

*Application of Information Technology* ■

## A Statistical Approach to Scanning the Biomedical Literature for Pharmacogenetics Knowledge

DANIEL L. RUBIN, MD, MS, CAROLINE F. THORN, PhD, TERI E. KLEIN, PhD, RUSS B. ALTMAN, MD, PhD

**Abstract** **Objective:** Biomedical databases summarize current scientific knowledge, but they generally require years of laborious curation effort to build, focusing on identifying pertinent literature and data in the voluminous biomedical literature. It is difficult to manually extract useful information embedded in the large volumes of literature, and automated intelligent text analysis tools are becoming increasingly essential to assist in these curation activities. The goal of the authors was to develop an automated method to identify articles in Medline citations that contain pharmacogenetics data pertaining to gene–drug relationships.

**Design:** The authors built and evaluated several candidate statistical models that characterize pharmacogenetics articles in terms of word usage and the profile of Medical Subject Headings (MeSH) used in those articles. The best-performing model was used to scan the entire Medline article database (11 million articles) to identify candidate pharmacogenetics articles.

**Results:** A sampling of the articles identified from scanning Medline was reviewed by a pharmacologist to assess the precision of the method. The authors' approach identified 4,892 pharmacogenetics articles in the literature with 92% precision. Their automated method took a fraction of the time to acquire these articles compared with the time expected to be taken to accumulate them manually. The authors have built a Web resource (<http://pharmdemo.stanford.edu/pharmdb/main.spy>) to provide access to their results.

**Conclusion:** A statistical classification approach can screen the primary literature to pharmacogenetics articles with high precision. Such methods may assist curators in acquiring pertinent literature in building biomedical databases.

■ *J Am Med Inform Assoc.* 2005;12:121–129. DOI 10.1197/jamia.M1640.

A challenge for biomedical researchers in the postgenomic era is to use prior knowledge about pharmacogenetics to understand the genetic basis for drug response and to predict drug response in individual patients. Pharmacogenetics research studies investigate how variations in particular genes alter the efficacy and toxicity of drugs. These studies contribute knowledge about which genes are involved in producing or altering a drug effect (mechanism of action or metabolism) and how genetic variations alter observable responses. There are more than 3 million citations in Medline mentioning the term *gene* or *drug*, and a resource that summarizes the gene–drug relationships that have been established and the

level of support for them in the literature would help researchers identify gaps in the current knowledge and plan new research studies. Other text sources such as drug patent literature or drug information databases could also provide useful data for this resource.

Many databases collect genetic and protein sequences, structures, expression, and phenotype data, but few contain information that would permit scientists to explore possible connections between genes and drugs. Pharmacogenetics data are just beginning to become available online, and most are in unstructured free-text format. Those few resources that contain pharmacogenetically relevant data focus on a subset of pharmacogenetics studies.<sup>1,2</sup> A comprehensive resource of information that relates all genes and drugs does not yet exist.

We have been building PharmGKB, a resource for pharmacogenetics.<sup>3,4</sup> One of the services PharmGKB provides is a curated repository of gene–drug relationships, annotated with literature support, contributed by curators and the scientific community.<sup>5</sup> While growth of our repository has been increasing progressively, the ability to create a comprehensive resource is limited by the enormity of the literature and the time it takes curators to identify relevant articles.

The medical literature has been the primary venue for distributing the results of pharmacogenetics studies. One possibility for building a pharmacogenetics resource is to populate it directly from the literature by screening all published articles for pertinent pharmacogenetics content. Some biomedical

---

Affiliations of the authors: Section of Medical Informatics, Stanford University, Stanford, CA (DLR, TEK, RBA); Department of Genetics, Stanford Medical Center, Stanford, CA (CFT, TEK, RBA).

This work is supported by grants from the National Institute of General Medical Sciences (NIGMS), Human Genome Research Institute (NHGRI), National Library of Medicine (NLM), and the NIH/NIGMS Pharmacogenetics Research Network and Database (U01GM61374).

The authors thank John Conroy for his expertise and assistance in accessing the PharmGKB database.

Correspondence and reprints: Daniel L. Rubin, MD, MS, Section of Medical Informatics, MSOB X-215, Stanford, CA 94305; e-mail: <[rubin@smi.stanford.edu](mailto:rubin@smi.stanford.edu)>.

Received for publication: 06/16/04; accepted for publication: 10/20/04.

database resources incorporate references from the scientific literature (Saccharomyces Genome Database\*, Flybase†, and PharmGKB‡). The references cited in these resources were acquired by curators performing the following activities: (1) literature searches, (2) reading all articles in relevant current journals, and (3) obtaining feedback from others who found omissions in the resource. Because of time constraints, it is not feasible to review all articles in the literature (more than 11 million articles in PubMed). Building a new resource, such as one to catalog gene–drug relationships in pharmacogenetics studies, using manual approaches is an expensive and time-consuming task requiring a large curation staff.

An automatic or even semi-automatic method that screens the literature to identify candidate articles of interest to curators would be very helpful for building a pharmacogenetics resource. Such a method would be useful not only for building new resources but also could be used by existing databases, helping to make them more complete, and could reduce the work of identifying new articles of interest.

A variety of methods has been developed to retrieve articles of interest (information retrieval) or to classify them automatically (text classification).<sup>6</sup> Information retrieval methods identify relevant documents based on keyword or concept searches. These approaches generally use indexing or vector space models<sup>7</sup> to identify relevant documents. Text classification includes a variety of methods that assign predetermined category labels to documents. Much of the effort related to processing the biomedical literature has focused on automating information extraction using curated lexicons or ontologies to identify particular phrases in text.<sup>6</sup> We are not aware of work to develop automated methods to screen the biomedical literature to identify pharmacogenetics articles.

