, 2009.

47. Camon EB, Barrell DG, Dimmer EC, et al. An evaluation of GO annotation retrieval for BioCreAtIvE and GOA. *BMC Bioinformatics*. 2005;6 Suppl 1:S17. Available at: Accessed December 3, 2009.

48. Couto FM, Silva MJ, Lee V, et al. GOAnnotator: linking protein GO annotations to evidence text. *J Biomed Discov Collab*. 2006;1:19. Available at: Accessed December 3, 2009.

49. Hill DP, Smith B, McAndrews-Hill MS, Blake JA. Gene Ontology annotations: what they mean and where they come from. *BMC Bioinformatics*. 2008;9 Suppl 5:S2. Available at: Accessed December 3, 2009.

50. Winnenburg R, Wächter T, Plake C, Doms A, Schroeder M. Facts from text: can text mining help to scale-up high-quality manual curation of gene products with ontologies? *Brief. Bioinformatics*. 2008;9(6):466-478. Available at: Accessed December 3, 2009.

51. Baumgartner WA, Cohen KB, Fox LM, Acquaah-Mensah G, Hunter L. Manual curation is not sufficient for annotation of genomic databases. *Bioinformatics*. 2007;23(13):i41-48. Available at: Accessed December 3, 2009.

52. Rothschild AS, Lehmann HP, Hripcsak G. Inter-rater agreement in physician-coded problem lists. *AMIA Annu Symp Proc*. 2005:644-648. Available at: Accessed December 3, 2009.

53. Caporaso JG, Deshpande N, Fink JL, et al. Intrinsic evaluation of text mining tools may not predict performance on realistic tasks. *Pac Symp Biocomput*. 2008:640-651. Available at: Accessed December 3, 2009.

54. Berardini TZ, Mundodi S, Reiser L, et al. Functional annotation of the Arabidopsis genome using controlled vocabularies. *Plant Physiol*. 2004;135(2):745-755. Available at: Accessed December 3, 2009.

55. Crosby MA, Goodman JL, Strelets VB, Zhang P, Gelbart WM. FlyBase: genomes by the dozen. *Nucleic Acids Res*. 2007;35(Database issue):D486-491. Available at: Accessed December 3, 2009.

56. Bodenreider O, Smith B, Kumar A, Burgun A. Investigating subsumption in SNOMED CT: an exploration into large description logic-based biomedical terminologies. *Artif Intell Med*. 2007;39(3):183-195. Available at: Accessed December 3, 2009.

57. Schulz S, Suntisrivaraporn B, Baader F. SNOMED CT's problem list: ontologists' and logicians' therapy suggestions. *Stud Health Technol Inform*. 2007;129(Pt 1):802-806. Available at: Accessed December 3, 2009.

58. Rector AL, Brandt S. Why do it the hard way? The case for an expressive description logic for SNOMED. *J Am Med Inform Assoc*. 2008;15(6):744-751. Available at: Accessed December 3, 2009.

59. Chapman WW, Dowling JN. Inductive creation of an annotation schema for manually indexing clinical conditions from emergency department reports. *J Biomed Inform*. 2006;39(2):196-208. Available at: Accessed December 3, 2009.

60. Hripcsak G, Heitjan DF. Measuring agreement in medical informatics reliability studies. *J Biomed Inform*. 2002;35(2):99-110. Available at: Accessed December 3, 2009.

61. Hripcsak G, Rothschild AS. Agreement, the f-measure, and reliability in information retrieval. *J Am Med Inform Assoc*. 2005;12(3):296-298. Available at: Accessed December 3, 2009.

62. Giuse NB, Giuse DA, Miller RA, et al. Evaluating consensus among physicians in medical knowledge base construction. *Methods Inf Med*. 1993;32(2):137-145. Available at: Accessed December 3, 2009.

63. Chapman WW, Dowling JN, Hripcsak G. Evaluation of training with an annotation schema for manual annotation of clinical conditions from emergency department reports. *Int J Med Inform*. 2008;77(2):107-113. Available at: Accessed December 3, 2009.

64. Leitner F, Krallinger M, Rodriguez-Penagos C, et al. Introducing meta-services for biomedical information extraction. *Genome Biol*. 2008;9 Suppl 2:S6. Available at: Accessed December 3, 2009.

65. Garten Y, Altman RB. Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text. *BMC Bioinformatics*. 2009;10 Suppl 2:S6. Available at: Accessed December 3, 2009.

