Biomedical Informatics Insights

OPEN ACCESS
Full open access to this and thousands of other papers at http://www.la-press.com.

ORIGINAL RESEARCH

Text Categorization of Heart, Lung, and Blood Studies in the Database of Genotypes and Phenotypes (dbGaP) Utilizing *n*-grams and Metadata Features

Mindy K. Ross[1,2], Ko-Wei Lin[2], Karen Truong[3], Abhishek Kumar[4] and Mike Conway[2]

[1]Department of Pediatrics, Division of Respiratory Medicine, University of California, San Diego, USA. [2]Department of Medicine, Division of Biomedical Informatics, University of California, San Diego, USA. [3]University of California, Los Angeles, USA. [4]Department of Computer Science and Engineering, University of California, San Diego, USA.
Corresponding author email: mkross@ucsd.edu

Abstract: The database of Genotypes and Phenotypes (dbGaP) allows researchers to understand phenotypic contribution to genetic conditions, generate new hypotheses, confirm previous study results, and identify control populations. However, effective use of the database is hindered by suboptimal study retrieval. Our objective is to evaluate text classification techniques to improve study retrieval in the context of the dbGaP database. We utilized standard machine learning algorithms (naive Bayes, support vector machines, and the C4.5 decision tree) trained on dbGaP study text and incorporated n-gram features and study metadata to identify heart, lung, and blood studies. We used the χ^2 feature selection algorithm to identify features that contributed most to classification performance and experimented with dbGaP associated PubMed papers as a proxy for topicality. Classifier performance was favorable in comparison to keyword-based search results. It was determined that text categorization is a useful complement to document retrieval techniques in the dbGaP.

Keywords: text classification, text categorization, database, genome-wide association studies, GWAS, natural language processing

Biomedical Informatics Insights 2013:6 35–45

doi: 10.4137/BII.S11987

This article is available from http://www.la-press.com.

© the author(s), publisher and licensee Libertas Academica Ltd.

This is an open access article published under the Creative Commons CC-BY-NC 3.0 license.

Introduction

In 2003, the National Institute of Health (NIH) required that certain funded projects include a data-sharing plan. Motivated by the idea of understanding phenotypic influence on genetic disease, these policies were later expanded to include NIH funded genome-wide associated studies (GWAS).[1,2] In order to facilitate implementation, a central data repository, the database of Genotypes and Phenotypes (dbGaP), was created by the National Center for Biotechnology Information (NCBI) to provide researchers access to the genotypic and phenotypic information. The database includes specific phenotype variables, statistical summaries of genetic information, and offers potential to access individual level data if approved by an NIH Data Access Committee. The database is growing at a rapid pace. In August 2011, 187 top-level studies (a study comprised of sub-studies) were archived in dbGaP, and by December 2012 there were 357 top-level studies.[3]

The existence of a publicly accessible database, however, does not guarantee information is available in a suitable form for efficient retrieval, study replication, identification of control populations, or new hypothesis generation.[4] Major contributing factors to suboptimal study reuse are that phenotypic variable names are not standardized and related concepts are not effectively mapped. Approximately 130,000 variable names exist in the database and many are redundant. For example, systolic blood pressure is represented by SBP, systolic_BP, and other variations such that a string match-based search may miss these synonyms. Alternatively, if the search term "white" is entered as a keyword search instead of "Caucasian," for example, the retrieved results for each search do not match. Genetic studies are expensive and time-consuming, hence maximizing data reuse is of paramount importance.[5]

We are approaching this problem by aligning phenotype variable descriptions to a standard information model in two ways. First, our group has developed PhenDisco—Phenotype Discoverer (http://pfindr.net/)—a robust tool for researchers to query and upload studies in a standardized fashion by retrofitting phenotypes in dbGaP with ontologies using natural language processing. Second, we are enhancing PhenDisco by integrating automatic document classification techniques into the system. This paper is focused on the second objective as we explore approaches to automatic document classification, motivated by the need to provide enhanced search functionality to the PhenDisco system. The exploding number of scientific publications in recent decades and an increasing number of databases makes automated document classification in biomedicine extremely important to provide accurate data retrieval, organize topics of interest for research, and streamline costs of data curation.

While there is a growing body of literature in the field of biomedical text classification, a search of PubMed and Google Scholar revealed no publications about text classification applied to dbGaP. We aim to (1) describe in detail the attributes of dbGaP studies, and (2) improve text categorization utilizing *n*-grams and metadata features. We focus on heart (cardiac), lung (pulmonary and/or respiratory), and blood (heme) studies. These categories were chosen due to their importance to the dbGaP host organization, the National Heart, Lung, and Blood Institute (NHLBI).

Methods

Three hundred and seventeen studies were available in dbGaP on July 1, 2012. Each title and abstract was manually reviewed and annotated by MKR and KWL into heart, lung, blood, and other categories. Inter-rater reliability was calculated using the R programming language (package IRR) implementation of Cohen's kappa (two raters) and Fleiss's kappa (three raters).[6] We confirmed the need for enhanced retrieval by performing a simple manual keyword search experiment. Search terms (asterisk = wildcard search) for heart studies were heart and/or card*. Lung study terms were lung and/or pulm* and/or resp*. The terms blood and/or heme* were entered for blood studies.

