Research Paper ■

Creating a Text Classifier to Detect Radiology Reports Describing Mediastinal Findings Associated with Inhalational Anthrax and Other Disorders

WENDY WEBBER CHAPMAN, PHD, GREGORY F. COOPER, MD, PHD, PAUL HANBURY, BS, BRIAN E. CHAPMAN, PHD, LEE H. HARRISON, MD, MICHAEL M. WAGNER, MD, PHD

Abstract **Objective:** The aim of this study was to create a classifier for automatic detection of chest radiograph reports consistent with the mediastinal findings of inhalational anthrax.

Design: The authors used the Identify Patient Sets (IPS) system to create a key word classifier for detecting reports describing mediastinal findings consistent with anthrax and compared their performances on a test set of 79,032 chest radiograph reports.

Measurements: Area under the ROC curve was the main outcome measure of the IPS classifier. Sensitivity and specificity of an initial IPS model were calculated based on an existing key word search and were compared against a Boolean version of the IPS classifier.

Results: The IPS classifier received an area under the ROC curve of 0.677 (90% CI = 0.628 to 0.772) with a specificity of 0.99 and maximum sensitivity of 0.35. The initial IPS model attained a specificity of 1.0 and a sensitivity of 0.04.

Conclusion: The IPS system is a useful tool for helping domain experts create a statistical key word classifier for textual reports that is a potentially useful component in surveillance of radiographic findings suspicious for anthrax.

■ **J Am Med Inform Assoc.** 2003;10:494–503. DOI 10.1197/jamia.M1330.

Early detection of a covert anthrax release is an important problem. As seen in the October 2001 releases involving the U.S. postal service, once a release of anthrax is discovered, it is easier to find and prevent infection in other people who have been exposed to the pathogen.[1] Untreated inhalational anthrax has a case fatality rate of nearly 100% with death usually occurring within only a few days.[2] Recent anthrax cases have shown that antibiotic treatment of the disease in its early phase greatly increases the chance of survival.[3] Therefore, timely discovery of an anthrax epidemic is crucial and may, in the case of a large aerosolized release, result in savings of billions of dollars and many saved lives.[4]

Chest radiograph findings have been a sensitive indicator of inhalational anthrax in recent cases.[3] Many hospital information systems store chest radiograph reports for patient care purposes, facilitating detection of radiologic findings that occur in the early stages of inhalational anthrax infection.[5] Unfortunately, these reports are stored as free text that cannot be easily used by automatic detection systems. This study evaluates our ability to automatically identify patients with mediastinal findings consistent with (but not specific to) inhalational anthrax.

Background

Radiologic Evidence of Inhalational Anthrax

When aerosolized *Bacillus anthracis* endospores enter the body through inhalation, the spores deposit in the alveolar spaces and then are engulfed by alveolar macrophages and transported to the mediastinal, peribronchial, and hilar lymph nodes.[6] After germination, the pathogen produces anthrax toxin, causing hemorrhagic lymphadentis and mediastinitis. The toxin also is transported by systemic circulation, resulting in edema, hemorrhage, necrosis, septic shock, and death.[7]

Symptoms of the disease occur in two phases[7] with death usually occurring less than a week after the onset of symptoms. The first phase is a prodromal period that continues for an average of four days. The prodromal stage resembles an influenza infection with symptoms such as fever, chills, myalgia, malaise, fatigue, and nonproductive cough. A widened mediastinum on chest radiograph that represents mediastinal lymphadenopathy also occurs in the early phase of the disease.[3,7,8] A fulminant second phase lasts

Affiliations of the authors: Center for Biomedical Informatics, University of Pittsburgh, Pittsburgh, Pennsylvania (WWC, GFC, PH, MMW); RODS Laboratory, University of Pittsburgh, Pittsburgh, Pennsylvania (WWC, GFC, MMW); Department of Radiology, University of Pittsburgh, Pittsburgh, Pennsylvania (BEC); Infectious Diseases Epidemiology Research Unit, Department of Medicine and Department of Epidemiology, University of Pittsburgh, Pittsburgh, Pennsylvania (LHH).

This work was supported in part by NLM training grant T15 LM07059 and CDC grant UPO/CCU 318753-02. The authors thank the physicians who read and classified the reports: John Dowling, Jeremy Espino, Jim Hinderup, Kim Mast, and Stylianos Kakoullis. The authors also thank Zhongwei Lu for his assistance in collecting reports and Melissa Saul for her advice.

Correspondence and reprints: Wendy W. Chapman, PhD, Center for Biomedical Informatics, University of Pittsburgh, Suite 8084 Forbes Tower, Pittsburgh, PA 15213; e-mail: <chapman@cbmi.upmc.edu>.

Received for publication: 01/21/03; accepted for publication: 05/13/03.

approximately 24 hours and develops suddenly with the onset of acute respiratory distress, hypoxemia, cyanosis, and, in most cases, death.

The earliest specific clinical findings of inhalational anthrax are radiologic, including mediastinal lymphadenopathy and mediastinal widening. Early detection of patients with mediastinal anthrax findings on chest radiographs may be an important element in automated detection of a covert anthrax release.

Automatic Identification of Chest Radiograph Reports Using the IPS System

Medical language processing systems have successfully extracted radiologic findings from textual reports[9] and have been used to help detect patients with findings consistent with pneumonia,[10–12] stroke,[13] and tuberculosis.[14] In this study we used a statistical text classifier called the Identify Patient Sets (IPS) system[15,16] to detect reports describing mediastinal findings consistent with inhalational anthrax.

A text classification system automatically classifies a set of documents, such as Medline articles or chest radiograph reports, into one of a discrete set of possible categories.[17] The IPS system is a text classifier that helps a user construct a statistical, key word–based model to classify a set of textual documents relative to a target category. We used the IPS system to create a classifier to classify chest radiograph reports based on whether the reports describe mediastinal findings consistent with inhalational anthrax. Because the IPS system is a statistical classifier, it computes the probability that the document belongs to the target category. A probabilistic threshold then can be applied to determine whether the document should be classified into that category.

Description of the IPS System

Figure 1 gives an overview of model creation and application using the IPS system. In this section we describe how a model is created using the IPS system, the components of an IPS model, and how an IPS model is used to predict the target class of an unclassified document.

Creating a Model with the IPS System

To begin constructing a model with the IPS system, the user supplies an unclassified set of documents that acts as a training set. The IPS system displays the training set one document at a time to the user, who classifies the documents as either belonging or not belonging to the target category. Our target category was a report that described mediastinal findings consistent with inhalational anthrax.

