Downloaded from jamia.bmj.com on May 7, 2013 - Published by group.bmj.com



OPEN ACCESS

# Identifying medical terms in patient-authored text: a crowdsourcing-based approach

Diana Lynn MacLean, Jeffrey Heer

▶ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/amiajnl-2012-001110).

Department of Computer Science, Stanford University, Stanford, California, USA

**Correspondence to**
Diana Lynn MacLean, Department of Computer Science, Stanford University, 372 Gates Hall (3B Wing), Stanford, CA 94305-9035, USA; malcdi@stanford.edu

Received 23 May 2013
Revised 4 March 2013
Accepted 15 April 2013

**To cite:** MacLean DL, Heer J. *J Am Med Inform Assoc* Published Online First: [*please include* Day Month Year] doi:10.1136/amiajnl-2012-001110

## ABSTRACT

**Background and objective** As people increasingly engage in online health-seeking behavior and contribute to health-oriented websites, the volume of medical text authored by patients and other medical novices grows rapidly. However, we lack an effective method for automatically identifying medical terms in patient-authored text (PAT). We demonstrate that crowdsourcing PAT medical term identification tasks to non-experts is a viable method for creating large, accurately-labeled PAT datasets; moreover, such datasets can be used to train classifiers that outperform existing medical term identification tools.

**Materials and methods** To evaluate the viability of using non-expert crowds to label PAT, we compare expert (registered nurses) and non-expert (Amazon Mechanical Turk workers; Turkers) responses to a PAT medical term identification task. Next, we build a crowd-labeled dataset comprising 10 000 sentences from MedHelp. We train two models on this dataset and evaluate their performance, as well as that of MetaMap, Open Biomedical Annotator (OBA), and NaCTeM's TerMINE, against two gold standard datasets: one from MedHelp and the other from CureTogether.

**Results** When aggregated according to a corroborative voting policy, Turker responses predict expert responses with an F1 score of 84%. A conditional random field (CRF) trained on 10 000 crowd-labeled MedHelp sentences achieves an F1 score of 78% against the CureTogether gold standard, widely outperforming OBA (47%), TerMINE (43%), and MetaMap (39%). A failure analysis of the CRF suggests that misclassified terms are likely to be either generic or rare.

**Conclusions** Our results show that combining statistical models sensitive to sentence-level context with crowd-labeled data is a scalable and effective technique for automatically identifying medical terms in PAT.

## OBJECTIVE

As people rely increasingly on the internet as a source of medical knowledge, online health communities, along with the volume of potentially valuable patient-authored text (PAT) they contain, are growing. This shift is attributed mostly to changes in the healthcare system (including decreased access to healthcare professionals and higher costs of healthcare) and increased technological literacy in the patient population.[1] While PAT may not contain scientifically accurate or systematic data, it comprises rich descriptions of hundreds of patients' experiences over a wide range of conditions, in real time. Already, projects such as Google Flu[2] and HealthMap[3] have shown that PAT is a reliable data source for tracking disease trends; moreover, novel insights into co-morbidities and drug-treatment effects have been discovered on sites like CureTogether[4] and PatientsLikeMe.[5] In these cases, however, the supporting data were curated: attempts to mine large, organic PAT corpora for medical insights have been noticeably limited. We believe this is due, in part, to the lack of an effective method for extracting medical terms from PAT.

Identifying medical concepts in text is a long-standing research challenge that has spurred the development of several software toolkits.[6] Toolkits like MetaMap and the Open Biomedical Annotator (OBA) focus primarily on mapping words from text authored by *medical experts* to concepts in biomedical ontologies. Despite recent efforts to develop an ontology suitable for PAT—the open and collaborative Consumer Health Vocabulary (OAC) CHV[7–9]—we suspect that these tools will remain ill-suited to the task due to structural differences between PAT and text authored by medical experts. Such differences include lexical and semantic mismatches,[10 11] mismatches in consumers' and experts' understanding of medical concepts,[10 12] and mismatches in descriptive richness and length.[10–12] Consider, for example, the text snippets below, both discussing the predictive value of a family history of breast cancer. The first snippet is from a medical study by De Bock *et al*[13]:

> In our study, at least two cases of female breast cancer in first-degree relatives, or having at least one case of breast cancer in a woman younger than 40 years in a first or second-degree relative were associated with early onset of breast cancer.

The second (unedited) snippet is from the MedHelp Breast Cancer community:

> im 40 yrs old and my mother is a breast cancer surivor. i have had a hard knot about an inch long. the knot is a little movable. the knot has grew a little over the past year and on the edge closest to my underarm. i am scared and dnt want to worry my mom ..

Our goal is to automatically and accurately identify medically relevant terms in PAT. Note that we do not attempt to map terms to ontological concepts; we view this as a separate and complementary task. We make the following contributions:

- ▶ We show that crowdsourcing PAT medical word identification tasks to non-experts achieves results comparable in quality to those given by medical experts—in our case, registered nurses.
- ▶ We present a comparative performance analysis of MetaMap, OBA, TerMINE, and two models—a dictionary and a conditional random field (CRF)—trained on 10 000 crowd-labeled sentences.



Copyright 2013 by American Medical Informatics Association.

Downloaded from jamia.bmj.com on May 7, 2013 - Published by group.bmj.com



- We make our trained CRF classifier, ADEPT (Automatic Detection of Patient Terminology) freely available as a web service from our website (http://vis.stanford.edu/projects/adept). ADEPT is trained on 10 000 crowd-labeled sentences, to our knowledge the largest labeled corpus of its kind.