Three machine learning algorithms used successfully for text classification in the past were applied in this work: naïve Bayes (NB), support vector machines (SVM), and the C4.5 decision tree learning algorithm (Weka v. 3.6.8 using default parameters). The NB algorithm is frequently utilized for text classification with good results despite its assumption that features are independent. The SVM algorithm is commonly used because it is robust and resilient to over-fitting. Decision tree algorithms are another commonly applied tool, but incur the risk of mistakes early in the training process.[7]

Below we describe the experiments in detail. First, we explain the study metadata features. Second, we detail text classification experiments integrating *n*-grams and metadata features. Finally, we outline the χ^2 feature selection algorithm experiments. In addition to text classification based directly on dbGaP studies, we also experimented with using PubMed listed abstracts associated with each study as proxy representation of dbGaP documents. All classifiers were evaluated using Weka with stratified 10-fold cross validation. Evaluation metrics used were accuracy, precision, recall, and F-measure (Table 1). The F-measure is the harmonic mean between precision and recall. Results of the manual keyword search demonstrate the opportunity for improvement in accuracy, precision, recall, and F-measure (Table 2).

Metadata features

Each dbGaP study Web page is organized into descriptive sections (Fig. 1). The sections are: study description, authorized access, publicly available data, study inclusion/exclusion criteria, molecular data, study history, selected publications, disease related to study (MeSH terms), links to other NCBI resources, authorized data access requests, and study attribution (principal/co-investigators and funding source). We focused on journal publications, MeSH terms, principal/co-investigators, and funding source as these features have been shown to increase classification accuracy in previous studies.[8–10]

n-gram and metadata experiments

The first set of experiments used *n*-gram document representation derived from study titles and descriptions in addition to selected metadata. *n*-grams provide a means of facilitating statistical analysis by characterizing text in terms of n-consecutive sequences of tokens. For example, performed pulmonary function test separates into unigrams of performed, pulmonary, function, and test and bigrams of performed pulmonary, pulmonary function, and function test, which assists identification as a pulmonary test. Another test, such as cardiac function stress test, can be differentiated by unigrams of cardiac, function, stress, and test and bigrams cardiac function, function stress, and stress test.

Table 1. Contingency table and measurement definitions.

	Correct label	Incorrect label
Assigned label	a	b
Not assigned label	c	d

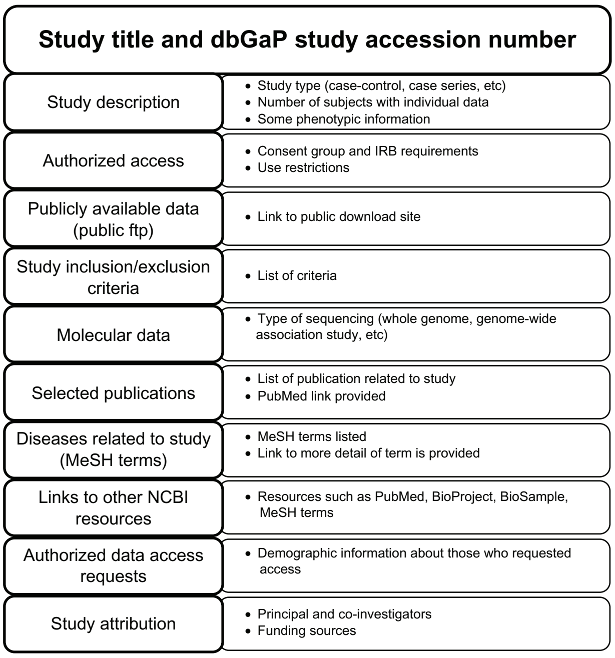
Accuracy = a + d/a + b + c + d
Precision = a/a + b
Recall = a/a + c
F-measure = 2 × precision × recall/precision + recall

Table 2. dbGaP keyword search results.

	Heart	Lung	Blood
Accuracy	0.64	0.66	0.41
Precision	0.31	0.14	0.07
Recall	0.72	0.75	0.76
F-measure	0.43	0.23	0.13

Notes: dbGaP manual keyword search results of heart, lung and blood studies based on gold-standard label assignment by investigators MKR and KWL.

Metadata for each study was automatically extracted from the dbGaP website (http://www.ncbi.nlm.nih.gov/gap) using Python scripts. For our first set of experiments we employed unigrams, bigrams, and the study metadata (journals, MeSH terms, principal and co-investigators, and funding source). Accuracy, precision, recall, and F-measure were calculated using Weka.



Study title and dbGaP study accession number

Study description	• Study type (case-control, case series, etc) • Number of subjects with individual data • Some phenotypic information
Authorized access	• Consent group and IRB requirements • Use restrictions
Publicly available data (public ftp)	• Link to public download site
Study inclusion/exclusion criteria	• List of criteria
Molecular data	• Type of sequencing (whole genome, genome-wide association study, etc)
Selected publications	• List of publication related to study • PubMed link provided
Diseases related to study (MeSH terms)	• MeSH terms listed • Link to more detail of term is provided
Links to other NCBI resources	• Resources such as PubMed, BioProject, BioSample, MeSH terms
Authorized data access requests	• Demographic information about those who requested access
Study attribution	• Principal and co-investigators • Funding sources

Figure 1. Metadata features of studies in dbGaP.
Note: Each dbGaP study is organized into descriptive sections.
Abbreviations: dbGaP, Database of Genotypes and Phenotypes; IRB, Institutional Review Board; ftp, File Transfer Protocol; MeSH, Medical Subject Headings; NCBI, National Center for Biotechnology Information.