Our hypothesis is that a computational approach can be used to screen the biomedical literature automatically, retrieving pharmacogenetics articles describing gene–drug relationships. Our objective was to develop a tool to scan the biomedical literature and find the pharmacogenetics articles of interest using statistical classification methods. After a subset of relevant articles is identified, this manageable number of articles could then be manually reviewed by curators. In addition, our goal was to develop a tool to extract the gene, drug, and the type of pharmacogenetics evidence supporting their relationship and populate a database of pharmacogenetics gene–drug relationships to augment our pharmacogenetics literature database and help pharmacogenetics researchers review the evidence relating various genes and drugs.

## Methods

### A Pharmacogenetics Corpus for Statistical Learning

We built a corpus of articles containing representative pharmacogenetics articles of the kind we would like to find while scanning all of the biomedical literature. These are articles that describe a gene and drug pharmacogenetics association,

such as pharmacokinetics, pharmacodynamics, or altered drug effect (the types of pharmacogenetics associations are difficult to categorize, but a network of 10 pharmacogenetics research groups in Rubin et al.<sup>5</sup> recently found a set of working categories of pharmacogenetics articles—“categories of pharmacogenetics evidence”). The corpus also contained articles that were not relevant to pharmacogenetics but that could be confused with articles that are relevant. This corpus was used for building and testing our candidate statistical models.

Thus, the corpus contained two categories of pharmacogenetics articles: “positive” and “negative” articles. A positive article contains information describing a gene and drug related to at least one of the categories of pharmacogenetics evidence used by PharmGKB to classify pharmacogenetics articles.<sup>5</sup> These categories are used by curators of PharmGKB to recognize pharmacogenetics articles and were not directly used in this study other than to define a positive article. For example, a positive article may describe a gene that alters the effect of a drug or alter its pharmacokinetics or pharmacodynamics. Another example is an article describing that drug metabolism depends on a particular gene. Negative articles are those that do not discuss genes and drugs, or that talk about drugs and genes individually but not in association with each other, or that talk about genes and drugs together but not related to each other by one of the categories of pharmacogenetics evidence. For example, a report describing the sequencing of a gene and incidentally mentioning a drug is a negative article.

We compiled a set of positive articles from the following sources: (1) all articles in the PharmGKB repository of submitted gene–drug relationships<sup>5</sup> ( $N = 417$ ), a resource containing articles that were manually curated; and (2) pharmacogenetics articles cited by recent reviews and PharmGKB<sup>3,8,9</sup> ( $N = 318$ ). After discarding duplicate articles and those containing no Medical Subject Headings (MeSH), 426 positive articles were identified.

A set of negative articles was compiled by downloading all articles cited in the Gene Ontology annotations for human genes<sup>10</sup> ( $N = 24,148$ ). We assumed that most of these articles are relevant to biological processes, molecular functions, and cellular components for genes, rather than pharmacogenetics of genes. This was not a perfect assumption, because seven of these articles were contained in the positive set. After removing duplicate articles and those having no MeSH annotations, the negative set comprised 9,722 articles.

A third set of articles (which we called a “background corpus”) was created to represent a background distribution of articles from Medline,<sup>11,12</sup> the database used for scanning the biomedical literature. The Medline database (consisting of 396 XML files in 2003) was downloaded. A total of 30,000 consecutive recent articles from the Medline database were selected by randomly choosing one of the recent XML distribution files. These articles made up the background set. They included both positive and negative articles but in a proportion that approximated the actual distribution across all of Medline.

### Article Preprocessing

For each article, the title, abstract, and MeSH terms were extracted. The title and abstract were combined and tokenized on space and punctuation. Stop words (such as “a” or “and”) were removed. Words were stemmed using the

\*<http://www.yeastgenome.org/>

†<http://flybase.bio.indiana.edu/>

‡<http://pharmgkb.org/>

Porter stemmer.<sup>13</sup> A vector representation of documents based on word content was created, in which each distinct word is an orthogonal dimension in the vector space.<sup>7</sup>

The MeSH terms associated with an article also were analyzed. There are three types of MeSH terms<sup>14</sup>: main headings, supplementary concepts, and qualifiers. MeSH terms assigned to articles contain three components (Figure 1): (1) a main subject heading or supplementary concept (always present); (2) zero, one, or several subheadings; and (3) major topic flags (can be used to tag the heading or subheading). MeSH terms were split into the component main heading and subheadings, and each was treated as a separate term. Major topic flags were discarded. For example, if an article had the MeSH term, "Methotrexate/\*administration & dosage/adverse effects," then it was assigned three terms: "Methotrexate," "administration & dosage," and "adverse effects." The number of combinations of headings, subheadings, and major topic flags is enormous, making the feature space very sparse. The MeSH terms were split in this manner to reduce the dimensionality and sparsity in the feature space. Dimensionality reduction is advantageous in machine learning and classification applications.<sup>15,16</sup>

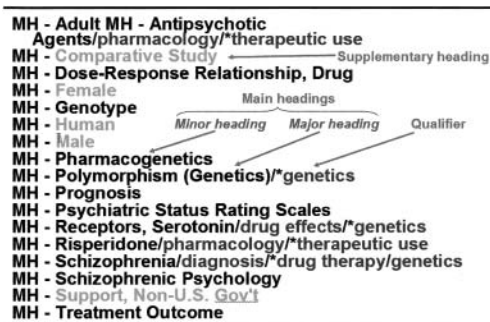
### Features Describing Pharmacogenetics Articles

For computational purposes, articles analyzed in this study were represented by a collection of individual numerical attributes (features). We studied two different types of features associated with articles: (1) patterns of word usage in the title and abstract and (2) patterns of MeSH terms assigned to articles. The scientific content of articles is carried in the vocabulary and term usage; thus, we expected that word usage in the citation text would be different between positive and negative articles.