## BACKGROUND AND SIGNIFICANCE

### Medical term identification

MetaMap, arguably the best-known medical entity extractor, is a highly configurable program that relates words in free text to concepts in the UMLS Metathesaurus.[6 14] MetaMap sports an array of analytic components, including word sense disambiguation, lexical and syntactical analysis, variant generation, and POS tagging. MetaMap has been widely used to process datasets ranging from email to MEDLINE abstracts to clinical records.[6 15 16]

The Open Biomedical Annotator (OBA) is a more recent biomedical concept extraction tool under development at Stanford University. OBA is based on MGREP: a concept recognizer developed at the University of Michigan.[17] Like MetaMap, OBA maps words in free text to ontological concepts; its workflow, however, is significantly simpler, comprising a dictionary-based concept recognition tool and a semantic expansion component, which finds concepts semantically related to those present in the exact text.[17]

A handful of studies compare MetaMap and/or OBA to human annotators. Ruau *et al* evaluated automated MeSH annotations on PRoteomics IDEntification (PRIDE) experiment descriptions against manually assigned MeSH annotations. MetaMap achieved precision and recall scores of 15.66% and 79.44%, while OBA achieved 20.97% and 79.48%, respectively.[18] Pratt and Yetisgen-Yildiz compare MetaMap's annotations to human annotations on 60 MEDLINE titles: they found that MetaMap achieved exact precision and recall scores of 27.7% and 52.8%, and partial precision and recall scores of 55.2% and 93.3%, respectively. They note that several failures result from missing concepts in the UMLS.[19]

In addition to ontological approaches, there are several statistical approaches to medical term identification. NaCTeM's TerMINE is a domain-independent tool that uses statistical scoring to identify technical terms in text corpora.[20] Given a corpus, TerMINE produces a ranked list of candidate terms. In a test on eye pathology medical records, precision was highest for the top 40—as ranked by C-value—terms (~75%) and decreased steadily down the list (~30% overall). Absolute recall was not calculated, due to the time-consuming nature of having experts verify true negative classifications in the test corpus; recall relative to the extracted term list was ~97%.[20]

Takeuchi and Collier use a support vector machine to classify text in MEDLINE abstracts to ontological concepts, achieving an F-score of 74% in 10-fold cross validation.[21] Along a similar vein, several statistical, supervised models achieved F scores in the 70% range for the 2004 BioNLP/NLPBA shared task for identifying five medical terminology types in the GENIA corpus.[22–24]

The general trend of statistical models outperforming MetaMap and OBA on generic input suggests that such methods may be more appropriate for PAT medical word identification tasks. Finally, a significant limitation of the stated prior work is the small size of annotated datasets used for training and evaluation. Our results are based on 2000 expert-labeled and 10 000 crowd-labeled sentences.

### Consumer health vocabularies

A complementary and closely related branch of research to ours is Consumer Health Vocabularies: ontologies that link laymen and UMLS medical terminology.[8 25] Supporting motivations include: narrowing knowledge gaps between consumers and providers,[8 9] coding data for retrieval and analysis,[7] improving the 'readability' of health texts for lay consumers,[26] and coding 'new' concepts that were missing from the UMLS.[27 28] We are currently aware of two consumer health vocabularies: the MedlinePlus Consumer Health Vocabulary, and the open and collaborative Consumer Health Vocabulary—(OAC) CHV—which was included in UMLS as of May 2011.

To date, most research in this area has focused on uncovering new terms to add to the (OAC) CHV. In an analysis of 376 patient-defined symptoms from PatientsLikeMe, Smith and Wicks found that only 43% of unique terms had either exact or synonymous matches in the UMLS; of the exact matches, 93% were contributed by SNOMED CT.[28] In 2007, Zeng *et al* compared several automated approaches for discovering new 'consumer medical terms' from MedlinePlus query logs. Using a logistic regression classifier, they achieved an AUC of 95.5% on all n-grams not present in the UMLS.[9] More recently, Doing-Harris and Zeng proposed a computer-assisted update (CAU) system to crawl PatientsLikeMe, suggesting candidate terms for the (OAC) CHV to human reviewers.[26] By filtering CAU terms by C-value[20] and termhood[9] scores, they were able to achieve a 4 : 1 ratio of valid to invalid terms; however, this also resulted in discarding over 50% of the original valid terms.[26] Given the goals of the CHV movement, our CRF model for PAT medical word identification may prove to be an effective method for generating new candidates terms for the (OAC) CHV.

## MATERIALS AND METHODS

We present two hypotheses. The first is that a non-expert crowd can identify medical terms in PAT as proficiently as experts. The second is that we can use large, crowd-labeled datasets to train classifiers that will outperform existing medical term identification tools.

### Datasets

MedHelp (http://www.medhelp.com) is an online health community designed to aid users in the diagnosis, exploration, and management of personal medical conditions. The site boasts a variety of tools and services, including over 200 condition-specific user communities. Our dataset comprises the entire, anonymized discussion history of MedHelp's forums. The raw dataset contains approximately 1 250 000 discussions. After cleaning and filtering (described below), the dataset comprises approximately 950 000 discussions from 138 forums: a total of 27 230 721 sentences.

CureTogether (http://www.curetogether.com) is an online health community where members share primarily categorical and quantitative data, but also hold short discussions. Our dataset comprises about 3000 user comments from a variety of forums. Both our MedHelp and CureTogether data were acquired through research agreements with the respective institutions.

#### Data preparation

We analyze our data at the sentence level. This promotes a fairer comparison between machine taggers, which break text into independent sentences or phrases before annotating, and human

Downloaded from jamia.bmj.com on May 7, 2013 - Published by group.bmj.com



taggers, who may otherwise transfer context across several sentences. We use Lucene (lucene.apache.org) to tokenize the text into sentences. For consistency, we exclude sentences from MedHelp forums that the researchers agreed were tangentially medical (eg, 'Relationships'), over-general (eg, 'General Health'), or that contain fewer than 1000 sentences.

We randomly sample 10 000 sentences from the MedHelp dataset to use as a training corpus, and 1000 additional sentences to use as a gold standard. Finally, we sample 1000 sentences from the CureTogether comment database as an addition gold standard independent of MedHelp.

### Metrics

We evaluate our results using five metrics: F1 score, precision, recall, accuracy, and Matthews Correlation Coefficient (MCC). Our goal is to maximize classifier performance on F1 score. *F1 score* is the harmonic mean of precision and recall; a high F1 score implies that precision and recall are both high and balanced. *Precision* (positive predictive value) measures the proportion of model predictions that are correct. *Recall* (specificity) measures the proportion of correct instances that were predicted. *Accuracy* measures the fraction of correct predictions overall. Accuracy can be misleading, as the medical to non-medical term ratio in the MedHelp corpus is approximately 1:4. *MCC* reflects the correlation between true values and model-predicted values; as it accounts for different class sizes it is a more informative metric than accuracy.

### Hypothesis 1: non-expert crowds can replace experts

Crowdsourcing is the act of allocating a series of small tasks (often called 'micro-tasks') to a 'crowd' of online workers, typically via a web-based marketplace. When the workflow is properly managed (eg, via quality control measures such as aggregate voting), the combined results are often comparable in quality to those obtained via more traditional task completion methods.[29 30] Crowdsourcing is particularly attractive for obtaining results faster and at lower cost than other participant recruitment schemes.