Feature selection experiment

Our next experiment focused on feature set optimization. Yang and Pedersen have shown that feature selection (in particular the χ^2 and information gain feature selection algorithms) can improve classification accuracy for some text classification tasks.[11,12] We used the Weka implementation of the χ^2 feature selection algorithm in this work. All *n*-grams (unigrams/bigrams) and metadata features were combined and the feature selection threshold (optimal number of features for best performance) was determined experimentally by incrementing the number of features by 100 until the maximum number of features had been reached. We ran feature selection experiments using the NB and SVM algorithms. Accuracy, precision, recall, and F-measure were calculated for each additional feature combination and the optimal threshold was determined.

PubMed experiment

Due to the relative paucity of training data in dbGaP, our last experiment examined feasibility of using study-associated PubMed indexed articles as a representation for topicality (eg, if 20 PubMed heart studies are associated with a dbGaP study, but only one lung study, then the study is likely to be a heart study). A similar method was applied to classifying company and university websites with some success by Ghani et al.[13] To develop our training corpus, MKR manually chose 100 PubMed articles associated with dbGaP studies at random in each category of heart, lung, and blood and 300 PubMed studies unrelated to heart, lung, or blood topics for a total of 600 PubMed studies. A binary classifier was used for each category. For example, a study would be categorized as heart or other, lung or other, and blood or other. In each dbGaP study, the associated PubMed studies are found in the selected publications section under which the study authors, title, and journal article are recorded with a PubMed hyperlink. Of the chosen studies, approximately 60 PubMed studies directly associated with dbGaP studies in each category of heart, lung, and blood. However, not all dbGaP studies have an associated PubMed study; therefore, 40 studies with the same topicality were chosen at random from the PubMed database (with same search terms used in the keyword experiments), for a total of 100 studies per heart, lung, and blood category. MC and KWL categorized MKR's chosen studies and inter-rater reliability was calculated using Fleiss' Kappa for three raters. Each discrepant classification was assigned a label after discussion and majority vote. A corpus was created from the PubMed study title and abstract and categorized using NB and SVM classifiers with *n*-gram (unigram and bigram) based feature representation.

Results

We detail our results in four sections. First, we report on some salient characteristics of the dbGaP metadata. Second, we describe the *n*-gram and metadata classification experiments. Third, we present results from the feature selection experiment. Finally, we state the findings from our PubMed proxy article experiment.

Metadata features

In this section we report details of four types of metadata associated with dbGaP studies: journals, MeSH terms, principle and co-investigators, and funding sources.

Journals

There were 4707 journals publications linked to all dbGaP studies. Of these, 606 were unique instances. The mean number of articles per dbGaP study was 15.4 ± 101.81, ranging from 0–1514. The journals most frequently linked to dbGaP studies were of general topicality, cardiology, epidemiology, or stroke (Table 3).

Medical Subject Heading terms

On average, dbGaP studies are associated with 2.24 ± 5.46 Medical Subject Heading (MeSH) terms, ranging from 0–69. There were 771 total terms, with cardiovascular disease, stroke, obesity, and smoking being the most common topics. Terms with a frequency of at least six are represented in Table 4.

Principal and co-investigators

There were 903 principal investigators and co-principal investigators associated with studies in dbGaP. On average each investigator was associated with a mean of 3.99 ± 8.22 distinct studies. Fifty-one investigators were associated with three or more dbGaP studies.

Table 3. Most frequent journals linked to dbGaP studies.

Journal name	Frequency	Percentage (n = 4707)
American Journal of Epidemiology	269	5.7
Circulation	228	4.8
American Journal of Cardiology	146	3.1
JAMA: Journal of the American Medical Association	127	2.7
Atherosclerosis	125	2.7
Stroke	124	2.6
American Heart Journal	118	2.5
The New England Journal of Medicine	116	2.5
Nature Genetics	110	2.3
Diabetes Care	100	2.1
Annals of Epidemiology	91	1.9

Note: Cardiovascular-related subject matter was most common among journals.

Funding sources

The funding agency most frequently linked to dbGaP studies supported 111 studies. The mean number of studies supported per funding body was 3.44 ± 0.93 ranging from 1–111. Studies were predominately funded by the National Human Genome Research Institute (36%) and the National Heart, Lung, and Blood Institute (12%). The most frequent funding sources are listed in Table 5.

Table 4. Most frequent MeSH terms in dbGaP.

MeSH terms	Frequency	Percentage (n = 771)
Myocardial infarction	22	2.9
Cardiovascular diseases	20	2.6
Stroke	20	2.6
Obesity	15	2.0
Smoking	10	1.3
Diabetes mellitus, type 2	8	1.0
Prostatic neoplasms	7	0.9
Breast neoplasms	6	0.8
Cardiovascular system	6	0.8
Cholesterol	6	0.8
Dementia	6	0.8
Diabetic nephropathies	6	0.8
Heart disease	6	0.8
Intermittent claudication	6	0.8
Lung neoplasms	6	0.8
Osteoporosis	6	0.8
Parkinson disease	6	0.8
Risk factors	6	0.8

Notes: Medical Subject Heading (MeSH) terms utilized with frequency of least six in dbGaP. The most frequent terms are those relating to cardiovascular disease, obesity, and smoking.

n-gram and metadata experiments

Results from the *n*-gram (unigram and bigram) and metadata feature experiments were favorable. We constructed our training corpus from dbGaP study descriptions, titles and histories. With all corpus study descriptions combined, there were a total of 78,709 words with an average of 258.06 words ± 229.63 (range 0–1419) per study. The study histories contained a total of 24,706 words with a mean of 81 ± 186.97 (range 0–1822). We considered unigrams (one word units) and bigrams (two-word units) as potential features.