MeSH term usage is a different feature of articles than word usage for detecting pharmacogenetics articles. Word usage is controlled by the diversity of authors of each scientific report. Every reference in Medline is indexed by a small number of librarians using several MeSH terms. MeSH is a controlled vocabulary used to index the medical literature,<sup>12,17</sup> and these annotations provide a multidimensional digest of the subject matter of the article. Like word usage, we expect MeSH term usage to be different for positive and negative articles.

We used the  $\chi^2$  test to select the most informative features (individual words and MeSH terms) that distinguish positive and negative articles. We calculated the  $\chi^2$  value by comparing the occurrence of positive and negative articles with the occurrence of the feature (Table 1). To calculate the  $\chi^2$  value, a cross tabulation was created for each feature, counting the number of positive and negative articles containing the feature and comparing these counts with a background distribution expected if the presence of the feature is independent from the type of article.

A very low  $\chi^2$  probability indicates that a feature and the type of article in which it occurs are not independent, though it does not suggest whether the presence or absence of that feature produces this dependence. To address this, we also calculated the positive predictive value (PPV) and negative predictive value (NPV) for each feature from the same cross tabulation. PPV is the percentage of articles containing features that are positive, and NPV is the percentage of articles lacking the fea-



**Figure 1.** Medical Subject Headings (MeSH terms) associated with an article. Some MeSH terms are designated "supplementary headings" and identified as such in separate MeSH documentation. Subheadings are additional terms that can be associated with MeSH headings, separated from the main heading by a slash. Major topics (main heading or subheading) are indicated with an asterisk.

**Table 1 ■ Chi Square Feature Selection**

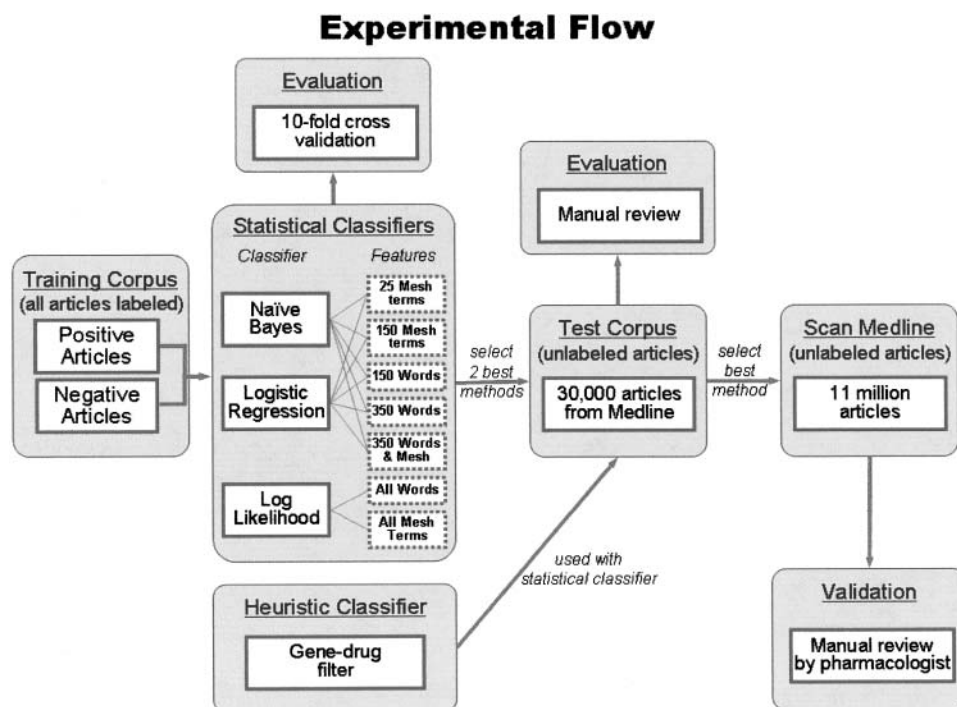
	Article Has Feature	Article Lacks Feature	Total
Article contains PG data	O = 5 E = 3.5	O = 2 E = 3.5	7
Article contains no PG data	O = 1 E = 2.5	O = 4 E = 2.5	5
Total	6	6	12

Chi-square calculation to determine if feature is predictive for pharmacogenetics articles. A total of 12 articles are counted in this example, and the observed (O) and expected (E) counts are compared to evaluate dependence of the feature on the type of article. A feature is considered to be a positive indicator if its positive predictive value (PPV) is greater than its negative predictive value (NPV). In this case, the feature is a positive indicator (PPV = 5/6 exceeds NPV = 4/6).

tures that are negative (Table 1). For each feature, if the PPV was less than the NPV, then the probability for that feature was given a negative value. Then all features were ranked by the probability value. Using this approach, the top-ranking features had significant  $\chi^2$  values and also had a larger PPV than NPV. Varying numbers of top-ranked features were used to build the statistical models described below.

### Classification Methods

We implemented three types of statistical models and a heuristic method (a "gene-drug filter" described below) to detect pharmacogenetics articles (Figure 2). The statistical models included Naïve Bayes,<sup>18</sup> logistic regression,<sup>19</sup> and a log-likelihood method. For Naïve Bayes and logistic regression, feature vectors were created based on (1) frequency of word usage, (2) frequency of MeSH term usage, and (3) frequency of words and MeSH terms. We trained the Naïve Bayes and logistic regression models using 5, 10, 25, 50, 100, 350, 600, 1,000, and 1,200 of the highest-ranked words or MeSH terms by  $\chi^2$ . We used the WEKA toolkit<sup>20</sup> to train the models, and they were evaluated using 10-fold cross validation. In this procedure, nine tenths of the corpus was used to build the statistical model, and it was tested with the remaining one tenth of articles. In this way, the model was tested with articles that did not participate in training the model. The procedure was repeated 10 times, each time using different



**Figure 2.** Diagram of the flow of experiments. A training corpus of labeled articles was used to evaluate different classification methods. The two best methods were subsequently tested on an unlabeled subset of Medline in conjunction with gene–drug filtering to select the best method to be used in scanning all of Medline. The results of applying this method to all of Medline were then validated manually.

subsets of articles for training and testing, and the results were averaged.