While classifying documents, the IPS system displays statistical properties associated with words and phrases (hereafter called *terms*) that appear in the classified documents. From the list of all terms in the classified documents, the user can select terms that discriminate between the two categories of documents. The selected terms and their statistical properties become a *model* for classifying unseen documents.

Components of an IPS Model

An IPS model is comprised of terms and their statistical properties.

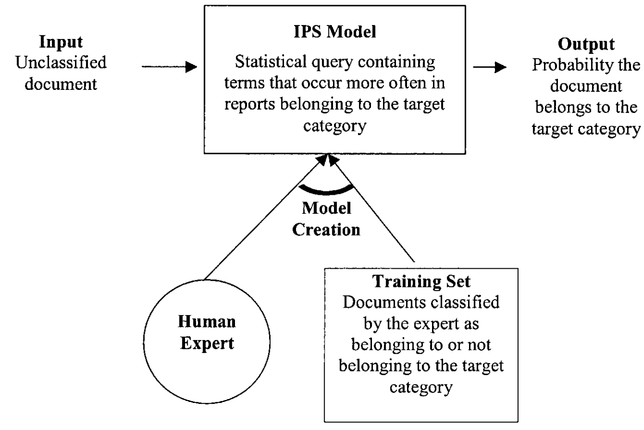

Figure 1. The Identify Patient Sets (IPS) system helps a human user create a statistical, keyword-based classification model. As the user classifies training documents according to whether the documents belong to the target category, the IPS system displays statistical information about the terms occurring in the classified documents. Based on the statistical information and on the users' knowledge of the domain, the user selects terms that discriminate best between the two categories of documents. The selected terms and their statistical properties produce a statistical model that can be applied to an unclassified report to generate a probability that the report belongs to the target category.

Terms in an IPS model. A classifier created with the IPS system consists of a probabilistic model of terms that discriminate between documents classified into the target category and those not classified into the target category. Terms in an IPS model can be single words such as *wide* or phrases from the UMLS such as *wide mediastinum*. The user also can specify a disjunction of terms, hereafter called a *concept*. For example, a concept indicating widening could be represented with the following disjunction: *widening* or *widened* or *wide* or *wider*. The IPS model would consider the widening concept to be present if any of the individual terms in the disjunction appeared in the document.

The fact that a term appears in a document does not necessarily indicate that the dictating physician believed the finding was present in the patient; in clinical reports many terms are used in a negative context, as in *the mediastinum is not widened*. The IPS system interprets terms as negated if they occur in one of several negation patterns that IPS recognizes. The IPS system treats *widened* and *widened* (*negated*) as completely different terms. The algorithm for determining whether a term is negative is simple but performs with fairly high accuracy, as described in Chapman et al.[18,19]

Statistical properties of terms. An IPS model is a combination of concepts and statistical properties based on how frequently the terms in the concept occur in the classified documents. The user can view the raw frequencies of any concept or individual term in classified and unclassified documents. In addition, the frequencies are used to calculate likelihood ratios (LR+ and LR−) for all terms in the document and for concepts in the current model. The LR+ of a term is the odds that the term occurs in a document belonging to the target category; the LR− of a term is the odds that the term does not occur in a document belonging to the target category:

$$LR+(term \in D) = \frac{P(term \in D | D \in T)}{P(term \in D | D \notin T)} \quad (1)$$

$$LR-(term \in D) = \frac{P(term \notin D | D \in T)}{P(term \notin D | D \notin T)} \quad (2)$$

where *term* is a word, phrase, or concept comprised of several terms, *D* is a document, and *T* is the target category. The probabilities in equations 1 and 2 are estimated from the raw frequencies using a Bayesian prior that effectively "smoothes" the estimates. Such smoothing is especially helpful when the raw frequencies are small. The likelihood ratios of terms in the model are derived automatically from the training documents or can be manually specified by the user. Alternatively, the user can specify that the terms be Boolean instead of probabilistic.

Figure 2 shows an example of an IPS model we created for illustrative purposes from part of the test set used in this study. The model consists of a conjunction of three concepts: a *widening* concept, a concept for the *mediastinal location*, and a concept for *normal presentation* (one concept is represented as one row on the IPS interface). The IPS system used the 106 classified documents to calculate the LR+ and LR− for the concepts in the model. (In Figure 4 the *widening* concept has a LR+ of 20.571, meaning the concept is 20.571 times more likely to be present in a document of interest, whereas the *normal presentation* concept is 3.950 times more likely to be absent in a document of interest.)