66. Aphinyanaphongs Y, Tsamardinos I, Statnikov A, Hardin D, Aliferis CF. Text categorization models for high-quality article retrieval in internal medicine. *J Am Med Inform Assoc*. 2005;12(2):207-216. Available at: Accessed December 3, 2009.

67. Mons B, Ashburner M, Chichester C, et al. Calling on a million minds for community annotation in WikiProteins. *Genome Biol*. 2008;9(5):R89. Available at: Accessed December 3, 2009.

68. Hoehndorf R, Bacher J, Backhaus M, et al. BOWiki: an ontology-based wiki for annotation of data and integration of knowledge in biology. *BMC Bioinformatics*. 2009;10 Suppl 5:S5. Available at: Accessed December 3, 2009.

69. Stokes TH, Torrance JT, Li H, Wang MD. ArrayWiki: an enabling technology for sharing public microarray data repositories and meta-analyses. *BMC Bioinformatics*. 2008;9 Suppl 6:S18. Available at: Accessed December 3, 2009.

70. Tonellato P. Methodology and Infrastructure for Translational Science. 2009. Available at: http://lpm.hms.harvard.edu/palaver/sites/default/files/092809.pdf. Accessed December 3, 2009.

71. Müller H-M, Kenny EE, Sternberg PW. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol*. 2004;2(11):e309. Available at: Accessed December 3, 2009.

72. Jin B, Muller B, Zhai C, Lu X. Multi-label literature classification based on the Gene Ontology graph. *BMC Bioinformatics*. 2008;9:525. Available at: Accessed December 3, 2009.

73. Névéol A, Shooshan SE, Humphrey SM, Mork JG, Aronson AR. A recent advance in the automatic indexing of the biomedical literature. *J Biomed Inform*. 2009;42(5):814-823. Available at: Accessed December 3, 2009.

74. Aphinyanaphongs Y, Statnikov A, Aliferis CF. A comparison of citation metrics to machine learning filters for the identification of high quality MEDLINE documents. *J Am Med Inform Assoc*. 2006;13(4):446-455. Available at: Accessed December 3, 2009.

75. Haynes RB, Wilczynski N, McKibbon KA, Walker CJ, Sinclair JC. Developing optimal search strategies for detecting clinically sound studies in MEDLINE. *J Am Med Inform Assoc*. 1994;1(6):447-458. Available at: Accessed December 3, 2009.

76. Rodriguez-Esteban R, Iossifov I, Rzhetsky A. Imitating manual curation of text-mined facts in biomedicine. *PLoS Comput. Biol*. 2006;2(9):e118. Available at: Accessed December 3, 2009.

77. Burkhardt K, Schneider B, Ory J. A biocurator perspective: annotation at the Research Collaboratory for Structural Bioinformatics Protein Data Bank. *PLoS Comput. Biol*. 2006;2(10):e99. Available at: Accessed December 3, 2009.

78. Eaton AD. HubMed: a web-based biomedical literature search interface. *Nucleic Acids Res*. 2006;34(Web Server issue):W745-747. Available at: Accessed December 3, 2009.

79. Hoffmann R, Valencia A. A gene network for navigating the literature. *Nat. Genet*. 2004;36(7):664. Available at: Accessed December 3, 2009.

80. Hoffmann R, Valencia A. Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics*. 2005;21 Suppl 2:ii252-258. Available at: Accessed December 3, 2009.

81. Rebholz-Schuhmann D, Kirsch H, Arregui M, et al. EBIMed--text crunching to gather facts for proteins from Medline. *Bioinformatics*. 2007;23(2):e237-244. Available at: Accessed December 3, 2009.

82. Doms A, Schroeder M. GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Res*. 2005;33(Web Server issue):W783-786. Available at: Accessed December 3, 2009.

83. Plake C, Schiemann T, Pankalla M, Hakenberg J, Leser U. AliBaba: PubMed as a graph. *Bioinformatics*. 2006;22(19):2444-2445. Available at: Accessed December 3, 2009.

84. Elkin PL, Tuttle MS, Trusko BE, Brown SH. BioProspecting: novel marker discovery obtained by mining the bibleome. *BMC Bioinformatics*. 2009;10 Suppl 2:S9. Available at: Accessed December 3, 2009.

85. Roberts A, Gaizauskas R, Hepple M, et al. The CLEF corpus: semantic annotation of clinical text. *AMIA Annu Symp Proc*. 2007:625-629. Available at: Accessed December 3, 2009.

86. Roberts A, Gaizauskas R, Hepple M, et al. Building a semantically annotated corpus of clinical texts. *J Biomed Inform*. 2009;42(5):950-966. Available at: Accessed December 3, 2009.