The precision, recall, and F measure were calculated for each statistical method. Precision was calculated as the percentage of classifier-positive articles that were positive articles, and recall was calculated as the percentage of positive articles that were classified as positive articles. The F measure combines precision and recall into a single measure:

$$F \text{ measure} = \frac{2(\text{Recall})(\text{Precision})}{\text{Recall} + \text{Precision}}$$

Our third statistical method was a log-likelihood calculation that was performed using (1) informative MeSH terms alone or (2) informative words alone (Figure 2). When using the method with MeSH terms, we discarded all MeSH supplementary concepts (a separate list of terms so designated by NLM that includes general concepts such as “Support, U.S. Gov’t,” and “Case Report”) that we had found from other experiments tend to add more noise than signal. We calculated a probability distribution of MeSH term usage in the positive corpus and the background corpus. Corpus-specific probabilities of MeSH terms occurring in these corpora were calculated separately as follows:

$$P(\text{MeSH term} | \text{Corpus}_i) = \frac{\text{Count}(\text{articles assigned to MeSH term})}{\text{Total number of articles in Corpus}_i}$$

where  $\text{Corpus}_i$  is the positive or background corpus.

For those MeSH terms that do not occur in a corpus, the probability distribution was smoothed by assigning that MeSH

term probability a very small value ( $10^{-8}$ ). A log-likelihood score for each MeSH term  $j$  was then calculated, which compares the distributions for positive and background corpora:

$$S_j = \ln \frac{P(\text{MeSH term}_j | \text{Corpus}_P)}{P(\text{MeSH term}_j | \text{Corpus}_B)}$$

where  $S_j$  is the score of MeSH term  $j$ ,  $\text{Corpus}_P$  and  $\text{Corpus}_B$  are the positive and background corpora, respectively. For each article in the test corpus (both positive and negative articles), a total log likelihood score was calculated:

$$S_{\text{article}} = \sum_{k=1}^T S_k$$

where  $S_k$  is the individual score for the  $k$ th MeSH term and  $T$  is the total number of MeSH terms.

A threshold value for an article’s total log-likelihood score of 10 was empirically chosen after test experiments were conducted with the PG corpus to determine its behavior. This threshold represented the best empirical balance between recall and precision for characterizing articles as being positive or negative. This method was tested on the pharmacogenetics corpus.

We also applied this log-likelihood method using the word content in the titles and abstracts of articles (Figure 2) and calculated an article’s score using similar equations to those above, substituting word usage for MeSH usage. The threshold value for the total log likelihood score was selected to produce the same recall with the method as when it was applied using MeSH terms as features to directly compare precision of the method using words and MeSH terms.

Another classification method we developed is a heuristic approach, called a “gene–drug filter,” intended to enhance the precision of some of the methods above. This filter selected articles containing one or more gene and drug name co-occurrences. To implement this filter, the title and abstract of articles classified as positive articles were parsed into sentences (the title was considered a sentence). For each sentence, we searched for the co-occurrence of a gene (or gene product) name and a drug name.

We used the GAPSCORE algorithm<sup>21</sup> with a p-value cutoff of 0.9 to identify gene and gene product names in text. GAPSCORE identifies gene names in text (including names that span several words), scoring words using a statistical model of gene names based on their appearance, morphology, and context. The algorithm produces a probability for each gene name identified; we chose the cutoff value empirically, representing a 90% probability that each name found is a gene name. To identify drug names, we looked for a lexical match of single text tokens against a drug name lexicon. This method finds only one-word drug names. The drug lexicon was assembled from a database of drug names and ingredients in PharmGKB,<sup>4,22</sup> most of which derive from the Veterans Administration National Drug File.<sup>23</sup> The terms “agent(s)” and “drug(s)” were added to the drug lexicon to capture generic unnamed pharmaceutical substances in articles.

### Scanning Medline

We compared the performance of the two best-performing methods (the logistic regression and log-likelihood methods) combined with the gene–drug filter for scanning a portion of the Medline database (the background corpus of 30,000 articles; Figure 2). A subset of the abstracts from candidate articles for each classification approach was reviewed manually to estimate how well the classifiers were performing the identification task. Subsequently, the best-performing classification approach was selected to scan all articles in the 2003 release of the Medline database (11,818,833 articles; Figure 2).

We also evaluated the performance of the gene–drug filter alone (the baseline classifier). We scanned the background corpus subset of Medline using only gene–drug co-occurrence as a selection criterion. This served as a check that the filtering step would not be sufficient when used alone, without the benefit of a classifier as a preselection step.

### Manual Validation

To validate the precision of the method, one third of the articles identified by the scanning method (1,649 abstracts) were reviewed by a pharmacologist to determine the proportion of the articles that were true-positive articles. The pharmacologist used the following criteria to determine whether an article was a positive article: (1) the title and/or abstract discuss one or more genes and drugs and (2) at least one gene and drug are described as being related in a way that is relevant to one or more of the categories of pharmacogenetics evidence (e.g., the gene alters the metabolism or action of drug).

## Results

### Performance of Models

Our pharmacogenetics training corpus contained 10,148 articles (426 positive articles and 9,722 negative articles). The

percentage of pharmacogenetics articles in this training corpus was 4.2%. There were seven articles (0.07%) that contained no MeSH terms.