A common barrier to both training and evaluating medical text annotators is the lack of sufficiently large, labeled datasets.[6 19] The challenge in building such datasets lies in sourcing medical experts with enough time to annotate text at a reasonably low cost.[19] Replacing such experts with non-expert crowds would address these concerns and allow us to build labeled datasets quickly and cheaply. To test the viability of replacing experts with non-expert crowds, we construct a PAT medical word identification task comprising 1000 MedHelp sentences.

#### PAT medical word identification task

Amazon's Mechanical Turk (http://www.mturk.com) is an online crowdsourcing platform where workers (Turkers) can browse 'human intelligence tasks' (or HITs) posted by requesters and complete them for a small payment. We ran several pilot studies with Turkers in order to determine a suitable interface and prompt for the task. Originally, we asked users to select all words/phrases relating to medical concepts from the given sentences. This generated several inconsistencies, including:

- *context*: users selected terms that had no medical relevance in the context of the given sentence, but might have medical connotations in other contexts. For example, 'I apologize if my post created any undue **anxiety**';
- *numerical measurements*: users inconsistently extracted numbers, units of measurement, dosages, or some combination of these;
- *concept granularity*: in a sentence like 'I have low blood sugar', users would not know whether to select 'low blood sugar' or just 'blood sugar'.

After several iterations, we arrived at a prompt (see figure 1) that produced consistent results. We discovered that asking users to tag words/phrases that they thought *doctors* would find interesting mitigated context and concept granularity inconsistencies. We also verified that 100 sentences is a reasonably sized task for most users to complete in one sitting.

#### Experiment design

We uniformly sampled 1000 sentences from our MedHelp dataset, deeming 1000 sufficiently large for an informative comparison between Nurse and Turker responses, but small enough to make expert annotation affordable. Per our pilot study observations, we split the sample into 10 groups of 100 sentences.

Our experts comprised 30 registered nurses from ODesk (http://www.odesk.com), an online professional contracting service. In addition to the registered nurse qualification, we required that each expert have perfectly rated English language proficiency. Each expert did one PAT medical word identification task (100 sentences), and each sentence group was tagged by three experts. The experts were reimbursed $5.00 for completing the task. All tasks were completed within 2 weeks at a cost of $150.

Our non-expert crowd comprised 50 Turkers recruited from Amazon's Mechanical Turk (AMT). We required that our Turkers have high English language proficiency, reside in the USA, and be certified to work on potentially explicit content. Each Turker performed a single PAT medical word identification task (100 sentences), and each sentence group was tagged by five Turkers. The Turkers were reimbursed $1.20 on faithful completion of the task. All tasks were completed within 17 hours at a cost of $60.

##### *Turkers versus gold standard*

We determine a gold standard for each sentence by taking a majority vote over the nurses' responses. Voting is performed at the *word* level, despite the prompt to extract words *or* phrases from the sentences. Figure 2 illustrates how this simplifies word identification by eliminating partial matching considerations over multi-word concepts. N-gram terms can be recovered by heuristically combining adjacent words.

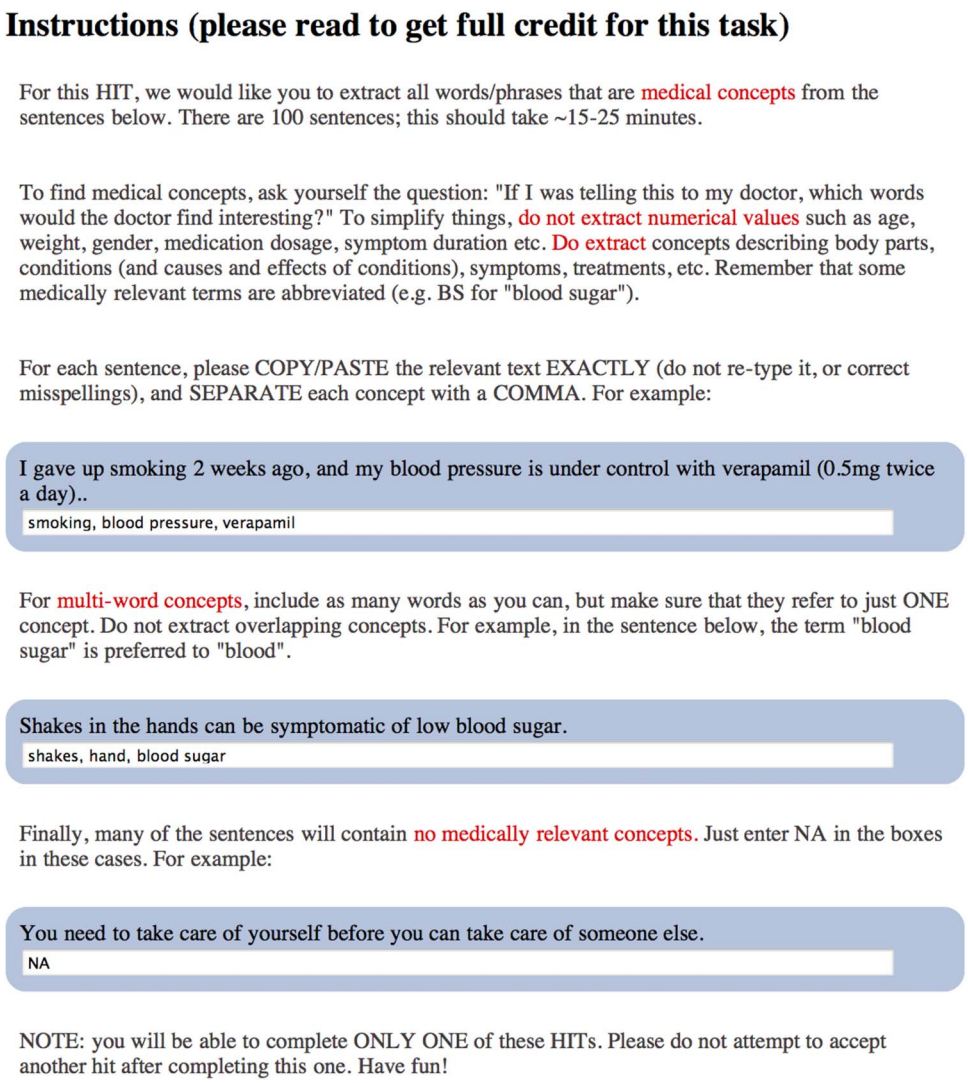
To test the feasibility of using non-expert crowds in place of experts, we compare Turker responses to Nurse responses directly, aggregating across possible Turker voting thresholds. This allows us both to evaluate the quality of aggregated Turker responses against the gold standard and to select the optimal voting threshold.