87. Coulet A, Shah N, Hunter L, Barral C, Altman RB. Extraction of genotype-phenotype-drug relationships from text: from entity recognition to bioinformatics application. *Pac Symp Biocomput*. 2010:485-487. Available at: Accessed December 3, 2009.

88. Garten Y, Tatonetti NP, Altman RB. Improving the prediction of pharmacogenes using text-derived drug-gene relationships. *Pac Symp Biocomput*. 2010:305-314. Available at: Accessed December 3, 2009.

89. Gerstein M, Seringhaus M, Fields S. Structured digital abstract makes text mining easy. *Nature*. 2007;447(7141):142. Available at: Accessed December 3, 2009.

90. Ceol A, Chatr-Aryamontri A, Licata L, Cesareni G. Linking entries in protein interaction database to structured text: the FEBS Letters experiment. *FEBS Lett*. 2008;582(8):1171-1177. Available at: Accessed December 3, 2009.

91. Chatr-aryamontri A, Ceol A, Palazzi LM, et al. MINT: the Molecular INTeraction database. *Nucleic Acids Res*. 2007;35(Database issue):D572-574. Available at: Accessed December 3, 2009.

92. Copestake A, Corbett P, Murray-Rust P, et al. An architecture for language processing for scientific texts. In: *Proceedings of the UK e-Science All Hands Meeting 2006.*; 2006. Available at: .

93. Leitner F, Valencia A. A text-mining perspective on the requirements for electronically annotated abstracts. *FEBS Lett*. 2008;582(8):1178-1181. Available at: Accessed December 3, 2009.

94. Cimiano P. *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. New York, USA: Springer; 2006. Available at: .

95. Krauthammer M, Nenadic G. Term identification in the biomedical literature. *J Biomed Inform*. 2004;37(6):512-526. Available at: Accessed December 3, 2009.

96. Nenadic G, Spasic I, Ananiadou S. Terminology-driven mining of biomedical literature. *Bioinformatics*. 2003;19(8):938-943. Available at: Accessed December 3, 2009.

97. Elkin PL, Brown SH, Husser CS, et al. Evaluation of the content coverage of SNOMED CT: ability of SNOMED clinical terms to represent clinical problem lists. *Mayo Clin. Proc.* 2006;81(6):741-748. Available at: Accessed December 3, 2009.

98. Hersh WR. *Information Retrieval: A Health and Biomedical Perspective*. Second. Springer; 2003. Available at: .

99. Baeza-Yates R, Ribeiro-Neto B. *Modern Information Retrieval*. Addison Wesley; 1999. Available at: .

100. Lourenço A, Carreira R, Carneiro S, et al. @Note: a workbench for biomedical text mining. *J Biomed Inform*. 2009;42(4):710-720. Available at: Accessed December 4, 2009.

101. Hripcsak G, Knirsch C, Zhou L, Wilcox A, Melton GB. Using discordance to improve classification in narrative clinical databases: an application to community-acquired pneumonia. *Comput. Biol. Med*. 2007;37(3):296-304. Available at: Accessed December 4, 2009.

102. Cano C, Monaghan T, Blanco A, Wall DP, Peshkin L. Collaborative text-annotation resource for disease-centered relation extraction from biomedical text. *J Biomed Inform*. 2009;42(5):967-977. Available at: Accessed December 4, 2009.

103. Ando Y, Saka H, Ando M, et al. Polymorphisms of UDP-glucuronosyltransferase gene and irinotecan toxicity: a pharmacogenetic analysis. *Cancer Res*. 2000;60(24):6921-6926. Available at: Accessed February 16, 2011.

104. Ciotti M, Chen F, Rubaltelli FF, Owens IS. Coding defect and a TATA box mutation at the bilirubin UDP-glucuronosyltransferase gene cause Crigler-Najjar type I disease. *Biochim. Biophys. Acta*. 1998;1407(1):40-50. Available at: Accessed February 16, 2011.

105. Bell DW, Lynch TJ, Haserlat SM, et al. Epidermal growth factor receptor mutations and gene amplification in non-small-cell lung cancer: molecular analysis of the IDEAL/INTACT gefitinib trials. *J. Clin. Oncol*. 2005;23(31):8081-8092. Available at: Accessed February 16, 2011.

106. Lynch TJ, Bell DW, Sordella R, et al. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *N. Engl. J. Med*. 2004;350(21):2129-2139. Available at: Accessed February 16, 2011.

107. Couzin J. Diagnostics. Amid debate, gene-based cancer test approved. *Science*. 2007;315(5814):924. Available at: Accessed February 17, 2011.