Chi square feature selection identified many informative words and MeSH terms, the presence of which suggested a pharmacogenetics article. The highest-scoring words and MeSH terms were intuitively reasonable as predictors of pharmacogenetics articles (Table 2).

Performance of the statistical models evaluated on the pharmacogenetics training corpus is summarized in Table 3. Precision in the pharmacogenetics training corpus ranged from 0.11 to 1.0, whereas recall was between 0.19 and 0.97. The F measure (combines recall and precision in a single efficiency measure) ranged from 0.20 to 0.88, with the log-likelihood method performing best overall. The logistic regression models generally performed better than the Naïve Bayes models for identifying pharmacogenetics articles, particularly in terms of superior recall.

The logistic regression model with 150 MeSH terms as features was considered the best model-based classifier for our task because it had a very high precision and adequate recall (Table 3). While logistic regression models had slightly higher F measure with 350 best word features (with and without

Table 2 ■ Highest-Ranking MeSH Terms and Words for Identifying Pharmacogenetics Articles

Words/Mesh Terms Ranked by $\chi^2$	MeSH Terms Ranked by Log-Likelihood Score
2b6	<b>Histamine N-methyltransferase</b>
2c19	<b>Sulfaphenazole</b>
2c9	<b>Purine-pyrimidine metabolism, inborn errors</b>
2d6	<b>Proto-oncogene proteins c-rel</b>
20	<b>Glucuronic acids</b>
<b>6-Mercaptopurine</b>	<b>p-Aminosalicylic acid</b>
6mp	<b>Thioinosine</b>
<b>Administration, oral</b>	<b>Reovirus 3</b>
<b>Area under curve</b>	<b>Protein subunits</b>
<b>Cytochrome P-450 CYP2D6</b>	<b>Orphenadrine</b>
<b>Debrisoquin</b>	<b>Organic cation transport proteins</b>
<b>Double-blind method</b>	<b>Organic anion transport polypeptide C</b>
<b>Hydroxylation</b>	<b>Histamine N-methyltransferase</b>
<b>Metabolic clearance rate</b>	<b>Sulfaphenazole</b>
<b>Omeprazole</b>	<b>Purine-pyrimidine metabolism, inborn errors</b>
<b>Pharmacogenetics</b>	<b>Proto-oncogene proteins c-rel</b>
<b>Treatment outcome</b>	<b>Nordazepam</b>
2b6	<b>Hexobarbital</b>
2c19	<b>Hepatocytes</b>
2c9	<b>Equilibrative nucleoside transporter 1</b>
2d6	<b>Histamine N-methyltransferase</b>
a118g	<b>Sulfaphenazole</b>
Acei	<b>Purine-pyrimidine metabolism, inborn errors</b>
Albuterol	<b>Proto-oncogene proteins c-rel</b>

Top-ranked word (regular type) and MeSH term (bold type) features, by  $\chi^2$  score (left column; all p values <0.001) and by log-likelihood score (right column; scores range from 13.1 to 14.2), calculated such that term presence is an indicator of a pharmacogenetics article.

Table 3 ■ Classification Methods

Method	Training Feature	Precision	Recall	F measure
Naïve Bayes	25 best MeSH terms	1.0	0.2	0.33
Logistic regression	25 best MeSH terms	1.0	0.43	0.60
Naïve Bayes	150 best words	0.98	0.53	0.69
Logistic regression	150 best words	0.95	0.66	0.78
Naïve Bayes	350 best words	0.98	0.53	0.69
Logistic regression	350 best words	0.96	0.75	0.84
Naïve Bayes	150 best MeSH terms	1.0	0.19	0.32
<b>Logistic regression</b>	<b>150 best MeSH terms</b>	<b>1.0</b>	<b>0.7</b>	<b>0.82</b>
Naïve Bayes	350 best words and MeSH terms	0.99	0.46	0.63
Logistic regression	350 best words and MeSH terms	0.96	0.79	0.87
Log likelihood	All words	0.11	0.97	0.20
<b>Log likelihood</b>	<b>All MeSH terms</b>	<b>0.80</b>	<b>0.97</b>	<b>0.88</b>
<i>Log likelihood</i>	<i>All MeSH terms with gene-drug filtering</i>	<i>0.89</i>	<i>0.72</i>	<i>0.80</i>

Performance of different statistical models evaluated in this study built to identify pharmacogenetics articles. Results were obtained on the pharmacogenetics training corpus used in this study (methods in bold are the models tested in scanning the Medline subset; the method in bold italic was used to scan all Medline entries).

MeSH terms), precision was less; precision is somewhat more important than recall in a screening task, so the 150 MeSH term model was preferred.

The log-likelihood model using MeSH terms as features also performed well with high recall and precision, and it had the highest F measure overall. When this method was combined with gene–drug filtering, recall decreased to a value similar to the best performing logistic regression model, but precision increased to nearly 0.9 (Table 3). When the log-likelihood method was used with article words instead of MeSH terms, precision was very low at the same level of recall. The logistic regression model and the MeSH log-likelihood models were subsequently used to scan the Medline database.

### Scanning Medline

The logistic regression and the MeSH log-likelihood methods were tested initially by scanning the background corpus (30,000 sequential articles in Medline) with the gene–drug filter (Figure 2). The logistic regression model identified 1,889 (6.3% of the articles scanned), but nearly one fourth were false-positive articles, whereas the MeSH log-likelihood method identified 24 of 30,000 articles (0.08% of the articles scanned), all of which were true-positives. When the gene–drug filter was used in isolation of any classifier for scanning the background corpus for pharmacogenetics articles, 3,708 of 30,000 (12.4% of the article scanned) articles were identified, and more than half were false-positives; thus, the filter alone was not an effective method for screening for relevant articles. Thus, the MeSH log-likelihood method with the gene–drug filter was selected for scanning the entire Medline database.