Inter-rater agreement

MKR and KWL manually classified dbGaP studies into the categories of heart, lung, and blood.

Table 5. Most frequent funding sources in dbGaP.

Funding source agency	Frequency	Percentage (n = 359)
National Human Genome Research Institute	111	36.4
National Heart, Lung, and Blood Institute	38	12.5
National Cancer Institute	21	6.9
National Institute on Aging	7	2.3
Carlos Slim Health Institute	7	2.3
Geisinger Clinic	4	1.3
National Center for Research Resources	4	1.3
National Institute of Diabetes and Digestive and Kidney Diseases	4	1.3
National Institute of Drug Abuse	4	1.3
National Institute of Neurological Disorders and Stroke	4	1.3
The Canadian Institutes for Health Research	4	1.3
Prostate Cancer Foundation	3	1.0
Wellcome Trust	3	1.0
National Institute of Allergy and Infectious Diseases	3	1.0
Amyotrophic Lateral Sclerosis Association	3	1.0

Notes: Of the most frequently encountered funding sources in dbGaP, the majority are from the National Human Genome Research Institute, the National Heart, Lung, and Blood Institute, and the National Cancer Institute.

Cohen's Kappa score was 0.86, 0.72, and 0.77, respectively indicating acceptable agreement. The discrepancies were reviewed, discussed and a final category was determined.

Heart studies

There were 46 heart studies in the database. The best performing algorithm was the C4.5 approach with accuracy of 92.5% and F-measure of 76.3. The second-best performing algorithm was SVM with 90.2% accuracy and F-measure of 65.9. The unigram feature yielded the best result regardless of the metadata feature combination (Table 6). This is substantial improvement over the keyword search accuracy of 64% and F-measure of 43.

Lung studies

There were 20 studies classified as lung in the dbGaP database. In this case, the C4.5 algorithm performed adequately with regards to accuracy, but was not the highest overall performing learning algorithm in terms of F-measure. The SVM classifier achieved the highest overall score of 95.1% for accuracy and 44.4 for F-measure when the metadata features funds, MeSH terms, and journals were combined (Table 7). Similar to heart studies, this was a noticeable improvement over keyword search for lung studies, which provided an accuracy of 66% and F-measure of 23.

Blood studies

There were 28 blood studies in the training set. The best performing classifier was SVM in conjunction with MeSH features, with 92.1% accuracy and 33.3 F-measure. The best performing combination of features was unigrams, funds, and journals with an accuracy of 92.1% and F-measure of 29.4 as shown in Table 8. The scores decreased as increasing features were added likely due to over-fitting of the model. While the results did not increase to the degree of the heart and lung cases, there was an improvement over keyword accuracy of 41% and F-measure of 13.

Feature selection experiment

Although the unigrams and metadata feature-based classifiers out-performed the keyword search available in dbGaP, we continued with the feature selection experiment to attempt further performance increase. The 20 most discriminating features in each category as determined by the χ^2 feature selection algorithm are noted in Table 9. Performance of the SVM and NB learning algorithms was improved by determining the optimal number of features in cross-validation experiments (Table 10). For heart studies, the threshold was 500 features with best performance by SVM algorithm (F-measure, 83.1). In lung studies the threshold was 2400 features with best results by NB algorithm (F-measure, 78.8). The blood study threshold was 2200 features with best performance by NB algorithm (F-measure, 72.7). All groups achieved substantial improvement, particularly over keyword search (Fig. 2).

PubMed experiment

We found PubMed studies effective metadata to represent topicality of dbGaP studies (Table 11).

Table 6. Results from *n*-gram and metadata experiments: heart studies.

Feature combination	Accuracy			F-measure		
	C4.5	SVM	NB	C4.5	SVM	NB
Unigrams	**92.5**	90.2	85.6	**76.3**	65.9	48.8
Bigrams	88.9	89.8	83.6	59.5	58.7	24.2
Funding Sources	83.3	83.2	79.7	0.0	10.5	3.1
Journals	83.6	82.0	82.6	16.7	20.3	34.6
MeSH	89.8	89.8	84.6	56.3	58.7	27.7
Principal Investigator	86.2	86.6	78.7	30.0	34.9	90.2
Unigrams_Funding Sources_Journals	**92.5**	89.8	85.6	**76.3**	47.6	65.2
Unigrams_Funding Sources	**92.5**	90.2	85.6	**76.3**	65.9	47.6
Unigrams_Journals	**92.5**	89.8	85.6	**76.3**	65.2	47.6
Unigrams_MeSH_Journals	**92.5**	89.8	85.6	**76.3**	65.2	47.6

Notes: Best-performing combinations of metadata (Journals, MeSH Terms, Principal and Co-Investigators, Funding Sources) and *n*-grams (Unigrams and Bigrams) featured in heart category. The underscore separates each feature. In addition to results highlighted in table, all other combinations of features involving unigrams with C4.5 algorithm provided good performance with accuracy of 92.5% and F-measure of 76.3. MeSH = Medical Subject Headings. The highest accuracies and F-measures are in bold typeset.