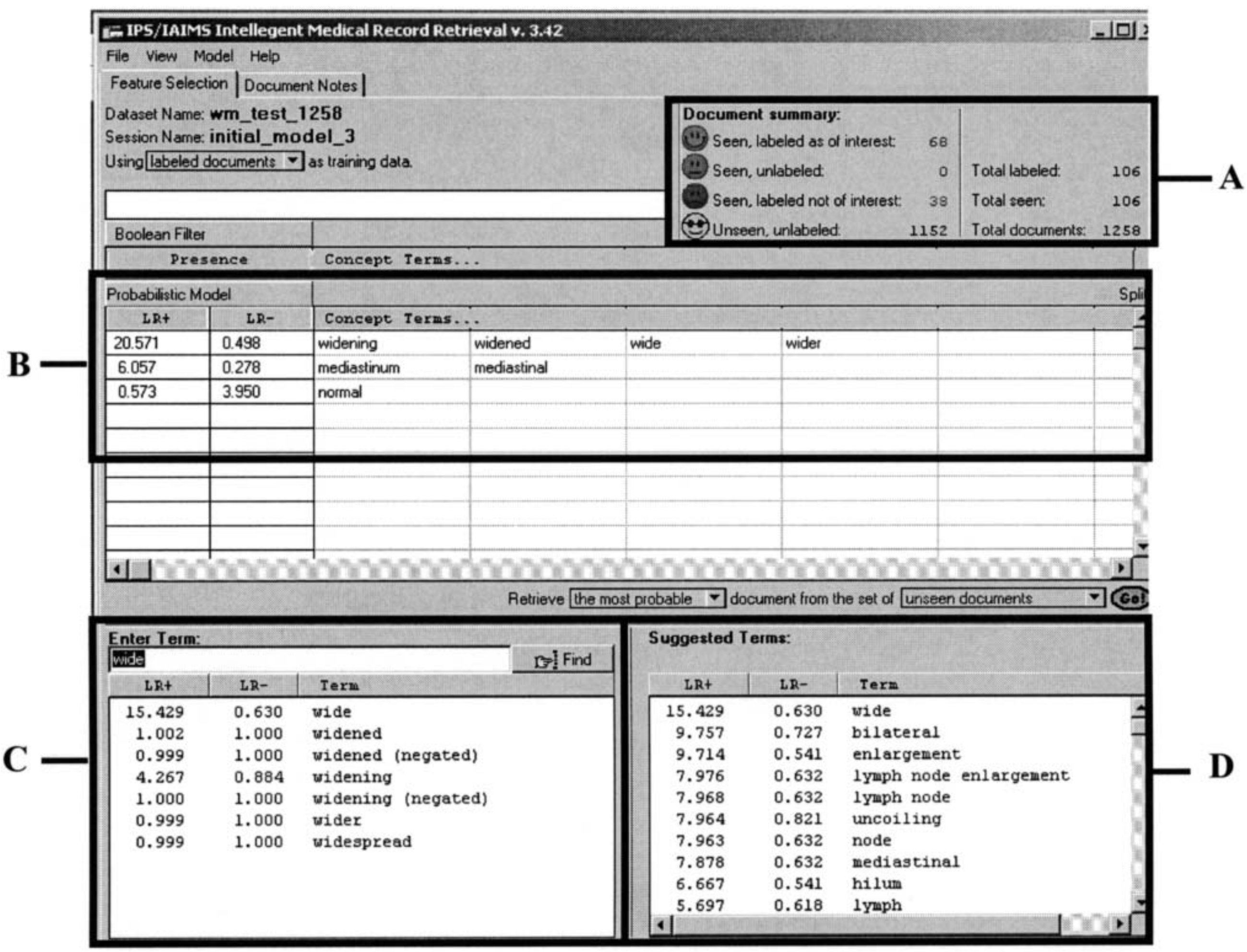

Figure 2. An example Identify Patient Sets (IPS) probabilistic model for detecting chest radiograph reports describing anthrax findings. (**A**) Distribution of documents labeled (i.e., classified) by the user. (**B**) The probabilistic model contains a conjunct of three concepts: a *widening* concept that is a disjunction of four terms; a *mediastinal location* concept that is a disjunction of two terms; and a *normal presentation* concept that is a single term. The model also displays the odds that those terms are present or absent in a document labeled of interest. For example, the IPS system calculated that the *widening* concept is 20.571 times more likely to be present in a document of interest, and the *normal presentation* concept is 3.95 times more likely to be absent. (**C**) Users can search on terms in the document set. C displays all terms in all documents that begin with the search term *wide*. The user can drag and drop any of these terms into the model. The IPS system differentiates between terms used in the document in a positive context and in a negative context. (**D**) The IPS system displays a list of terms that differentiate between documents labeled of interest and documents labeled not of interest. The user can drag terms from the Suggested Terms box to the model.

When a domain expert uses the IPS system to create a model, the knowledge of the expert is enhanced with statistical information from terms in the documents. The resulting model integrates expert and statistical knowledge, and the model may be more complete than a key word–based model that the expert could have created without the IPS system.

Classifying an Unseen Document

An IPS model can be viewed as a probabilistic query that can be applied to a previously unclassified document *D*, generating the probability that the document belongs to the target category *T*. A document containing the terms in the model would be assigned a higher probability of belonging to the target category than a document without the terms. The posterior probability is calculated from the prior odds that the document belongs to the target category *T* and the likelihood ratios for every concept $c_i \in C$ in the model:

$$P(T|D) = \frac{\frac{P(T)}{1-P(T)} \times \Pi_{c_i \in C} LR(c_i)}{1 + \left(\frac{P(T)}{1-P(T)} \times \Pi_{c_i \in C} LR(c_i)\right)} \qquad (3)$$

Equation 3 assumes that the concepts in the model are independent, conditioned on whether the document is of interest.

Methods

We created an IPS model to classify chest radiograph reports according to whether the reports described mediastinal findings consistent with inhalational anthrax. To evaluate the performance of the IPS model, we applied it to a set of unclassified reports and compared its classifications against physician classifications of the same reports.

Below we describe the definition of mediastinal anthrax findings used for this project, the classifiers we compared, gold standard judgments of the reports, and the evaluations we performed.

Definition of Target Category

Due to lack of chest radiograph reports for actual anthrax patients, our goal was to identify radiograph reports describing findings that might be seen on the radiograph of a patient with inhalational anthrax. We defined the target category based on literature descriptions of anthrax previous to the 2001 anthrax attacks, before we had an idea of what weapons-grade inhalational anthrax infections looked like. Case studies before 2001 showed that nearly all inhalational anthrax cases had either radiographic or autopsy confirmation of mediastinal widening.[7,8] We designed this study to detect radiologic mediastinal abnormalities consistent with anthrax and defined relevant abnormalities as mediastinal, paratracheal, peribronchial, and hilar lymphadenopathy or mediastinal widening. Our goal was not to detect patients with anthrax but to detect reports indicating that any of these findings might be present on the radiograph. The differential diagnosis for these findings includes other disorders in addition to anthrax, including lymphoma or other neoplasms, congenital disorders, or aortic dissection. However, a chest radiograph showing mediastinal lymphadenopathy would certainly raise suspicion for anthrax.

Descriptions of Classifiers

IPS Initial Model

As described in the Background section, the IPS system selects documents from the training set to be classified by the user based on how well the terms in the current IPS model match the documents. Our aim was to detect chest radiograph reports with mediastinal findings consistent with anthrax, so we created an initial model to detect the most evident mediastinal anthrax finding on chest radiograph—mediastinal widening. The model consisted of two concepts: the first concept represented widening and contained the term *wide*; the second concept represented the mediastinal location and contained the terms *mediastinum* and *mediastinal*. We based our initial model on a key word classifier developed independently and currently in use at University of Pittsburgh Medical Center (UPMC). The key word classifier was created by a physician who manually read through a set of training reports to determine the best key words for capturing patients with mediastinal adenopathy consistent with inhalational anthrax and has been in use for three years.

IPS Refined Model

Authors WWC and LHH used the IPS system to modify and hopefully improve the initial model.* The potential training set consisted of 69,508 chest radiograph reports stored on the MARS hospital information system[20] at the Presbyterian Hospital of UPMC from January 1 through December 31, 1999. One advantage of using the IPS system to create a classification model is the system's ability to enrich the training set with positive examples by selecting training cases more likely belonging to the target category for manual classification. Although the potential training set was large (69,508 reports), we manually classified only a subset of these documents (1,652 reports). Reports we manually classified consisted of (1) reports assigned the highest probability of belonging to the target category, (2) reports assigned the lowest probability of belonging to the target category (training a good model requires negative training examples), and (3) randomly selected reports. Below we describe the feedback loop in which we classified the training set and incrementally changed the initial IPS model:

1. Based on the terms in the current model, the IPS system selected from the pool of potential training documents the unclassified chest radiograph report that received the highest probability (or lowest, depending on what we requested) of describing mediastinal findings of interest to the study.
2. We read the report and classified it as a document of interest (i.e., consistent with mediastinal findings of interest) or not of interest.
3. The statistical properties of the terms in the classified report(s) were updated based on whether the current report was classified of interest or not.
4. The IPS system displayed statistical properties of the terms in the current model and of all other terms found in

*Initially, WWC and LHH classified reports together. After classifying a few hundred reports together, WWC classified the reports alone, consulting LHH when necessary.

the classified reports. Based on the displayed information, we decided whether to add terms to or remove terms from the current model.