108. Mathew JP, Taylor BS, Bader GD, et al. From bytes to bedside: data integration and computational biology for translational cancer research. *PLoS Comput. Biol*. 2007;3(2):e12. Available at: Accessed February 17, 2011.

109. Anon. Broad Institute - Datasets. 2008. Available at: http://www.broad.mit.edu/tools/data.html.

110. Anon. Broad Institute - Software. 2008. Available at: http://www.broad.mit.edu/tools/software.html.

111. Wright G, Tan B, Rosenwald A, et al. A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma. *Proc. Natl. Acad. Sci. U.S.A.* 2003;100(17):9991-9996. Available at: Accessed February 18, 2011.

112. Vose JM. Current approaches to the management of non-Hodgkin's lymphoma. *Semin. Oncol*. 1998;25(4):483-491. Available at: Accessed February 18, 2011.

113. Alizadeh AA, Eisen MB, Davis RE, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*. 2000;403(6769):503-511. Available at: Accessed February 18, 2011.

114. Rosenwald A, Wright G, Chan WC, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N. Engl. J. Med.* 2002;346(25):1937-1947. Available at: Accessed February 18, 2011.

115. Anon. A predictive model for aggressive non-Hodgkin's lymphoma. The International Non-Hodgkin's Lymphoma Prognostic Factors Project. *N. Engl. J. Med.* 1993;329(14):987-994. Available at: Accessed February 18, 2011.

116. Shipp MA, Ross KN, Tamayo P, et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat. Med.* 2002;8(1):68-74. Available at: Accessed February 18, 2011.

117. Li L. Survival prediction of diffuse large-B-cell lymphoma based on both clinical and gene expression information. *Bioinformatics*. 2006;22(4):466-471. Available at: Accessed February 18, 2011.

118. Pepe MS, Etzioni R, Feng Z, et al. Phases of biomarker development for early detection of cancer. *J. Natl. Cancer Inst*. 2001;93(14):1054-1061. Available at: Accessed February 18, 2011.

119. Erban JK, Lau J. On the toxicity of chemotherapy for breast cancer--the need for vigilance. *J. Natl. Cancer Inst*. 2006;98(16):1096-1097. Available at: Accessed February 18, 2011.

120. Hassett MJ, O'Malley AJ, Pakes JR, Newhouse JP, Earle CC. Frequency and cost of chemotherapy-related serious adverse effects in a population sample of women with breast cancer. *J. Natl. Cancer Inst*. 2006;98(16):1108-1117. Available at: Accessed February 18, 2011.

121. van de Vijver MJ, He YD, van't Veer LJ, et al. A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med*. 2002;347(25):1999-2009. Available at: Accessed February 18, 2011.

122. Weigelt B, Hu Z, He X, et al. Molecular portraits and 70-gene prognosis signature are preserved throughout the metastatic process of breast cancer. *Cancer Res*. 2005;65(20):9155-9158. Available at: Accessed February 18, 2011.

123. Buyse M, Loi S, van't Veer L, et al. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *J. Natl. Cancer Inst*. 2006;98(17):1183-1192. Available at: Accessed February 18, 2011.

124. Porter MF. An algorithm for suffix stripping. *Program: electronic library and information systems*. 2006;40(3):211-218. Available at: Accessed February 28, 2011.

125. Leopold E, Kindermann J. Text Categorization with Support Vector Machines. How to Represent Texts in Input Space? *Machine Learning*. 2002;46:423-444. Available at: .

126. Burges CJC. A tutorial on Support Vector Machines for pattern recognition. *DATA MINING AND KNOWLEDGE DISCOVERY*. 1998;2(2):121-167. Available at: .

127. Muller K-R, Mika S, Ratsch G, Tsuda K, Scholkopf B. An introduction to kernel-based learning algorithms. *Neural Networks, IEEE Transactions on*. 2001;12(2):181-201. Available at: .

128. Denny JC, Irani PR, Wehbe FH, Smithers JD, Spickard A. The KnowledgeMap project: development of a concept-based medical school curriculum database. *AMIA Annu Symp Proc*. 2003:195-199. Available at: Accessed December 3, 2009.

129. Denny JC, Smithers JD, Miller RA, Spickard A. "Understanding" medical school curriculum content using KnowledgeMap. *J Am Med Inform Assoc*. 2003;10(4):351-362. Available at: Accessed February 28, 2011.

130. Burgun A. Desiderata for domain reference ontologies in biomedicine. *J Biomed Inform*. 2006;39(3):307-313. Available at: Accessed February 18, 2011.