Table 7. Results from *n*-gram and metadata experiments: lung studies.

Features	Accuracy			F-measure		
	C4.5	SVM	NB	C4.5	SVM	NB
Unigrams	91.8	94.1	90.8	24.2	25.0	6.7
Bigrams	91.5	94.1	91.2	13.3	25.0	0.0
Funding Sources	93.4	93.4	90.5	0.0	0.0	0.0
Journals	93.8	93.8	88.5	17.4	17.4	14.6
MeSH	94.1	94.8	91.5	35.7	38.5	7.1
Principal and Co-Investigators	94.4	94.8	90.2	26.1	33.3	0.0
Funding Sources_Principal and Co-Investigators	94.4	94.8	94.8	33.3	26.1	33.3
Funding Sources_MeSH_Journals	93.4	94.8	91.5	33.3	23.1	7.1
Funding Sources_MeSH	94.1	**95.1**	91.8	35.7	**44.4**	7.4
MeSH_Principal and Co-Investigators	94.8	94.8	91.5	38.5	38.5	7.1

Notes: Best-performing combinations of metadata (Journals, MeSH Terms, Principal and Co-Investigators, Funding Sources) and *n*-gram (Unigrams and Bigrams) features in lung category. The underscore separates each feature. MeSH = Medical Subject Headings. The highest accuracies and F-measures are in bold typeset.

The SVM algorithm using unigrams achieved the highest F-measure for heart, lung, and blood studies (98.2, 97.1, and 95.7, respectively). Accuracy was also highest with the SVM algorithm and unigrams (97.0, 95.1, and 92.6, respectively).

Inter-rater agreement

Inter-rater agreement was calculated across three raters (MKR, KWL, and MC). Both MKR and KWL have a background in clinical medicine, while MC's background is in informatics. The Fleiss' Kappa for *m* raters was 0.97 with percent-agreement of 95.9.[6] Cohen's kappa statistic was applied for pairwise comparison. The Cohen's Kappa and percent-agreement between MKR/KWL, MKR/MC, and KWL/MC were 98.7 and 99, 95.7 and 96.7, and 94.9 and 96.1, respectively. The discrepancies were examined. Most were about classification of diagnosis that fall into multiple categories. For example, stroke and hypertension were classified by some raters as other because they were considered neurologic disease rather than cardiovascular disease.

Discussion

For all categories—heart, lung, and blood—we found that text classification methods improved document identification compared to keyword based approach. By utilizing *n*-grams and metadata, the highest F-measures achieved for heart, lung, and blood studies were 65.9, 44.4, and 33.3, respectively. With feature

Table 8. Results from *n*-gram and metadata experiments: blood studies.

Features	Accuracy			F-measure		
	C4.5	SVM	NB	C4.5	SVM	NB
Unigrams	89.5	91.8	87.21	33.3	24.2	4.9
Bigrams	89.2	91.2	88.2	26.7	12.9	0.0
Funding Sources	90.2	88.9	86.9	0.0	0.0	0.0
Journals	90.5	90.2	84.9	0.0	0.0	8.0
MeSH	91.2	**92.1**	89.8	12.9	**33.3**	16.2
Principal and Co-Investigators	90.8	91.5	85.3	0.0	23.5	0.0
Unigrams_Funding Sources_Journals	89.5	92.1	87.9	33.3	29.4	5.1
Unigrams_Funding Sources	89.5	91.8	87.5	33.3	24.3	5.0
Unigrams_Journals	89.5	91.8	87.9	33.3	24.2	5.1
Funding Sources_MeSH_Principal and Co-Investigators	91.2	92.1	87.9	33.3	9.8	12.9

Notes: Best-performing combinations of metadata (Journals, MeSH Terms, Principal and Co-Investigators, Funding Sources) and *n*-gram (Unigrams and Bigrams) features in blood category. The underscore separates each feature. MeSH = Medical Subject Headings. The highest accuracies and F-measures are in bold typeset.

Table 9. Most frequently encountered discriminating features.

Heart	Lung	Blood
coronary	lung	LEUKEMIA
infarction	pulmonary	leukemia
myocardial infarction	the_lung	leukemia_all
coronary_artery	of_smoking	acute
artery	of_lung	tumor
myocardial	pulmonary_disease	lymphoblastic_leukemia
heart_disease	lung_cancer	lymphoblastic
nhlbi	controlled clinical trials	acute_lymphoblastic
heart	esp	transformation
MYOCARDIAL INFARCTION	was_evaluated	altered_in
hypertension	cessation_and	patient_samples
cardiovascular	expiratory	ACUTE LYMPHOBLASTIC LEUKEMIA
institute_nhlbi	seattlego_the	ACUTE MYELOID LEUKEMIA
atherosclerosis	trial_with	number_analysis
STROKE	lung_function	leukemia_cll
Stroke	SMOKING CESSATION	CHRONIC LYMPHOCYTIC LEUKEMIA
risk_factors	smoking_cessation	platelet
atherosclerosis_risk	into_two	cll
in_communities	obstructive_pulmonary	landscape_of
of_atherosclerosis	bronchodilator	we_performed

Notes: Notable discriminating features in heart, lung, and blood categories. Words in capital letters are MeSH terms. Underscore indicates bigram.
Abbreviations: esp, Exome Sequencing Project; Seattlego_the, participating group in ESP; cll, Chronic Lymphocytic Leukemia; all, Acute lymphoblastic Leukemia; nhlbi, National Heart, Lung, and Blood Institute.

selection threshold implemented, the best F-measures were 83.1, 78.8, and 72.7, respectively.