### Hypothesis 2: classifiers trained on crowd-labeled data perform better

To test our second hypothesis, we create a crowd-labeled dataset comprising 10 000 MedHelp sentences, and an expert-labeled dataset comprising 1000 CureTogether sentences. Using the procedures described above, this cost approximately $600 and $150, respectively. We train two models—a dictionary and a CRF—on the MedHelp dataset, and evaluate their performance via fivefold cross validation; we compare MetaMap, OBA, and TerMINE's output directly. Finally, we compare the performance of all five models against the CureTogether gold standard.

Downloaded from jamia.bmj.com on May 7, 2013 - Published by group.bmj.com





**Instructions (please read to get full credit for this task)**

For this HIT, we would like you to extract all words/phrases that are medical concepts from the sentences below. There are 100 sentences; this should take ~15-25 minutes.

To find medical concepts, ask yourself the question: "If I was telling this to my doctor, which words would the doctor find interesting?" To simplify things, do not extract numerical values such as age, weight, gender, medication dosage, symptom duration etc. Do extract concepts describing body parts, conditions (and causes and effects of conditions), symptoms, treatments, etc. Remember that some medically relevant terms are abbreviated (e.g. BS for "blood sugar").

For each sentence, please COPY/PASTE the relevant text EXACTLY (do not re-type it, or correct misspellings), and SEPARATE each concept with a COMMA. For example:

I gave up smoking 2 weeks ago, and my blood pressure is under control with verapamil (0.5mg twice a day)..

smoking, blood pressure, verapamil

For multi-word concepts, include as many words as you can, but make sure that they refer to just ONE concept. Do not extract overlapping concepts. For example, in the sentence below, the term "blood sugar" is preferred to "blood".

Shakes in the hands can be symptomatic of low blood sugar.

shakes, hand, blood sugar

Finally, many of the sentences will contain no medically relevant concepts. Just enter NA in the boxes in these cases. For example:

You need to take care of yourself before you can take care of someone else.

NA

NOTE: you will be able to complete ONLY ONE of these HITs. Please do not attempt to accept another hit after completing this one. Have fun!

**Figure 1** Patient-authored text (PAT) medical word identification task instructions and interface.

### MetaMap, OBA, and TerMINE

We used the Java API for MetaMap 2012 (metamap.nlm.nih.gov), running it under three conditions: default; restricting the target ontology to SNOMED CT, as a high percentage of 'consumer health vocabulary' is reputedly contained in SNOMED CT[28]; and restricting the target ontology to the (OAC) CHV.

We used the Java client for OBA,[17] running it under two conditions: default; and restricting the target ontology to SNOMED CT (the OAC (CHV) was not available to the OBA at the time of writing).

For TerMINE, we used the online web service (http://www.nactem.ac.uk/software/termine). In all cases, we consider the *words* extracted in the result set, ignoring any particulars of the mappings themselves (illustrated in figure 2).

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *Nurse 1:* | shakes | in | the | hands | can | be | symptompatic | of | low | blood | sugar |
| *Nurse 2:* | shakes | in | the | hands | can | be | symptompatic | of | low | blood | sugar |
| *Nurse 3:* | shakes | in | the | hands | can | be | symptompatic | of | low | blood | sugar |
| *Results:* | shakes | | | hands | | | symptompatic | | | blood | sugar |

**Figure 2** An illustration of our corroborative, word-level voting policy. Stopwords (like 'of') are excluded from the vote.

### Dictionary

A dictionary is one of the simplest classifiers we can build using labeled training data. Our dictionary compiles a vocabulary of *all* words tagged as 'medical' in the training data according to the corroborative voting policy; it then scans the test data, and tags any words that match a vocabulary element. Our dictionary implements case-insensitive, space-normalized matching.

### ADEPT: a CRF model

CRFs are probabilistic graphical models particularly suited to labeling sequence data.[31] Their suitability stems from the fact that they relax several independence assumptions made by Hidden Markov Models; moreover, they can encode arbitrarily related feature sets without having to represent the joint dependency distribution over features.[31] As such, CRFs can incorporate sentence-level context into their inference procedure. Our CRF training procedure takes, as input, labeled training data coupled with a set of feature definitions, and determines model feature weights that maximize the likelihood of the observed annotations. We use the Stanford Named Entity Recognizer package (http://nlp.stanford.edu/software/CRF-NER.shtml), a trainable, Java implementation of a CRF classifier, and its default feature set. Examples of default features include word

Downloaded from jamia.bmj.com on May 7, 2013 - Published by group.bmj.com



**Table 1** Turker performance against the Nurse gold standard along Turker voting thresholds

| Turker vote threshold | F1 | Precision | Recall | Accuracy | MCC |
|---|---|---|---|---|---|
| 1 | 78.45 | 67.15 | **94.31** | 93.96 | 0.77 |
| 2 | **84.43** | 82.53 | 86.41 | 96.29 | **0.82** |
| 3 | 83.80 | 91.67 | 77.18 | **96.52** | **0.82** |
| 4 | 76.61 | 95.70 | 63.87 | 95.46 | 0.76 |
| 5 | 59.81 | **97.99** | 43.04 | 93.26 | 0.62 |

A corroborative vote of 2 or more yields high scores across the board, and maximizes F1 score.

substrings (eg, 'ology' from 'biology') and windows (previous and trailing words); the full list is detailed in online supplementary Appendix A. We refer to our trained CRF model as ADEPT (Automatic Detection of Patient Terminology).

## RESULTS

### Replacing experts with crowds

Both the Nurse and the Turker groups achieve high inter-rater reliability scores: 0.709 and 0.707, respectively, on the Fleiss κ measure. Table 1 compares aggregated Turker responses against the MedHelp gold standard; voting thresholds dictate the number of Turker votes required for a word to be tagged as 'medical'. F1 score is maximized at a voting threshold of 2. We call this a *corroborated vote*, and select 2 as the appropriate threshold for our remaining experiments. Overall, Turker scores are sufficiently high that we regard corroborated Turker responses as an acceptable approximation for expert judgment.