Steps 1 through 4 were repeated until we were satisfied that the model was accurately detecting reports of interest. Deciding you are satisfied with a model is a fuzzy judgment related to the user's belief about the quality and comprehensiveness of the terms in the model and to the relevance of the reports being ranked by the current model. Once the reports selected by the IPS system as having the highest probability of belonging to the target class obviously did not belong to the target class, we believed we had already retrieved all the reports with mediastinal findings of interest to the study. To test our belief, we applied an IPS feature that shows the user randomly selected documents from the set of unlabeled documents. Using a power calculation for proportions with a power of 0.95 and alpha of 0.05, we determined 250 randomly selected documents would adequately estimate the true population of the unlabeled documents. Review of 250 randomly selected reports showed no reports describing mediastinal findings of interest. In all, we spent approximately 60 person-hours classifying reports and refining the IPS model.

The final training set consisted of 1,682 (910 positive and 772 negative) chest radiograph reports selected by the IPS system from the larger pool of 69,508 reports. Throughout the training, we added new terms to the model based on the Suggested Terms list displayed by the IPS system (e.g., Fig. 2D). New terms we added were related locations (e.g., *hilar* and *paratracheal*), lexical variants or misspellings found in the training set (e.g., *bihilar* and *peritracheal*), and negated terms that seldom occurred in positive documents (e.g., *lymphadenopathy* [*negated*]). The final model created with the IPS system is shown in Figure 3. We manually set Concept 1 (*Mediastinum*) and Concept 2 (*Widening*) to be Boolean concepts (i.e., the concepts must be present for a document to be classified as positive). Concepts 3 and 4 are only applied to the set of documents classified as positive by the Boolean concepts and effectively order the documents retrieved with the Boolean concepts from highest to lowest probability.

Concept 1: *Mediastinum* **Boolean**

Terms: "bihilar" or "bronchial lymph node" or "bronchial" or "hilar lymph nodes" or "hilar lymph node" or "hilar" or "hila" or "hilum" or "mediastinal lymph nodes" or "mediastinal lymph node" or "mediastinal" or "mediastinum" or "paramediastinal" or "paratracheal lymph nodes" or "paratracheal" or "peritracheal" or "tracheobronchial lymph nodes" or "tracheobronchial"

Concept 2: *Widening* **Boolean**

Terms: "enlarged lymph nodes" or "lymph node enlargement" or "lymphadenopathy" or "widened" or "widening" or "wide"

Concept 3: *Absence of Mediastinal Widening* **LR-:** 1.334

Terms"lymphadenopathy (negated)" or "mediastinal lymph node (negated)" or "mediastinal lymph node enlargement (negated)" or "mediastinal lymphadenopathy (negated)" or "mediastinal widening (negated)"

Concept 4: *Mediastinal Widening* **LR+:** 12.442

Terms: "hilar lymphadenopathy" or "mediastinal lymph node enlargement" or "mediastinal lymphadenopathy" or "mediastinal widening"

Figure 3. The final Identify Patient Sets (IPS) Refined model for identifying reports describing mediastinal anthrax findings. A Boolean concept must be present for a document to be classified as positive. The likelihood ratio (LR+) represents the odds that the term appears in a positive document.

Test Set

We tested the performance of the classifiers on a test set of all chest radiograph reports dictated at Presbyterian Hospital between January 1 and December 31, 2000 (n = 79,032). This work was done with approval from the University of Pittsburgh's Institutional Review Board.

Gold Standard Judgments

The gold standard determination of whether a report described findings consistent with anthrax was the majority vote of three physicians. Recruiting physicians to read and classify 79,032 documents would have been impractical. The prevalence of radiographic findings consistent with anthrax was estimated to be very low (1.3% in the training documents), so random selection of documents would provide only a small number of positive documents and would be insufficient for making valid conclusions. To minimize the number of documents physicians needed to classify while maximizing the number of positive documents in the sample, we used an unbiased sampling strategy that we describe below to select a subset of 1,258 documents for physicians to judge and then extrapolated our findings to the entire test set of 79,032 documents.

Five internists with variable experience levels (three years to 25 years) were trained to judge whether a chest radiograph report described anthrax findings. For training, each physician read and judged a common set of 25 practice reports selected from the training set used to create the IPS model. Differences in the physician judgments on the practice documents were discussed among the group of physicians until everyone agreed on the judging task. Four of the physicians then read and judged the subset of test documents, while one physician acted as a mediator on disagreements.

Each of the four physicians classified half of the 1,258 test documents. Thus, every document was classified by two physicians. To avoid potential bias due to pairing of physicians, documents were assigned so that every physician classified an equal number of documents in common with every other physician. The fifth physician determined the correct classification for documents receiving conflicting classifications.

Sampling Method

We used a sampling method combining random and enriched sampling to select a subset of documents from the entire test set for physician review. Both the IPS Initial and the IPS Refined classifiers were applied to the entire test set of 79,032 documents shown as the rectangle in Figure 4. The ovals P_I and P_R respectively represent documents classified as positive (i.e., describing mediastinal findings of interest) by the IPS Initial and IPS Refined classifiers. Physicians read and classified every document contained in the ovals. We sampled randomly from documents not classified positive by either of the classifiers, represented as the gray area of the rectangle. Thus, the set of documents classified by physicians included (1) all documents from the test set that were classified positive by either of the classifiers (n = 629) and (2) an equally sized

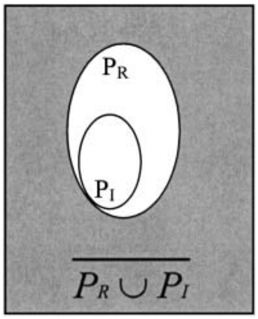

Figure 4. Schematic representation of the test set. The rectangle represents all chest radiograph reports stored in MARS for the year 2000 (n = 79,032). The ovals P_R and P_I respectively represent reports classified as positive by the Identify Patient Sets (IPS) Refined and IPS Initial models. The gray area represents reports not classified as positive by either of the classifiers. P_I is a subset of P_R, because the terms used in the IPS Initial model are also in the IPS Refined model.