Text classification of biomedical-related documents is an active area of research and our dbGaP database results are comparable to previous approaches in the literature. Donaldson et al[14] identified PubMed literature on the topic of protein-protein interactions using an SVM algorithm, gaining a classification accuracy of 90%.[14] Dobrokhotov et al[15] used probabilistic classification to classify and rank PubMed literature according to human genes of interest, achieving 59% precision and 69% recall.[15] Miotto et al[16] applied classification and regression trees (CART) and artificial neural networks (ANN) machine learning algorithms to PubMed abstracts to identify allergen cross-reactivity papers, finding that a bag-of-words document representation performed best overall.[16] In 2007, Wang et al[9] used a NB text classifier to demonstrate improved automated document classification in an immune epitope database. Features used were authors, journal, and MeSH headings with a sensitivity (precision) of 95% and specificity of 51.1%.[9] In 2008, Poulter et al[17] created an information retrieval system for Medline with a NB classifier incorporating MeSH terms and journals titles to retrieve articles of interest in specific domains. The average precision varied from 0.69 to 0.92 depending on the topic.[17] In 2009, Conway et al[18] demonstrated that a combination of *n*-gram and semantic features in conjunction with a NB classifier yielded the best classification results for disease outbreak reports.[18] Botsis et al[19] identified cases of anaphylaxis to influenza vaccine from a

Table 10. Optimal number of features.

	Optimal # of features	F-measure (feature selection)	F-measure (keyword search)	F-measure (*n*-grams and metadata)
Heart	500	83.1 (SVM)	43.0	65.9
Lung	2400	78.8 (NB)	23.0	44.4
Blood	2200	72.7 (NB)	13.0	33.3

Notes: Comparison of experiment results for heart, lung, and blood studies reported based on highest scoring learning algorithm and F-measure.
Abbreviations: SVM, Support Vector Machines; NB, Naïve Bayes.

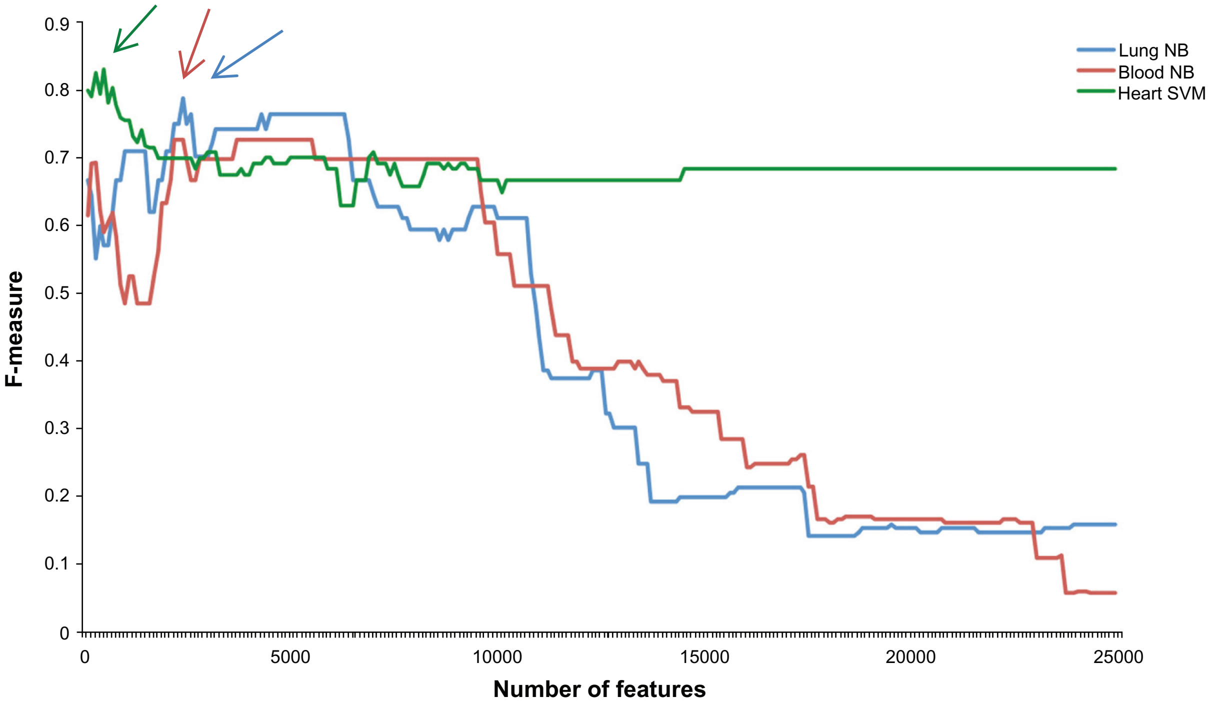


Figure 2. F-Measure feature selection thresholds: heart, lung, blood.
Notes: Best performing χ^2 feature selection algorithm results for heart, lung and blood studies. The arrow indicates the cutoff for best performance: Heart = 500 features, lung = 2400, and blood = 2400.
Abbreviations: NB, Naïve Bayes; SVM, Support Vector Machines.