### Classifiers trained on crowd-labeled data

Table 2 shows the performance of MetaMap, OBA, TerMINE, the dictionary model, and ADEPT on the 10 000 sentence crowd-labeled corpus, as well as against both gold standard datasets. ADEPT achieves the maximum score in every metric, bar recall. Moreover, its high performance carries over onto the CureTogether test corpus, suggesting adequate generalization from the training data. Figure 3 provides illustrative examples of ADEPT's performance on sample sentences from the MedHelp gold standard.

To verify the statistical significance of these results, for each annotator we bootstrap 1000 sets of 1000 F1 scores sampled with replacement from each gold standard dataset. We then apply a paired t-test to each annotator pair. All annotator F1 scores were significantly distinct from one another, with $p \leq 0.001$, for both the MedHelp and the CureTogether gold standards (figure 4).

### ADEPT failure analysis

While ADEPT's results are promising, it is also important to assess failure cases. Figure 5 plots term classification accuracy against logged term frequency in both test corpora. We observe that while most terms are classified correctly all of the time, a number of terms (~650) are never classified correctly; of these, almost all (>90%) appear only once in the test corpora.

A LOWESS fit to the points representing terms that were *misclassified at least once* shows that classification accuracy increases with term frequency in the test corpora (and by logical extension, term frequency in the training corpus). As we might expect, over half (~51%) of the misclassified terms occur with frequency one in the test corpora. A review of these terms reveals no obvious term type (or set of term types) likely to be incorrectly classified. Indeed, many are typical words with conceivable medical relevance (eg, *gout*, *aggravates*, *irritated*). Such misclassifications would likely improve with more training data, which would allow ADEPT to learn new terms and patterns.

What remains is to investigate terms that are both frequent and frequently misclassified. Table 3 gives examples of terms that occur more than once in the test corpora and are misclassified more often than not. Immediately obvious is the presence of terms that are medical but generic, for example *doctor*, *doctors*, *drs*, *physician*, *nurse*, *appointment*, *condition*, *health*, etc. These misclassifications likely stem from ambivalence in the training data. If so, either specific instructions to human annotators on how to handle generic terms, or rule-based post processing of annotations, could improve classifier performance.

## DISCUSSION

We explored two hypotheses in this work. The first was that we can reliably replace experts with non-expert crowds for PAT medical word identification tasks. Both Nurses and Turkers achieved high inter-rater reliability scores in the task. We

**Table 2** Annotator performance against the crowd-labeled dataset and the gold standards

| Validation dataset | Annotator | F1 | Precision | Recall | Accuracy | MCC | Parameters |
|---|---|---|---|---|---|---|---|
| MedHelp, Crowd-labeled *10 000 sentences* | MetaMap | 32.64 | 21.88 | 64.20 | 70.44 | 0.24 | Default |
| | | 34.97 | 25.45 | 55.85 | 76.83 | 0.26 | SNOMED CT |
| | | 34.88 | 24.48 | 60.63 | 74.75 | 0.26 | CHV |
| | OBA | 43.77 | 30.20 | 79.53 | 77.21 | 0.39 | Default |
| | | 43.23 | 36.15 | 53.76 | 84.25 | 0.35 | SNOMED CT |
| | Dictionary | 46.18 | 32.34 | **80.75** | 79.02 | 0.42 | |
| | ADEPT | **78.41** | **82.66** | 74.59 | **95.42** | **0.76** | |
| MedHelp, Gold Standard *1000 sentences* | MetaMap | 37.73 | 28.03 | 57.67 | 77.82 | 0.29 | SNOMED CT |
| | OBA | 45.78 | 32.10 | **79.31** | 78.04 | 0.41 | SNOMED CT |
| | TerMine | 42.35 | 52.67 | 35.41 | 88.77 | 0.37 | |
| | Dictionary | 37.30 | 26.34 | 63.89 | 74.98 | 0.29 | |
| | ADEPT | **78.33** | **82.55** | 74.53 | **95.20** | **0.76** | |
| CureTogether, Gold Standard *1000 sentences* | MetaMap | 39.12 | 29.33 | 58.57 | 74.13 | 0.27 | SNOMED CT |
| | OBA | 47.28 | 33.56 | **79.91** | 74.74 | 0.40 | SNOMED CT |
| | TerMine | 43.09 | 53.11 | 36.25 | 86.43 | 0.37 | |
| | Dictionary | 38.74 | 27.53 | 65.35 | 70.65 | 0.27 | |
| | ADEPT | **77.74** | **78.82** | 76.69 | **93.78** | **0.74** | |

Downloaded from jamia.bmj.com on May 7, 2013 - Published by group.bmj.com



| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *ADEPT:* | it | says | proliferative | ductal | hyperplasia | without | atypia | and | non-proliferative | duct | ecstasia | without | carcinoma |
| *Dictionary:* | it | says | proliferative | ductal | hyperplasia | without | atypia | and | non-proliferative | duct | ecstasia | without | carcinoma |
| *MetaMap:* | it | says | proliferative | ductal | hyperplasia | without | atypia | and | non-proliferative | duct | ecstasia | without | carcinoma |
| *OBA:* | it | says | proliferative | ductal | hyperplasia | without | atypia | and | non-proliferative | duct | ecstasia | without | carcinoma |
| *TerMINE:* | it | says | proliferative | ductal | hyperplasia | without | atypia | and | non-proliferative | duct | ecstasia | without | carcinoma |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *ADEPT:* | last | summer | i | was | at | home | with | my | daughter | who | is | now | 2 |
| *Dictionary:* | last | summer | i | was | at | home | with | my | daughter | who | is | now | 2 |
| *MetaMap:* | last | summer | i | was | at | home | with | my | daughter | who | is | now | 2 |
| *OBA:* | last | summer | i | was | at | home | with | my | daughter | who | is | now | 2 |
| *TerMINE:* | last | summer | i | was | at | home | with | my | daughter | who | is | now | 2 |

| | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *ADEPT:* | in | my | case | the | woman | my | husband | had | an | affair | with | reassured | him | twice | she | had | no | stds |
| *Dictionary:* | in | my | case | the | woman | my | husband | had | an | affair | with | reassured | him | twice | she | had | no | stds |
| *MetaMap:* | in | my | case | the | woman | my | husband | had | an | affair | with | reassured | him | twice | she | had | no | stds |
| *OBA:* | in | my | case | the | woman | my | husband | had | an | affair | with | reassured | him | twice | she | had | no | stds |
| *TerMINE:* | in | my | case | the | woman | my | husband | had | an | affair | with | reassured | him | twice | she | had | no | stds |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *ADEPT:* | i | had | a | chest | xray | done | and | they | said | there | was | something | in | my | lung |
| *Dictionary:* | i | had | a | chest | xray | done | and | they | said | there | was | something | in | my | lung |
| *MetaMap:* | i | had | a | chest | xray | done | and | they | said | there | was | something | in | my | lung |
| *OBA:* | i | had | a | chest | xray | done | and | they | said | there | was | something | in | my | lung |
| *TerMINE:* | i | had | a | chest | xray | done | and | they | said | there | was | something | in | my | lung |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *ADEPT:* | mgmt | retail | sales | not | overweight | good | almost | great | posture |
| *Dictionary:* | mgmt | retail | sales | not | overweight | good | almost | great | posture |
| *MetaMap:* | mgmt | retail | sales | not | overweight | good | almost | great | posture |
| *OBA:* | mgmt | retail | sales | not | overweight | good | almost | great | posture |
| *TerMINE:* | mgmt | retail | sales | not | overweight | good | almost | great | posture |