When applied to the Medline database without the gene–drug filter, the MeSH log-likelihood method found 11,922 of 11,818,833 articles (0.1%), a similar identification rate to that seen with the corpus of 30,000 background articles. Review of a sample of these articles revealed frequent false-positive articles. When the MeSH log-likelihood method was combined with the gene–drug filter to select articles describing

particular drugs and genes, there were 4,892 articles identified (0.04%). It took approximately 24 hours to scan all of Medline.

One third (1,649) of the 4,892 articles identified by the MeSH log-likelihood method were reviewed by a pharmacologist. A total of 1,513 of these articles (92% precision) were pharmacogenetically relevant (true-positive articles). The 92% precision ascertained by the pharmacologist for scanning Medline is similar to the 89% precision measured with this method using the pharmacogenetics training corpus (Table 3).

### A Resource for Pharmacogenetics

We created a database containing the 4,892 pharmacogenetically relevant articles identified by our scanning method as well as the score assigned to those articles. The database also contains 25,120 gene–drug co-occurrences detected in those articles. The database can be queried by gene or drug (Figure 3). It also shows the literature supporting gene–drug relationships, which helps researchers differentiate gene–drug associations that are real from those that are likely to be coincidental.

### Discussion

The amount of published pharmacogenetics data is growing at an explosive pace. A database resource summarizing current pharmacogenetics knowledge relating genes and drugs is needed to help scientists find gaps in current knowledge and formulate new hypotheses. Most important pharmacogenetics studies are ultimately published in the biomedical literature, so a pharmacogenetics database should be built from that literature. The problem is that Medline contains over 11 million articles, and identifying a sizeable number of relevant pharmacogenetics articles is a challenge. While domain experts may be able to suggest many relevant articles in the field, these are generally the best known articles, and there will likely be articles that are missed because of the complexity of the biomedical literature.

THE 25 MOST FREQUENT GENE-DRUG RELATIONSHIPS IN PUBMED		
GENE	DRUG	SUPPORTING ARTICLES
cytochrome P450	xenobiotics	468
CYP2D6	xenobiotics	347
CYP3A4	xenobiotics	232
CYP2D6	debrisoquine	183
CYP2C19	xenobiotics	167
CYP2C9	xenobiotics	163
CYP2C19	mephenytoin	157
CYP1A2	xenobiotics	150
CYP3A	xenobiotics	145
CYP2D6	dextromethorphan	143
angiotensin	xenobiotics	141
CYP2D6	quinidine	131
cytochrome P450	debrisoquine	114
cytochrome P450	phenobarbital	105
CYP3A4	ketoconazole	101

**Figure 3.** Web resource storing gene-drug relationships and pharmacogenetics literature supporting these relationships. The number of articles supporting each relationship can be a useful indicator of the strength of the association.

We have developed and evaluated statistical methods for screening the literature to identify pharmacogenetics articles pertaining to gene-drug relationships. An important first step is to find features in articles that are useful for the discrimination task. We found that MeSH headings are useful for identifying articles that contain pharmacogenetic data. For the log likelihood method, precision is much higher when using MeSH terms than when using just words at the same level of recall (precision of 0.8 vs. 0.11). MeSH terms may have produced higher precision because the variety of word usage and the dimensionality of the feature space are much smaller for MeSH than for words. In addition, MeSH are manually assigned to index the content of articles, so each term may convey more specific information about an article's content than a particular word in the article. The value of MeSH terms is less apparent for the naïve Bayes and logistic regression classifiers; classifiers using the 350 best words had little improvement when MeSH terms were added (Table 3). The value of the feature thus appears to depend on the classifier in which it is used; in our application, MeSH terms with the log-likelihood method appear to be more useful than words alone for identifying pharmacogenetics articles.

MeSH headings are index terms manually assigned to articles by the National Library of Medicine, with an average of 10 MeSH terms applied to each Medline citation by professional indexers based on reading the full text of the article. Thus, MeSH terms serve as a digest of the content in the article, and they appear to be more informative features for identifying pharmacogenetics articles than simply using individual words that they contain. Because classification performance using MeSH terms was good, the profile of MeSH term usage appears to be distinctive for these articles, while being different for nonpharmacogenetics articles.

There are potential disadvantages to using MeSH. Because MeSH terms are not applied to all articles, only those that have been indexed with these terms can be recognized by a classifier that uses them as features. In our experiments with the training corpus, we found only 7 of 10,148 (0.07%) articles that contained no MeSH terms, so it is unlikely that

absence of MeSH terms is a large factor. A more significant limitation of MeSH is that it is only applicable to Medline articles, because these terms are only used for indexing the biomedical literature. Thus, MeSH features could not be used for screening other text resources, such as patent databases or drug inserts. For such resources, classifiers based on word features would need to be used, and Table 3 suggests the performance that might be expected.

In using MeSH terms as features of text, we chose to separate the headings and subheadings and treat them as independent features (Figure 1). We chose this approach because the number of potential combinations of headings and subheadings is very large and would result in a large feature space with sparsity of representative features unless the training corpus is huge.<sup>15</sup> A potential disadvantage of this approach is that semantic information conveyed by the attachment of subheadings to headings would be lost, which could reduce the informative value of the feature. We found that classifier performance was best using MeSH features; thus, even better results may be obtained using combined headings and subheadings. This would be interesting to pursue in future work.

Classifier performance was better for logistic regression and the MeSH log-likelihood method than Naïve Bayes. The former two methods are discriminative classifiers, whereas the latter is a generative classifier. Generative classifiers learn a model of the joint probability of the class and the features and then predict the class given the features. Discriminative classifiers model the probability of the class given the features directly, and it has been argued that they are preferred.<sup>24</sup> This difference in classifier performance is consistent with our results.