database of adverse events using a rule-based classifier that incorporated keywords of anaphylaxis as part of the feature set with sensitivity of 79% and specificity of 94%. The standard machine learning classifiers resulted in F-measures ranging from 0.70 to 0.81.[19] In 2009, Denecke and Baehr[20] utilized additional document metadata such as keywords, titles, journal, and conference information to achieve favorable results for a document classifier.[20] More recently Wei and Collier[10] demonstrated that for the task of classifying full-text research papers according to model organism, a NB classifier using features derived from MeSH terms, journal and gene names, provided the best classification accuracy.[10]

In our experiments, the highest accuracy and F-measure overall for heart studies was based upon the unigram feature alone. This was attributed to the fact that the heart category contained the highest number of studies and the most homogenous diagnoses, such as coronary artery disease and myocardial infarction. For lung studies, the funding source was the most discriminating attribute. One possible explanation for this was the relatively small size of the lung document training set, and to explore this we reviewed the C4.5 classifier output. We discovered many terms used to determine decision nodes were not specific to lungs such as the name of the clinical trial, whereas the heart study output better captures cardiac medical terms

Table 11. PubMed metadata.

	Heart		Lung		Blood	
	Accuracy	F-measure	Accuracy	F-measure	Accuracy	F-measure
Unigrams NB	89.5	94.0	85.4	91.9	87.5	92.9
Bigrams NB	83.4	91.0	83.6	91.1	83.4	90.9
Unigrams SVM	**97.0**	**98.2**	**95.1**	**97.1**	**92.6**	**95.7**
Bigrams SVM	95.9	97.6	88.7	93.6	87.5	93.0

Notes: Support Vector Machine (SVM) algorithm performed best with unigram features to identify topicality of the database of Genotypes and Phenotypes (dbGaP) studies based on related PubMed studies. The highest accuracies and F-measures are in bold typeset.

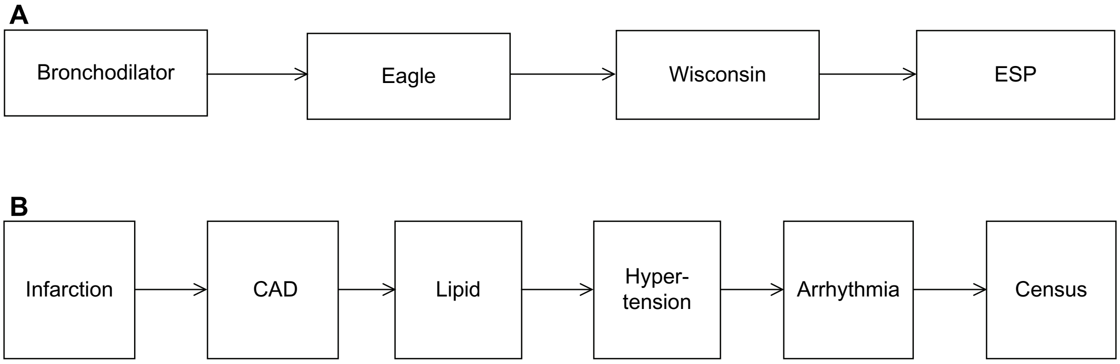


Figure 3. C4.5 decision tree nodes to determine heart vs. lung label.
Notes: The C4.5 decision tree nodes, represented by boxes, chosen for classification of (**A**) lung studies and (**B**) heart studies. The decision algorithm chooses a term at each node that best divides the training set into the desired category. For lung studies, the terms chosen to best identify the class are not exclusive to lung, rather they are names of study trials; whereas, the heart category terms are more generalizable. Therefore the decision tree algorithm performs the best for heart studies.
Abbreviations: Eagle, Environment and Genetics in Lung cancer Etiology study sponsored by the National Cancer Institute (NCI); ESP, Exome Sequencing Project sponsored by the National Heart, Lung, and Blood Institute; CAD, Coronary artery disease.

(Fig. 3). In regard to the blood studies, we speculate that the lower performance was because their topicality is comprised of heterogeneous terms including clotting disorders, leukemia, and platelets.

In regard to the inter-rater reliability of classifying the PubMed experiments, MKR and KWL's scores were more aligned because of their clinical backgrounds and experience with manual classification. The categorization of studies is not straightforward. There are diagnoses that may belong to multiple categories and require clinical expertise to determine appropriate classification. For example, although stroke involves the neurologic system, it is generally thought of as a cardiovascular disorder. Other indistinct diagnoses are pulmonary embolism and pulmonary hypertension. Although the root cause of embolism is typically a blood coagulation abnormality, and pulmonary hypertension is vascular in origin, these can be considered lung diseases as recognized by the American Lung Association. This concept of classifying studies into only one category when in reality diagnoses may fall under multiple categories, is a shortcoming of our methods. This is something to be addressed in the future as we optimize the algorithms and is important to consider when deciding which researchers will perform manual categorization of training corpus documents as domain knowledge has a decisive role.

Conclusion

Although relatively small, the number of studies in dbGaP is rapidly increasing. We demonstrated that using a document classifier based on *n*-grams and structured metadata yields better document retrieval results than the keyword-based search currently available in dbGaP. Without feature selection for heart studies, which had the largest amount of data in the training set, the C4.5 algorithm with unigrams had the best performance. For lung studies, the SVM classifier with funding sources and MeSH terms contributed the most to successful classification. The SVM algorithm was most effective when MeSH terms were utilized to identify blood studies, closely followed by the combination of unigrams, funding sources, and journals.