**Figure 3** A comparison of terms identified as medically-relevant (shown in black) by different models in five sample sentences. OBA and MetaMap are run using the SNOMED CT ontology.

attribute the fact that inter-rater reliability is not even higher to inherent task ambiguity.

Combining and aggregating Turker responses predicts Nurse responses with an F1 score of 84%. As crowds of non-experts are much easier to coordinate than medical experts, especially through interfaces like AMT, this opens up new avenues for building large, labeled PAT datasets both quickly and cheaply.

Our second hypothesis was that statistical models trained on large, crowd-labeled PAT datasets would outperform the current state of the art in medical word identification. Our CRF model

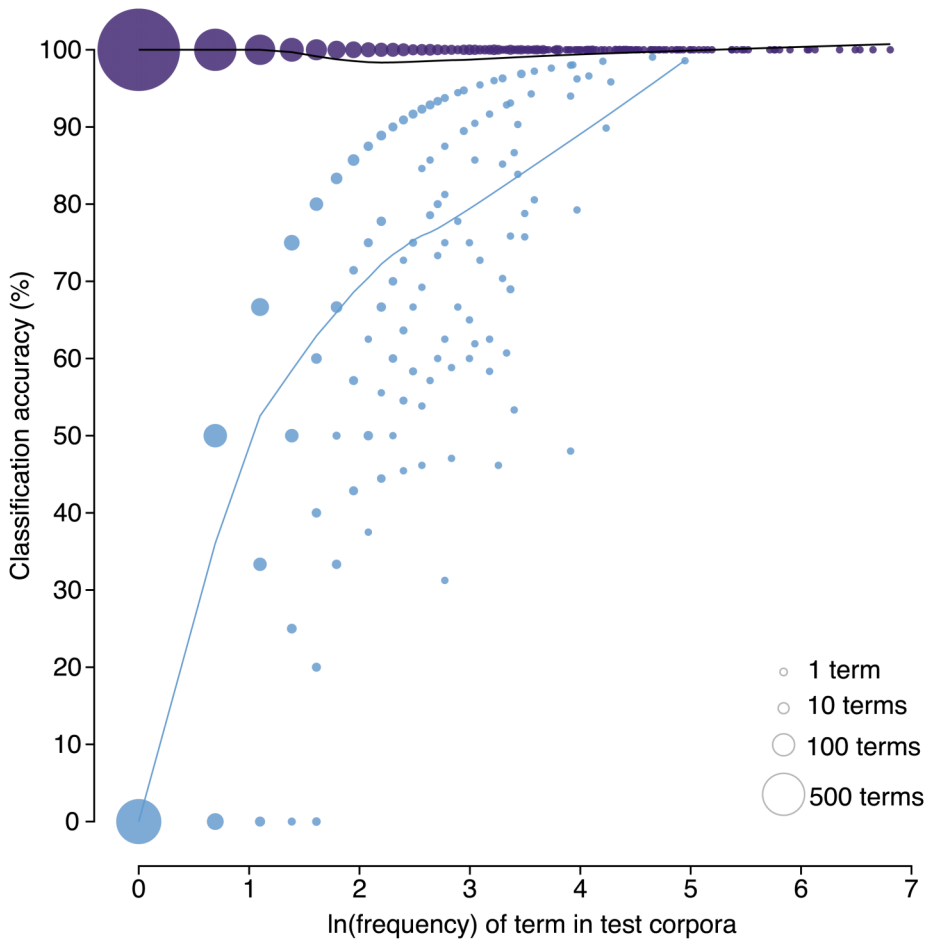



**Figure 4** Term classification accuracy plotted against logged term frequency in test corpora. Purple (darker) circles represent terms that are always classified correctly; blue (lighter) circles represent terms that are misclassified at least once. A LOWESS fit line to the entire dataset (black) shows that most terms are always classified correctly. A LOWESS fit line to the misclassified points (blue, or lighter) shows that classification accuracy increases with term frequency.

Downloaded from jamia.bmj.com on May 7, 2013 - Published by group.bmj.com



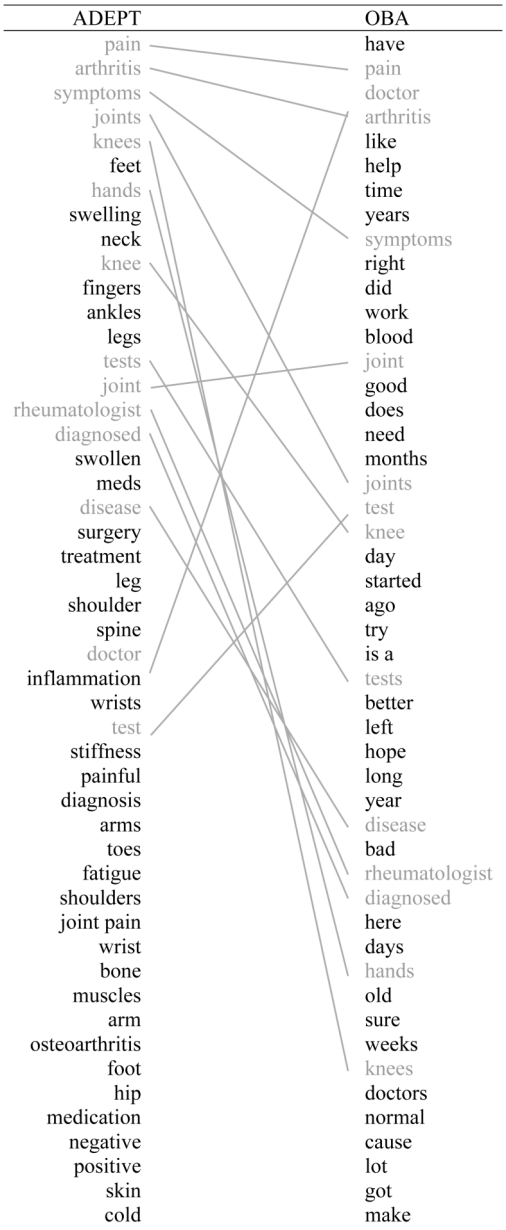



**Figure 5** Top 50 terms, ranked by frequency, derived for MedHelp's Arthritis forum as determined by ADEPT (left) and OBA (right). Terms unique to their respective portion of the list are shown in black. Terms occurring in both lists are linked with a line. The gradient of these lines show that all co-occurring terms, bar three, are ranked more highly by ADEPT.