Table 1 ■ Estimates of Outcome Measures

$$\text{Sensitivity} = \frac{TP}{TP + (FN \times SF)}$$

$$\text{Specificity} = \frac{(TN \times SF)}{(TN \times SF) + FP}$$

$$\text{Positive predictive value} = \frac{TP}{TP + FP}$$

$$\text{Negative predictive value} = \frac{(TN \times SF)}{(TN \times SF) + (FN \times SF)}$$

NOTE. SF: D_n/D_r = 124.65; D_n: Documents not classified as positive by IPS classifier = 79,032 − 629 = 78,403.
D_r: Documents randomly selected from D_n = 629.

random sample of documents that were not classified positive by either of the classifiers (n = 629), totaling 1,258 chest radiograph reports.

Measurements

We performed primary and secondary analyses on the classifiers. The primary analysis measured classification performance. Based on results of the primary analysis, we tuned the IPS model to be more sensitive and performed a secondary analysis.

Because output from the IPS model is probabilistic, we used the area under the receiver operating characteristic (ROC) curve[21,22] as the main outcome measure. We calculated sensitivity and specificity from the binary output of the IPS Initial classifier.

To generate the outcome measures, we calculated true-positive (TP), false-positive (FP), true-negative (TN), and false-negative (FN) counts for both classifiers. Physicians read only a subset of documents classified negative by the classifiers. We assumed that the proportion of TN and FN documents in the randomly sampled set represented the proportion of TN and FN documents in the entire test set. The ratio of negatively classified documents in the entire set to the number of negatively classified documents in the sampled set was 78,403:629 (124.65:1). Therefore, we estimated the number of TN and FN counts for the entire set by multiplying TN and FN counts from the subset by a sampling factor of 124.65. Table 1 describes calculations of outcome measures for the IPS Refined classifier based on our sampling method.

Outcome measures for the IPS Initial classifier were calculated similarly to those shown in Table 1, except that only TN and FN counts from the randomly selected subset (gray area in Figure 4) were multiplied by SF (TN and FN counts from P_R-P_I in Figure 4 were not multiplied by SF).

The outcome measures described in Table 1 were used to calculate the area under the ROC curve (AUC) using trapezoidal integration of the IPS Refined classifier. The AUC is a common measure of binary classification accuracy that ranges between 0.5 (chance classification) and 1.0 (perfect classification).[21,22] We calculated the AUC and 90% confidence intervals for the IPS Refined classifier using the bootstrapping method[23,24] as follows: From the 1,258 test documents that were classified by physicians we randomly selected with replacement a sample of 1,258 documents. Varying the classification threshold from 0 to 1 by 1% increments, we calculated the sensitivity and 1 − specificity rates (as described in Table 1) for the sampled documents. From the sensitivity and 1 − specificity rates, we calculated the AUC using trapezoidal integration. We repeated this process 2,000 times to generate 2,000 AUCs, which then were sorted in ascending order. We report here the median value and 90% confidence intervals for the AUC, which were taken directly from the sorted list of 2,000 bootstraps.

Secondary Analysis

We performed a secondary analysis to investigate how much positive predictive value (PPV) and specificity would decrease with a more sensitive IPS model. Based on manual review of documents falsely classified as negative in the primary analysis, we added terms to the IPS model that could detect additional positive documents. Adding more terms to a concept in an IPS model will result in a potentially more sensitive classifier, because more documents will be classified as positive. However, because some of the documents classified as positive may be false-positives, we measured the specificity and PPV resulting from the modifications using the outcome measures described in Table 1.

We created two modified IPS models. The first modification (IPS-M1) was created by adding four additional terms to the Concept 2 (*widening*) of the original IPS model shown in Figure 3 (*prominent*, *prominence*, *soft tissue*, and *mass*). The second modified classifier (IPS-M2) was created to be even more sensitive by adding a total of nine words to the Concept 2 of the original IPS model, including the four words in IPS-M1 and five additional words (*opacity*, *opacification*, *opacifications*, *density*, and *densities*).

IPS-M1 and IPS-M2 were applied to the 1,258 documents classified by physicians. Because we were interested in the most sensitive classifier, we calculated outcome measures for IPS-M1 and IPS-M2 at the most sensitive probability classification threshold, i.e., the threshold at which any document containing the Boolean terms was classified as positive.

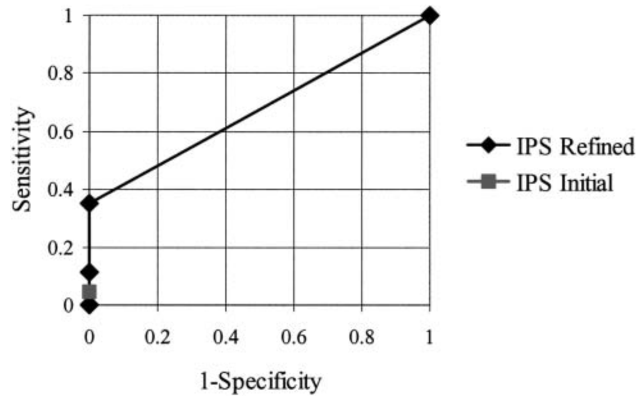

Figure 5. Receiver operating characteristic (ROC) curve for the Identify Patient Sets (IPS) Refined classifier and point on the ROC plot for the IPS Initial classifier.

Results

Physicians read and classified 1,258 chest radiograph reports, classifying 49% (616 of 1,258) of the reports as positive. Multiplying FN counts by SF, we estimated the prevalence of positive documents in the entire sample of 79,032 reports to be 2.19% (1,729 positives).† Interrater agreement of physician classifications was 97.7% (1,229/1,258).

Primary Analyses

The AUC of the IPS Refined classifier was 0.677 (90% CI = 0.628 to 0.772). At all probabilistic classification thresholds, the IPS Refined classifier maintained a specificity of 0.999; the highest sensitivity achieved was 0.351. The classification threshold yielding 0.351 sensitivity classified any document containing the Boolean concepts as positive and was used in subsequent comparisons with other models.

The IPS Initial classifier performed with a sensitivity of 0.043 and a specificity of 1.0. Figure 5 plots the IPS Refined classifier's ROC curve with the single performance point from the IPS Initial classifier.

Table 2 compares the performance of IPS Initial with IPS Refined. IPS Refined achieved higher sensitivity than IPS Initial while maintaining high specificity and PPV and a low false-negative rate (FNR = 1 − NPV).