The ability of our classification methods to generalize to new, previously unseen, articles differed among our classifiers. When we applied the two classifiers that performed best on the training corpus to a large subset of Medline (the 30,000-article background corpus), the MeSH log-likelihood method had fewer false-positives than logistic regression, even though they had similar performance in studies with the training corpus (Table 3). There are several possible reasons

for this result. First, the training corpus was likely not representative of the mix of articles in Medline as a whole. The frequency of positive articles in the training corpus (4.2%) was likely larger than in Medline. In addition, the number of positive articles was probably not completely representative of the spectrum of pharmacogenetics articles occurring in Medline. Our results demonstrate the importance of classifier generalizability as a critical factor in making it successful for scanning Medline.

While the different classifiers we evaluated with the training corpus had different performance (Table 3), small differences in performance of some of the classifiers are probably not statistically significant. For example, the performance of the logistic classifier with 150 MeSH terms and the logistic classifier with 350 words and MeSH terms may be the same. We chose to use the former classifier because it was a simpler model and it tended toward higher specificity.

We needed to use a gene–drug filter to reduce the number of false-positives when applying our classifier to Medline. We wanted to assess precision of our final result, and reducing the result set from 11,922 to 4,892 made this manual task more manageable because it effectively increased the stringency of our test for positive articles.

Adding the gene–drug filter likely reduced recall, and a limitation of our study is that we were not able to assess the impact on recall using the classifier with this filter. It would be difficult to assess the degree to which recall was reduced without manually reviewing all of the articles in Medline. However, for the task of identifying literature to be reviewed by curators, precision is more important than recall; curators can only review a certain number of articles, and as long as there is an excess of articles to review, it is preferable to keep the number of false-positives to a minimum. In fact, the need to streamline literature curation in the face of an expanding literature was the focus of a recent competition.<sup>25</sup> We were able to estimate the overall precision of our method by manually reviewing one third of the articles identified by our approach. Our precision of 92% in all of Medline is similar to the precision obtained using this method in the training corpus (89%; Table 3).

An advantage of our MeSH log-likelihood method is that it only needs to be run once to score all articles in Medline. The choice of score threshold for declaring an article “positive” is a postprocessing step, and it can be changed easily. By selecting different thresholds for an article’s score, the precision/recall tradeoff can be adjusted. Thus, if comprehensive retrieval of pharmacogenetics articles is desired, articles with lower scores could be identified, with the cost that a greater fraction of false-positives would be retrieved as well. We currently display all articles that exceed one particular score threshold (Figure 3); it would be possible to allow the user to adjust this threshold and obtain higher recall or precision according to their needs.

An additional limitation of our study is that it is possible that the performance of the classifiers we evaluated in the training corpus may be different in the task of scanning Medline. Consistency of such results depends on how well the classifiers generalize. At least in the case of our log-likelihood classifier, we found that the precision in the training corpus is similar to that obtained in Medline. However, we cannot be certain that one of the other classifiers in Table 3 would not

have performed better, and MeSH log-likelihood classifier may not be optimal for the task of identifying pharmacogenetics articles. However, the results with our current method are a promising starting point for this task.

Our method makes it possible to process and analyze the entire Medline database in a short time, and it certainly can be applied periodically to discover new references. This can contribute to and enhances a curator’s ability to discover relevant literature. Methods such as we describe can help focus literature curation efforts by narrowing the number of articles that need to be reviewed, and they can stimulate the growth of new and existing biomedical resources.

By combining gene–drug co-occurrence with MeSH profiles indicative of pharmacogenetics, we were able to create a database of pharmacogenetics articles pertaining to gene–drug relationships. This database can be perused using a Web interface (Figure 3), or its content can be mined using computational methods to summarize gene–drug relationships.<sup>26</sup> Although we did not formally evaluate the gene–drug extraction component, the accuracy of the GAPSCORE algorithm has been evaluated previously,<sup>21</sup> and this algorithm will identify multiword gene names. We have not yet evaluated the accuracy of our approach to drug name identification; however, because drug name identification is a part of our overall method of article identification, its performance is reflected in the overall precision result we assessed. Drug name identification was based on matching against a lexicon of drug names; thus, precision would be expected to be high, and recall would depend on the size of the lexicon.

Simple co-occurrence does not always imply a true gene–drug association; however, as the number of articles containing a particular co-occurrence increases, a true association becomes more likely. One of the benefits of our Web resource is that it allows researchers to view the amount of literature supporting gene–drug associations and view the specific articles to enable them to differentiate true and spurious associations (Figure 3). Since our method is automated, we can periodically rescan Medline to update our resources with current biomedical data.

There has been much prior work in applying statistical methods to analyzing the biomedical literature to extract biomedical data,<sup>6,27</sup> but relatively little prior work has addressed the task of screening the entire literature for particular types of articles, such as pharmacogenetics articles. The KDD Challenge Cup 2002 included a task to identify papers appropriate for scientific curation of *Drosophila* gene expression information.<sup>25</sup> This competition was an attempt to use automated techniques to identify relevant articles from a large archive. Differences between this task and our work are (1) the competition’s test task was on 213 reports, whereas we screened all 11 million citations in Medline and (2) the full text of articles was used in the competition, whereas our method was applied only to the title and abstract of articles. Several methods were evaluated in this competition, and the F measure ranged from 32% to 78%. The best performing method used an information extraction approach with manually constructed rules that matched predetermined patterns of interest.<sup>28</sup> Another method that performed well used keywords with a Naïve Bayes classifier.<sup>29</sup> By comparison, the F measure of our log-likelihood method was 0.88 (Table 3).