achieves an F1 score of 78%, dramatically outperforming existing annotation toolkits MetaMap and OBA, and statistical term extractor TerMINE. This performance carries over from cross-validation to validation against an independently sourced PAT gold standard from CureTogether.

We attribute ADEPT's success to the suitability of sentence-level context-sensitive learning models, like CRFs, to PAT medical word identification tasks. Our dictionary, trained on the same data as ADEPT, achieves high recall because it learns many medical terms from training data, but it achieves low precision because it cannot discriminate between relevant and irrelevant invocations of these words. Unlike ADEPT, the dictionary cannot learn, for example, that the word ' sugar' is of particular medical relevance when it co-occurs with the word 'diabetes'.

**Table 3** Examples of terms that occur more than once, and are misclassified more than 50% of the time

| | |
|---|---|
| Frequently misclassified (*FP*>1, *FN*>1) | baby, bc, condition, doctor, doctors, drs, health, ice, natural, relief, short, strain, weight |
| Mostly false positive (*FP*>1, *FN*≤1) | accident, decreased, drinks, drunk, exertion, external, healthy, heavy, higher, lie, lying, milk, million, pants, periods, prevention, solution, suicidal… *[37 more terms]* |
| Mostly false negative *(FP≤1, FN>1)* | appointment, clear, copd, hiccups, lack, ldn, massage, maxalt, missed, nurse, physician, pubic, rebound, silver, sleeping, smell, tea, treat, tree, tx … *[41 more terms]* |
| Infrequently misclassified (*FP*≤1, *FN*≤1) | cravings, generic, growing, hereditary, increasing, lab, limit, lunch, panel, pituitary, position, possibilities, precursor, taste, version, waves, weakness …*[118 more terms]* |

The third sentence in figure 3 suggests that context-based relevance detection may be problematic for MetaMap and OBA, too. In this sentence, the term *case* is annotated because of its membership in SNOMED CT as a medically relevant term pertaining to either a 'situation' or a 'unit of product usage'.

In spite of encouraging results, limitations to this work remain. Most notable is the fact that our technique simply identifies medically relevant terms in PAT: we do not attempt entity resolution or ontology mapping. A related limitation is ADEPT's lack of specificity: we have not trained it to pick out particular types (eg, drugs, body parts) of terms. An adaptation of the framework presented in this paper would likely generate suitable training data for such a task. Finally, ADEPT still fails in some cases. We expect ADEPT's performance to degrade as the corpus diverges from the training corpus in terms of generality and style. As discussed in the failure analysis section, classification accuracy on rare terms would likely be improved through providing additional training data; classification accuracy on frequent terms might be addressed via imposing a specific policy on generic term annotation.

As a final demonstration of the usefulness and efficacy of our method, consider the task of describing a MedHelp forum with its most important constituent medical terms. A natural first attempt would be to rank all relevant terms by their frequency, and select the top *N*. Figure 5 compares the top 50 medical terms in MedHelp's Arthritis forum as determined by ADEPT and the OBA. The terms recovered by ADEPT are both diverse and richly descriptive of arthritic conditions; in contrast, the majority of terms recovered by the OBA are spurious, and serve only to demote the rankings of relevant terms.

## CONCLUSION

We have shown that the combination of crowdsourced training data and statistical models sensitive to sentence-level context results in a powerful, scalable and effective technique for automatically identifying medical words in PAT. We have made our trained CRF model, named ADEPT (Automatic Detection of Patient Terminology), available to the public both for download and as a web service (http://vis.stanford.edu/projects/adept).

**Acknowledgements** The authors thank Atul Butte and Joel Dudley for their feedback on this work.

**Contributors** DLM and JH: conception and design. DLM and JH: data acquisition. DLM and JH: experiment design and execution. DLM and JH: analysis and interpretation of the data. DLM and JH: drafting of manuscript. DLM and JH: critical revision of the paper for important intellectual content. DLM and JH: final approval of the paper.

Downloaded from jamia.bmj.com on May 7, 2013 - Published by group.bmj.com



**Funding** This work was supported by NSF grant number 0964173 and NIH R01 GM079719-07.

**Competing interests** None.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Open Access** This is an Open Access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 3.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/3.0/