In future work, we plan to employ features derived from MetaMap[21] and other natural language processing tools in order to further improve classification accuracy and integrate this into the PhenDisco system. We plan to expand these experiments in topicality and build classifiers for other conditions of interest to dbGaP users (eg, asthma, chronic obstructive pulmonary disease, myocardial infarction, diabetes, etc.). The methods presented in this paper are not only suitable for dbGaP, but can also add structure to new databases or retrofit existing databases.

Acknowledgements

We thank Dr. Lucila Ohno-Machado for review of the manuscript. We thank Dr. Hyeon-eui Kim (University of California, San Diego), Dr. Wendy Chapman (University of California, San Diego), Dr. Son Doan

(University of California, San Diego), and Julianne Iacuaniello for helpful discussion.

Author Contributions

MKR conceived of the study, collected and analyzed data, and drafted the manuscript. KT and AK performed machine learning experiments and data analysis. KWL performed manual classification analysis. MC participated in study design and helped draft the manuscript. All authors read and approved the final manuscript.

Funding

The authors were funded in part by NIH grants UH2HL108785 and T15LM011271 (MKR).

Competing Interests

Author(s) disclose no potential conflicts of interest.

Disclosures and Ethics

As a requirement of publication the authors have provided signed confirmation of their compliance with ethical and legal obligations including but not limited to compliance with ICMJE authorship and competing interests guidelines, that the article is neither under consideration for publication nor published elsewhere, of their compliance with legal and ethical guidelines concerning human and animal research participants (if applicable), and that permission has been obtained for reproduction of any copyrighted material. This article was subject to blind, independent, expert peer review. The reviewers reported no competing interests.

References

1. Final NIH Statement on Sharing Research Data. Feb 26, 2003. http://grants.nih.gov/grants/guide/notice-files/not-od-03-032.html. Accessed Apr 23, 2013.
2. Implementation Guidance and Instructions for Applicants: Policy for Sharing of Data Obtained in NIH-Supported or Conducted Genome-Wide Association Studies (GWAS). Nov 16, 2007. http://grants.nih.gov/grants/guide/notice-files/not-od-08-013.html. Accessed Apr 23, 2013.
3. Mailman MD, Feolo M, Jin Y, et al. The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet*. 2007;39(10):1181–6.
4. Manconi A, Vargiu E, Armano G, Milanesi L. Literature retrieval and mining in bioinformatics: state of the art and challenges. *Adv Bioinformatics*. 2012;2012:573846.
5. Kraft P, Zeggini E, Ioannidis JP. Replication in genome-wide association studies. *Stat Sci*. 2009;24(4):561–73.
6. Artstein RaP, M. Inter-coder agreement for computational linguistics. *Computational Linguistics*. 2007.
7. Witten IH, Frank E, Hall MA. *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd ed. Burlington, MA: Morgan Kaufmann; 2011.
8. Trieschnigg D, Pezik P, Lee V, de Jong F, Kraaij W, Rebholz-Schuhmann D. MeSH Up: effective MeSH text classification for improved document retrieval. *Bioinformatics*. 2009;25(11):1412–8.
9. Wang P, Morgan AA, Zhang Q, Sette A, Peters B. Automating document classification for the Immune Epitope Database. *BMC Bioinformatics*. 2007;8:269.
10. Wei Q, Collier N. Towards classifying species in systems biology papers using text mining. *BMC Res Notes*. 2011;4(1):32.
11. Yang YaP, J. A Comparative Study on Feature Selection in Text Categorization. *Proceedings of ICML-97, 14th International Conference on Machine Learning*. 1997:412–20.
12. Lewis D. Representation and Learning in Inforamtion Retrieval. *PhD Thesis, Department of Computer Science, University of Massachusetts, Amherst, USA*. 1992.
13. Ghani RS, Slattery S, Yang Y. Hypertext categorization using hyperlink patterns and meta data. *Proceedings of the Eighteenth International Conference on Machine Learnning ICML*. 2001:178–85.
14. Donaldson I, Martin J, de Bruijn B, et al. PreBIND and Textomy—mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC Bioinformatics*. 2003;4:11.
15. Dobrokhotov PB, Goutte C, Veuthey AL, Gaussier E. Combining NLP and probabilistic categorisation for document and term selection for Swiss-Prot medical annotation. *Bioinformatics*. 2003;19 Suppl 1:i91–4.
16. Miotto O, Tan TW, Brusic V. Supporting the curation of biological databases with reusable text mining. *Genome Inform*. 2005;16(2):32–44.
17. Poulter GL, Rubin DL, Altman RB, Seoighe C. MScanner: a classifier for retrieving Medline citations. *BMC Bioinformatics*. 2008;9:108.
18. Conway M, Doan S, Kawazoe A, Collier N. Classifying disease outbreak reports using *n*-grams and semantic features. *Int J Med Inform*. 2009;78(12):e47–58.
19. Botsis T, Nguyen MD, Woo EJ, Markatou M, Ball R. Text mining for the Vaccine Adverse Event Reporting System: medical text classification using informative feature selection. *J Am Med Inform Assoc*. 2011;18(5):631–8.
20. Denecke KRT, Baehr T. Text Classification Based on Limited Bibliographic Metadata. *IEEE Xplore Digital Library*. 2009:1–6.
21. National Library of Medicine (NLM). UMLS Metathesaurus Fact Sheet. 2012. http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html. Accessed Apr 23, 2013.