Secondary Analysis

Table 2 also shows the performance of the modified IPS models, IPS-M1 and IPS-M2. With modifications learned from the primary analysis, IPS-M2 classified 0.856 of the positive documents correctly with a specificity of 0.97 and a PPV of 0.41.

Error Analysis

False-negative Classifications

The original IPS model generated nine false-negative classifications. Examining the false-negatives gave some insight into the reasons for a sensitivity of only 0.35. First, we realized the original IPS model was not tuned to achieving maximum sensitivity. In creating the model, we purposefully left out ambiguous words such as *mass* and *opacity* for fear of increasing the number of false-positives. However, we instructed the physicians to classify a document that described *any* finding that could be consistent with anthrax as positive, including uncertain and ambiguous findings.

Secondary Analysis to Address False Negatives. A secondary analysis examined the IPS classifier's performance when we included additional terms found in false-negative documents. As seen in Table 2, specificity and PPV decreased with addition of the terms, but sensitivity approached 0.86. IPS-M2 accurately detected all but two of the documents classified as positive by physicians. Because the secondary analysis relied on knowledge we gained from examining the test set, the estimated performance statistics indicate a "ceiling" level of performance. On a new test set we would expect the performance of IPS-M1 and IPS-M2 to fall somewhere between that of the IPS Refined and the performance we obtained for IPS-M1 and IPS-M2 on the current test set. In spite of experimental bias, the secondary analysis is informative, because we learned to what extent the specificity is decreased when attempting to boost sensitivity.

The second source of false negatives was an inconsistency in gold standard classifications of documents describing calcified perihilar lymph nodes, which most likely indicate a chronic—not acute—process. Six documents describing calcified lymph nodes in the test set were all classified as negative by the IPS classifier. Physicians classified two of the documents positive and the other four negative. If physicians had consistently classified all six documents as negative, sensitivity of IPS Refined would increase to 0.41, and sensitivity of IPS-M2 would be 1.0.

Review of the 22 documents in the test set that were falsely classified as positive identified two main causes. First, ten documents contained terms that were negated in the document but were not negated by the IPS system's negation processor. Failed negation tagging involved two patterns not currently used by the negation processor. The relevant mediastinal adenopathy in the following two sentences were not negated: "Evaluate for wide mediastinum" and "The mediastinal lymphadenopathy seen in the previous films has since resolved." Many of the remaining documents (7 of 22) were falsely classified as positive because of the IPS system's simple "bag-of-terms" approach to classification in which the document is viewed by the system as a bag of unrelated terms consisting of one or more consecutive words. A bag-of-terms approach matches terms from the model no matter where the terms appear in the document. For instance, one document described a "wide cardiac silhouette" and later said the "mediastinum was normal." Because the words *wide* and *mediastinum* both appeared in the document—even though the words were in separate sentences—the document was classified as positive.

Discussion

Improved biosurveillance is a national priority,[25] and timeliness of detection and initiation of therapy is especially crucial for those infected with inhalational anthrax.[3,4] Automated detection of suspicious findings using preexisting electronic data complements traditional disease reporting—especially for diseases physicians do not normally see or for early detection in the initial stages of infection when the symptoms

†Estimated prevalence $= \frac{Positives}{Total} = \frac{TP+(FN \times SF)}{Total} = \frac{1{,}729}{79{,}032} = 2.19\%$, where TP = 616, FN = 9, SF = 124.65.

Table 2 ■ Performance of IPS Models

Classifier	Sensitivity	Specificity	PPV	FNR
IPS Initial	0.043 (74/1,729)	1.0 (77,305/77,305)	1.0 (74/74)	0.021 (1−77,305/78,960)
IPS Refined	0.351 (607/1,729)	0.999 (77,283/77,305)	0.965 (607/629)	0.014 (1−77,283/78,405)
IPS-M1	0.712 (1,230/1,729)	0.985 (76,161/77,305)	0.518 (1,230/2,374)	0.007 (1−76,161/76,660)
IPS-M2	0.856 (1,480/1,729)	0.972 (75,164/77,305)	0.409 (1,480/3,621)	0.003 (1−75,164/75,413)

NOTE. IPS Initial: Based on UPMC key word search; *widening* ("wide") and *mediastinal* ("mediastinum" and "mediastinal") concepts.
IPS Refined: IPS Initial model refined by authors—outcome measures obtained by applying the most sensitive classification threshold.
IPS-M1: IPS Refined with four additional terms in the *widening* concept ("mass," "soft tissue," "prominent," and "prominence").
IPS-M2: IPS-M1 with five additional terms in the *widening* concept ("opacity," "opacification," "opacifications," "density," and "densities").

are not definitive. However, automated surveillance of inhalational anthrax is not straightforward, because the underlying prevalence of inhalational anthrax in the population is close to zero, making the positive predictive value of any detection system also zero until an actual anthrax release occurs.

The goal of anthrax surveillance may be to find the first case of anthrax as soon in the course of the disease as possible. The October 2001 attacks showed that once a single case was diagnosed, subsequent cases were diagnosed earlier in the course of their disease, patients at risk for infection presented to the emergency department earlier, and infected patients were treated more effectively[3]—showing the importance of early diagnosis not only for the patient being diagnosed but also for other infected people. In this way, public health disease detection differs from the diagnosis of disease in an individual patient.

Current anthrax surveillance is reliant on the astute clinician for diagnosis of the first case. After the 2001 anthrax attacks, physician awareness of anthrax has been heightened; whether physicians need automated methods to assist them in detecting anthrax outbreaks is an unanswered question. Automated anthrax detection may improve current surveillance results by either diagnosing undiagnosed cases or detecting patterns suspicious for a large-scale outbreak earlier than individual physicians. Because chest images currently provide the earliest specific evidence of anthrax,[3] detection of reports describing mediastinal findings consistent with anthrax is an important component of an automated anthrax detection system.

The IPS Initial classifier, which represents the current anthrax detection capability at our institution, only detected 4.3% of the positive documents. In our primary analysis, the IPS Refined classifier achieved a sensitivity of 35% with a PPV of 96.5%. and a 1.4% (1−NPV) false-negative alarm rate. With additional terms added to the IPS Refined model, the sensitivity increased to 85.6%, but the PPV also dropped to 41%. The utility of implementing a detection system with only 41% PPV depends on the perceived cost and benefit of detection. The cost–benefit ratio of anthrax detection is influenced by the detector's sensitivity and PPV, the prior belief that an anthrax attack is going to occur, and the resources available for investigation of suspicious cases.