The Genomics track of the TREC 2003 competition contained a task related to finding Medline references focusing on the basic biology of a given gene or its protein products. In this competition, MeSH features were found to be beneficial.<sup>30</sup>

We are using the results of this work to enhance the curation activities of PharmGKB (<http://pharmgkb.org>), a resource for the pharmacogenetics community.<sup>3,4,22</sup> One of the services PharmGKB provides is a database of curated pharmacogenetics literature citations associated with the categories pharmacogenetics knowledge, genes, drugs, and phenotypes.<sup>5</sup> Creating the initial entries in this database was a time-consuming effort. The articles identified by our classifier have provided a 100-fold increase in literature content to be added to our resources.

#### References ■

- Zhang MQ. Statistical features of human exons and their flanking regions. *Hum Mol Genet*. 1998;7(5):919–32.
- Flockhart DA. Drug Interaction Database. Available at: <http://medicine.iupui.edu/flockhart/>. Accessed January 4, 2005.
- Rubin DL, Woon M, Carillo M, et al. PharmGKB: A Resource to Link Genotype and Phenotype in Pharmacogenetics. In: *Medinfo*, 2004. San Francisco, CA; 2004.
- Klein TE, Chang JT, Cho MK, et al. Integrating genotype and phenotype information: an overview of the PharmGKB project. *Pharmacogenetics Research Network and Knowledge Base. Pharmacogenomics J*. 2001;1(3):167–70.
- Rubin DL, Carillo M, Woon M, Conroy J, Klein TE, Altman RB. A Resource to Acquire and Summarize Pharmacogenetics Knowledge in the Literature. In: *Medinfo*, 2004; San Francisco, CA; 2004.
- Shatkay H, Feldman R. Mining the biomedical literature in the genomic era: an overview. *J Comput Biol*. 2003;10(6):821–55.
- Salton G, Wong A, Yang CS. Vector-space model for automatic indexing. *Communications of the Acm*. 1975;18(11):613–20.
- Evans WE, Johnson JA. Pharmacogenomics: the inherited basis for interindividual differences in drug response. *Annu Rev Genomics Hum Genet*. 2001;2:9–39.
- Evans WE, Relling MV. Pharmacogenomics: translating functional genomics into rational therapeutics. *Science*. 1999; 286(5439):487–91.
- GO Annotations at EBI (Human). Available at: <http://www.geneontology.org/GO.current.annotations.shtml>. Accessed January 4, 2005.
- National Library of Medicine: The Medline Database. Available at: <http://www.nlm.nih.gov/pubs/factsheets/medline.html>. Accessed January 4, 2005.
- Bachrach CA, Charen T. Selection of MEDLINE contents, the development of its thesaurus, and the indexing process. *Med Inform (Lond)*. 1978;3(3):237–54.
- Porter MF. An algorithm for suffix stripping. *Program*. 1980; 14(3):130–7.
- National Library of Medicine: Medical Subject Headings. Available at: <http://www.nlm.nih.gov/mesh/meshhome.html>. Accessed January 4, 2005.
- Geman S, Bienenstock E, Doursat R. Neural networks and the bias variance dilemma. *Neural Computation*. 1992;4(1): 1–58.
- Yang Y, Pedersen J. A comparative study on feature selection in text categorization. In: Fisher DH (ed). *Proceedings of ICML-97*, 14th International Conference on Machine Learning. San Francisco, CA: Morgan Kaufmann, 1997, pp 412–20.
- Coletti MH, Bleich HL. Medical subject headings used to search the biomedical literature. *J Am Med Inform Assoc*. 2001;8(4): 317–23.
- John GH, Langley P. Estimating continuous distributions in bayesian classifiers. In: *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*. San Francisco, CA: Morgan Kaufmann, 1995, pp 338–45.
- Lecessie S, Vanhouwelingen JC. Ridge estimators in logistic-regression. *Applied Statistics—Journal of the Royal Statistical Society Series C*. 1992;41(1):191–201.
- Witten IH, Frank E. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. San Francisco, CA: Morgan Kaufmann; 2000.
- Chang JT, Schutze H, Altman RB. GAPSCORE: finding gene and protein names one word at a time. *Bioinformatics*. 2004;20(2): 216–25.
- Hewett M, Oliver DE, Rubin DL, et al. PharmGKB: the Pharmacogenetics Knowledge Base. *Nucleic Acids Res*. 2002;30(1): 163–5.
- Veterans Administration National Drug File (NDF). Available at: <http://www.va.gov/vdl/Clinical.asp?appID=89>. Accessed January 4, 2005.
- Ng AY, Jordan MI. On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes. In: *Advances in Neural Information Processing Systems (NIPS)*. 2002;14:609–16.
- Yeh AS, Hirschman L, Morgan AA. Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup. *Bioinformatics*. 2003;19(suppl 1):i331–9.
- Chang JT, Altman RB. Extracting and characterizing gene-drug relationships from literature. *Pharmacogenetics*. 2004;14(9): 577–86.
- Hirschman L, Park JC, Tsujii J, Wong L, Wu CH. Accomplishments and challenges in literature data mining for biology. *Bioinformatics*. 2002;18(12):1553–61.
- Regev Y, Finkelstein-Landau M, Feldman R. Rule-based extraction of experimental evidence in the biomedical domain—the KDD Cup 2002 (task 1). *SIGKDD Explorations newsletter*. 2003;4(2):90–2.
- Shi M, Edwin DS, Menon R, et al. A machine learning approach for the curation of biomedical literature. *Advances in Information Retrieval*. 2003;2633:597–604.
- Hersch W, Bhupatiraju RT. TREC Genomics Track Overview, NIST SP 500–255. In: *TREC Twelfth Text Retrieval Conference*; 2003; Gaithersburg, MD; 2003.