## REFERENCES

1 Neal L, Oakley K, Lindgaard G, *et al*. Online Health Communities. *Proc ACM SIGCHI Conference on Human Factors in Computing Systems* 2007, 2129–32.
2 Ginsberg J, Mohebbi MH, Patel RS, *et al*. Detecting influenza epidemics using search engine query data. *Nature* 2008;457:1012–14.
3 Freifeld CC, Mandl KD, Reis BY, *et al*. Healthmap: global infectious disease monitoring through automated classification and visualization of internet media reports. *J Med Internet Res* 2008;15:150–7.
4 Carmichael A. Infertility-Asthma link confirmed. Cure Together Blog. http://curetogether.com/blog/2011/03/07/infertility-asthma-link-confirmed. Updated March 7, 2011 (accessed 12 Jan 2012).
5 Wicks P, Vaughan TE, Massagli MP, *et al*. Accelerated clinical discovery using self-reported patient data collected online and a patient-matching algorithm. *Nat Biotechnol* 2011;29:411–14.
6 Aronson AR. An overview of MetaMap: historical perspective and recent advances. *J Med Internet Res* 2010;17:229–36.
7 Doing-Harris KM, Zeng-Treitler Q. Computer-assisted update of a consumer health vocabulary through mining of social network data. *J Med Internet Res* 2011;13:e37.
8 Zeng QT, Tse T. Exploring and developing consumer health vocabularies. *J Am Med Inform Assoc* 2006;13:24–9.
9 Zeng QT, Tse T, Divita G, *et al*. Term identification methods for consumer health vocabulary development. *J Med Internet Res* 2007;9:e4.
10 Zeng Q, Kogan S, Ash N, *et al*. Characteristics of consumer terminology for health information retrieval. *Meth Inform Med* 2002;41:289–98.
11 McCray AT, Loane RF, Browne AC, *et al*. Terminology issues in user access to web-based medical information. *Proc AMIA Symp* 1999:107–11.
12 Gibbs RD, Gibbs PH, Henrich J. Patient understanding of commonly used medical vocabulary. *J Fam Pract* 1987;25:176–8.
13 De Bock Geertruida JC, Caroline S, Elly KW, *et al*. A family history of breast cancer will not predict female early onset breast cancer in a population-based setting. *BMC Cancer* 2008;8:203.
14 Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Fall Symp* 2001:17–21.
15 Brennan PF, Aronson AR. Towards linking patients and clinical information: detecting UMLS concepts in e-mail. *J Biomed Inform* 2003;36:334–41.
16 Chapmann WW, Fiszman M, Dowling JN, *et al*. Identifying respiratory findings in emergency department reports for biosurveillance using MetaMap. *Medinfo* 2004;11:487–91.
17 Jonquet C, Shah NH, Musen MA. The open biomedical annotator. *Summit on Translat Bioinforma* 2009:56–60.
18 Ruau D, Mbagwu M, Dudley JT, *et al*. Comparison of automated and human assignment of MeSH terms on publicly-available molecular data sets. *J Biomed Inform* 2011;44:S39–43.
19 Pratt W, Yetisgen-Yildiz M. A study of biomedical concept identification: MetaMap vs. people. *Proc AMIA Symp* 2003:529–33.
20 Frantzi K, Ananiadou S, Mima H. Automatic recognition of multi-word terms: the C-value/NC-value method. *Int J Dig Libraries* 2000;3:115–30.
21 Takeuchi K, Collier N. Bio-medical entity extraction using support vector machines. *Artif Intelligence Med* 2005;33:125–37.
22 Finkel JR, Dingare S, Nguyen N, *et al*. Exploiting context for biomedical entity recognition: from syntax to the web. *Proc Intl Joint Workshop on NLP in Biomedicine and its Applications* 2004:88–91.
23 GuoDong Z, Jian S. Exploring deep knowledge resources in biomedical name recognition. *Proc Intl Joint Workshop on NLP in Biomedicine and its Applications* 2004:96–9.
24 Settles B. Biomedical named entity recognition using conditional random fields and rich feature sets. *Proc Intl Joint Workshop on NLP in Biomedicine and its Applications* 2004;104–7.
25 Fernandez-Luque L, Karlsen R, Bonander J. Review of extracting information from the social web for health personalization. *J Med Internet Res* 2011;13:e15.
26 Keselman A, Tse T, Crowell J, *et al*. Assessing consumer health vocabulary familiarity: an exploratory study. *J Med Internet Res* 2007;9:e5.
27 Keselman A, Smith CA, Divita G, *et al*. Consumer health concepts that do not map to the UMLS: where do they fit? *J Am Med Inform Assoc* 2008;15:496–505.
28 Smith CA, Wicks PJ. PatientsLikeMe: consumer health vocabulary as a folksonomy. *Proc AMIA Symp* 2008:682–6.
29 Kittur A, Chi EH, Suh B. Crowdsourcing user studies with Mechanical Turk. *Proc ACM SIGCHI Conf on Human Factors in Computing Systems* 2009:453–6.
30 Bostock J M. Crowdsourcing graphical perception: using Mechanical Turk to assess visualization design. *Proc ACM SIGCHI Conf on Human Factors in Computing Systems* 2010:203–12.
31 Lafferty JD, McCallum A, Pereira F. Conditional random fields: probabilistic models for segmenting and labeling sequence data. *Proc International Conference on Machine Leaning* 2001: 282–9.

Downloaded from jamia.bmj.com on May 7, 2013 - Published by group.bmj.com

# Identifying medical terms in patient-authored text: a crowdsourcing-based approach

Diana Lynn MacLean and Jeffrey Heer

*J Am Med Inform Assoc* published online May 5, 2013
doi: 10.1136/amiajnl-2012-001110

Updated information and services can be found at:
http://jamia.bmj.com/content/early/2013/05/04/amiajnl-2012-001110.full.html

*These include:*

**Data Supplement** *"Supplementary Data"*
http://jamia.bmj.com/content/suppl/2013/05/05/amiajnl-2012-001110.DC1.html

**References** This article cites 18 articles, 2 of which can be accessed free at:
http://jamia.bmj.com/content/early/2013/05/04/amiajnl-2012-001110.full.html#ref-list-1

**Open Access** This is an open-access article distributed under the terms of the Creative Commons Attribution Non-commercial License, which permits use, distribution, and reproduction in any medium, provided the original work is properly cited, the use is non commercial and is otherwise in compliance with the license. See: http://creativecommons.org/licenses/by-nc/3.0/ and http://creativecommons.org/licenses/by-nc/3.0/legalcode

**P<P** Published online May 5, 2013 in advance of the print journal.

**Email alerting service** Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

Advance online articles have been peer reviewed, accepted for publication, edited and typeset, but have not not yet appeared in the paper journal. Advance online articles are citable and establish publication priority; they are indexed by PubMed from initial publication. Citations to Advance online articles must include the digital object identifier (DOIs) and date of initial publication.

To request permissions go to:
http://group.bmj.com/group/rights-licensing/permissions

To order reprints go to:
http://journals.bmj.com/cgi/reprintform

To subscribe to BMJ go to:
http://group.bmj.com/subscribe/

Downloaded from jamia.bmj.com on May 7, 2013 - Published by group.bmj.com

**Topic Collections**

Articles on similar topics can be found in the following collections

Open access (88 articles)

**Notes**

Advance online articles have been peer reviewed, accepted for publication, edited and typeset, but have not not yet appeared in the paper journal. Advance online articles are citable and establish publication priority; they are indexed by PubMed from initial publication. Citations to Advance online articles must include the digital object identifier (DOIs) and date of initial publication.

To request permissions go to:
**http://group.bmj.com/group/rights-licensing/permissions**

To order reprints go to:
**http://journals.bmj.com/cgi/reprintform**

To subscribe to BMJ go to:
**http://group.bmj.com/subscribe/**