One possible scenario for locating the first patient infected with anthrax is to exhaustively screen all chest radiograph reports for mediastinal findings of interest. In 2000 at Presbyterian Hospital, 79,032 reports required review (1,520 reports per week), and 1,729 of the reports actually described mediastinal findings of interest (2.19% PPV). The benefit of an exhaustive review would be perfect sensitivity. A second scenario is to review only cases with chest radiographs showing mediastinal findings consistent with anthrax. If anthrax surveillance for the year 2000 consisted of reviewing all reports detected by the IPS classifier (IPS-M2), 3,621 reports would need to be reviewed (70 per week), of which, 1,480 (41%) would truly describe mediastinal findings of interest and may require further chart review to look for other signs of inhalational anthrax. Of the 1,729 cases with actual mediastinal findings, 249 would not be detected by the classifier (14%).

A third scenario would be to monitor output of the IPS classifier with a pattern detection algorithm that looks for temporal or spatial patterns different from what is expected during a non-outbreak period. Effective algorithms to detect unexpected clusters of cases currently are being developed and used for outbreak detection.[26,27] In this scenario a smaller number of reports would require review, because the detection algorithm would expect to see some number of patients with mediastinal findings of interest even during non-outbreak periods. The system would only generate an alarm if the number of cases detected as positive exceeded a threshold that would be set based on the costs of false positives and negatives and on the benefits of detecting true positives.

In clinical decision analysis, a diagnostic test typically requires high PPV to be useful. The utility of automated outbreak detection of anthrax depends not only on the PPV of the detection system but also on the prior belief of an anthrax attack and the amount of resources available for investigation of suspicious cases. If prior belief and resources were low, the threshold could be set high, and a low PPV would be less of a concern, because an alarm would be generated only with an extreme increase in the number of suspicious cases. An extreme increase in the number of positive cases may occur only during an actual large-scale anthrax attack. If prior belief of an anthrax attack were high, resources for investigating suspicious cases could be raised, and the threshold could be lowered. In this case, even a small-scale attack could be detected, and a low PPV would be considered worth the investment in resources. Moreover, as the prevalence of actual anthrax cases increases, the PPV of the detected cases will also increase.

Output of the IPS classifier could be supplemented with additional information that would increase the PPV and sensitivity of detection. For instance, PPV could be increased by using the IPS classifier in conjunction with gram-positive rods detection from the laboratory—a combination of findings that is rare and would cause immediate suspicion of anthrax. A more sensitive detector also would monitor other radiologic findings such as pneumonic infiltrates and pleural effusions, which were as predictive as mediastinal abnormalities in the 2001 weapons grade anthrax infections.[3] Accurate methods exist for detecting chest radiograph reports with pneumonic infiltrates and pleural effusions.[9,10,28]

Detecting anthrax in its earliest stages may be accomplished by monitoring flulike symptoms.[29] Research in syndromic surveillance is examining the usefulness and timeliness of various sources of data for automatically detecting patients with syndromes of interest to public health, including free-text triage chief complaints,[30,31] ICD-9 admission codes,[31,32] and over-the-counter medication sales.[33,34] Because flulike symptoms could indicate myriad health problems other than inhalational anthrax, an ideal anthrax detection system would monitor a combination of nonspecific syndromic symptoms; specific radiologic findings including infiltrates, pleural effusion, and mediastinal adenopathy; and microbiology findings.

The best methods for anthrax surveillance are not known. What is recognized is that a large-scale anthrax attack would result in an increase in the number of patients with some combination of syndromic, radiographic, and bacterial findings. We believe automated surveillance could facilitate detecting such an attack.

Sampling Method

The estimated sensitivity statistic reported in this article was influenced by the high proportion of negative documents in the full test set that were subject to multiplication by a large sampling factor. Only nine of the 629 documents classified as negative by the IPS classifier were falsely classified as negative. However, multiplying the nine false negatives by the sampling factor of 124.65 produced an estimate of 1,122 false negatives. Had we not used the sampling method described in this report and instead only randomly sampled reports for physician review, we could have calculated sensitivity in a more straightforward manner; however, the statistic might not have been reliable because of the small number of positive documents that would appear in a random sample. We believe the sampling method we used is a reasonable method that may be useful in other studies involving text classification of rare conditions.

The IPS System

Text processing systems are becoming accurate enough to be applied to real-world medical problems. NLP systems can accurately extract many types of radiologic findings, but not everyone has access to an NLP system. The IPS system is a fairly simple alternative that is potentially useful for constructing classifiers to detect outbreaks from textual documents such as patient medical reports, 911 transcripts, or Web-based queries. The IPS classifier has several advantages over a manually created key word–based classifier. First, the IPS system identifies additional terms from the documents that the user may want to include in the model, including terms a user had not initially thought of (e.g,. *paratracheal lymph nodes*), lexical variants (e.g., *mediastinal*), or even misspellings from the documents (e.g., *peritracheal*). Second, the ability to tag terms as being negated in the document allows the IPS classifier to generate fewer false-positive classifications. Third, the option of a probabilistic output allows the classifier to be fine-tuned for sensitivity or specificity, depending on the classification task.

Limitations and Future Work

The greatest weakness in this study was being too conservative in creating the IPS model when our goal was to build a sensitive screening system that could tolerate false-positive classifications. Eliciting more complex gold standard classifications from the physicians may have compensated for a conservative model. We asked physicians to classify radiograph reports as either describing “some evidence of anthrax” or “no evidence of anthrax.” Asking physicians to rate their certainty of the anthrax evidence so that we could break down our analysis by strength of evidence in the text would have provided a more complete understanding of the system’s performance. Knowing that a feature detection system can accurately detect documents with strong evidence of anthrax but not documents with weak evidence could help establish optimal use of the system.

The IPS system is a naïve Bayesian system that helps an expert create a text classifier by incrementally selecting new terms as he or she expands the training set. It is an open question whether statistical or machine learning algorithms that learn classification terms from text[35] without user intervention may give better results that those presented in this report. We have recently developed and begun to evaluate a fully automated version of the IPS system (AMC) that attempts to find an optimal IPS model for a training set.[36] AMC can be used alone to create a classification model or can be used as the initial IPS model that an expert can then refine. Future work involves comparing manually created IPS models with models created with AMC and other text learning algorithms.

We plan to compare different classification models with activity monitor operating characteristic (AMOC) curves[37] in which time to detect a case with mediastinal findings consistent with anthrax is plotted as a function of the average number of false-positive alarms per day. Time to detection characterizes benefit, and number of false positives per day characterizes cost.

Conclusion

In this study we evaluated our ability to automatically detect chest radiograph reports describing mediastinal findings consistent with inhalational anthrax. Even perfect sensitivity at detecting these radiologic findings does not ensure perfect sensitivity at detecting actual cases of anthrax, but automated detection of these findings is an important component of automated case or pattern detection for anthrax.

References ■

1. Inglesby TV, Henderson DA, Bartlett JG, et al. Anthrax as a biological weapon: medical and public health management. Working Group on Civilian Biodefense. JAMA. 1999;281: 1735–1745.
2. Meselson M, Guillemin J, Hugh-Jones M, et al. The Sverdlovsk anthrax outbreak of 1979. Science. 1994;266:1202–8.
3. Jernigan JA, Stephens DS, Ashford DA, et al. Bioterrorism-related inhalational anthrax: the first 10 cases reported in the United States. Emerg Infect Dis. 2001;7:933–44.

4. Wagner MM, Tsui FC, Espino JU, et al. The emerging science of very early detection of disease outbreaks. J Public Health Manag Pract. 2001;7(6):51–9.
5. Hashimoto S, Murakami Y, Taniguchi K, Nagai M. Detection of epidemics in their early stage through infectious disease surveillance. Int J Epidemiol. 2000;29:905–10.
6. Dixon TC, Meselson M, Guillemin J, Hanna PC. Anthrax. N Engl J Med. 1999;341:815–26.
7. Shafazand S, Doyle R, Ruoss S, Weinacker A, Raffin TA. Inhalational anthrax: epidemiology, diagnosis, and management. Chest. 1999;116:1369–76.
8. Abramova FA, Grinberg LM, Yampolskaya OV, Walker DH. Pathology of inhalational anthrax in 42 cases from the Sverdlovsk outbreak of 1979. Proc Natl Acad Sci U S A. 1993; 90:2291–4.
9. Hripcsak G, Friedman C, Alderson PO, DuMouchel W, Johnson SB, Clayton PD. Unlocking clinical data from narrative reports: a study of natural language processing. Ann Intern Med. 1995;122:681–8.
10. Fiszman M, Chapman WW, Aronsky D, Evans RS, Haug PJ. Automatic detection of acute bacterial pneumonia from chest x-ray reports. J Am Med Inform Assoc. 2000;7:593–604.
11. Chapman WW, Fizman M, Chapman BE, Haug PJ. A comparison of classification algorithms to automatically identify chest x-ray reports that support pneumonia. J Biomed Inform. 2001; 34(1):4–14.
12. Friedman C, Knirsch C, Shagina L, Hripcsak G. Automating a severity score guideline for community-acquired pneumonia employing medical language processing of discharge summaries. Proc AMIA Symp. 1999:256–60.
13. Elkins JS, Friedman C, Boden-Albala B, Sacco RL, Hripcsak G. Coding neuroradiology reports for the Northern Manhattan Stroke Study: a comparison of natural language processing and manual review. Comput Biomed Res. 2000;33(1):1–10.
14. Hripcsak G, Knirsch CA, Jain NL, Pablos-Mendez A. Automated tuberculosis detection. J Am Med Inform Assoc. 1997;4:376–81.
15. Cooper GF, Buchanan BG, Kayaalp M, Saul M, Vries JK. Using computer modeling to help identify patient subgroups in clinical data repositories. Proc AMIA Symp. 1998:180–4.
16. Aronis JM, Cooper GF, Kayaalp M, Buchanan BG. Identifying patient subgroups with simple Bayes'. Proc AMIA Symp. 1999:658–62.
17. Mitchell TM. Machine Learning. Boston, MA: McGraw-Hill, 1997.
18. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform. 2001;34:301–10.
19. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. Evaluation of negation phrases in narrative clinical reports. Proc AMIA Symp. 2001:105–9.
20. Young RJ, Vries JK, Councill CD. The medical archival system: an information retrieval system based on distributed parallel processing. Inf Process Manage. 1991;27:379.
21. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982; 143(1):29–36.
22. Metz CE. Basic principles of ROC analysis. Semin Nucl Med. 1978;8:283–98.
23. Dofrman DD, Berbaum KS, Lenth RV. Multireader, multicase receiver operating characteristic methodology: a bootstrap analysis. Acad Radiol. 1995;2:626–33.
24. Efron B. Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. Biometrika. 1981;68: 589–99.
25. Sanger DE. Bush plans early warning system for terror. New York Times, 2002;Feb 6:11.
26. Tsui FC, Wagner MM, Dato V, Chang CC. Value of ICD-9 coded chief complaints for detection of epidemics. Proc AMIA Symp. 2001:711–5.
27. Wong W, Moore AW, Cooper G, Wagner M. Rule-based anomaly pattern detection for detecting disease outbreaks. Proceedings of the 18th National Conference on Artificial Intelligence (AAAI-02). Edmonton, Alberta, Canada, 2002.
28. Fiszman M, Haug PJ. Using medical language processing to support real-time evaluation of pneumonia guidelines. Proc AMIA Symp. 2000:235–9.
29. Quenel P, Dab W, Hannoun C, Cohen JM. Sensitivity, specificity and predictive values of health service based indicators for the surveillance of influenza A epidemics. Int J Epidemiol. 1994; 23:849–55.
30. Chapman WW, Wagner M, Ivanov O, Olszewski R, Dowling JN. Syndromic surveillance from free-text triage chief complaints. J Urban Health. 2003(suppl) (in press).
31. Ivanov O, Wagner MM, Chapman WW, Olszewski RT. Accuracy of three classifiers of acute gastrointestinal syndrome for syndromic surveillance. Proc AMIA Symp. 2002:345–9.
32. Espino JU, Wagner MM. Accuracy of ICD-9-coded chief complaints and diagnoses for the detection of acute respiratory illness. Proc AMIA Symp. 2001:164–8.
33. Goldenberg A, Shmueli G, Caruana RA, Fienberg SE. Early statistical detection of anthrax outbreaks by tracking over-the-counter medication sales. Proc Natl Acad Sci. 2002;99: 5237–40.
34. Hogan WR, Tsui FC, Ivanov O, et al. Early detection of pediatric respiratory and diarrheal outbreaks from retail sales of electrolyte products. J Am Med Inform Assoc. In press.
35. Mladenic D. Text-learning and related intelligent agents: A survey. IEEE Intelligent Systems. 1999;14(4):44–54.
36. Visweswaran S, Hanbury P, Saul M, Cooper GF. Detecting adverse drug events in discharge summaries using variations on the simple Bayes model. Proc AMIA Symp. In press.
37. Fawcett T, Provost F. Activity monitoring: noticing interesting changes in behavior. In: Madigan C (ed). Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 1999:53–62.