

# Sussex Research

## Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning

Zhuoran Wang, Anoop D Shah, A Rosemary Tate, Spiros Denaxas, John Shawe-Taylor, Harry Hemingway

### Publication date

19-01-2012

### Licence

This work is made available under the **Copyright not evaluated** licence and should only be used in accordance with that licence. For more information on the specific terms, consult the repository record for this item.

### Citation for this work (American Psychological Association 7th edition)

Wang, Z., Shah, A. D., Tate, A. R., Denaxas, S., Shawe-Taylor, J., & Hemingway, H. (2012). *Extracting diagnoses and investigation results from unstructured text in electronic health records by semi-supervised machine learning* (Version 1). University of Sussex. <https://hdl.handle.net/10779/uos.23387396.v1>

### Published in

PLoS ONE

### Link to external publisher version

<https://doi.org/10.1371/journal.pone.0030412>

### Copyright and reuse:

This work was downloaded from Sussex Research Open (SRO). This document is made available in line with publisher policy and may differ from the published version. Please cite the published version where possible. Copyright and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners unless otherwise stated. For more information on this work, SRO or to report an issue, you can contact the repository administrators at [sro@sussex.ac.uk](mailto:sro@sussex.ac.uk). Discover more of the University's research at <https://sussex.figshare.com/>

# Extracting Diagnoses and Investigation Results from Unstructured Text in Electronic Health Records by Semi-Supervised Machine Learning

Zhuoran Wang<sup>1,2</sup>, Anoop D. Shah<sup>3\*</sup>, A. Rosemary Tate<sup>4</sup>, Spiros Denaxas<sup>3</sup>, John Shawe-Taylor<sup>1</sup>, Harry Hemingway<sup>3</sup>

**1** Department of Computer Science, University College London, London, United Kingdom, **2** School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh, United Kingdom, **3** Clinical Epidemiology Group, Department of Epidemiology and Public Health, University College London, London, United Kingdom, **4** Department of Informatics, University of Sussex, Brighton, United Kingdom

## Abstract

**Background:** Electronic health records are invaluable for medical research, but much of the information is recorded as unstructured free text which is time-consuming to review manually.

**Aim:** To develop an algorithm to identify relevant free texts automatically based on labelled examples.

**Methods:** We developed a novel machine learning algorithm, the ‘Semi-supervised Set Covering Machine’ (S3CM), and tested its ability to detect the presence of coronary angiogram results and ovarian cancer diagnoses in free text in the General Practice Research Database. For training the algorithm, we used texts classified as positive and negative according to their associated Read diagnostic codes, rather than by manual annotation. We evaluated the precision (positive predictive value) and recall (sensitivity) of S3CM in classifying unlabelled texts against the gold standard of manual review. We compared the performance of S3CM with the Transductive Vector Support Machine (TVSM), the original fully-supervised Set Covering Machine (SCM) and our ‘Freetext Matching Algorithm’ natural language processor.

**Results:** Only 60% of texts with Read codes for angiogram actually contained angiogram results. However, the S3CM algorithm achieved 87% recall with 64% precision on detecting coronary angiogram results, outperforming the fully-supervised SCM (recall 78%, precision 60%) and TSVM (recall 2%, precision 3%). For ovarian cancer diagnoses, S3CM had higher recall than the other algorithms tested (86%). The Freetext Matching Algorithm had better precision than S3CM (85% versus 74%) but lower recall (62%).

**Conclusions:** Our novel S3CM machine learning algorithm effectively detected free texts in primary care records associated with angiogram results and ovarian cancer diagnoses, after training on pre-classified test sets. It should be easy to adapt to other disease areas as it does not rely on linguistic rules, but needs further testing in other electronic health record datasets.

**Citation:** Wang Z, Shah AD, Tate AR, Denaxas S, Shawe-Taylor J, et al. (2012) Extracting Diagnoses and Investigation Results from Unstructured Text in Electronic Health Records by Semi-Supervised Machine Learning. PLoS ONE 7(1): e30412. doi:10.1371/journal.pone.0030412

**Editor:** Vladimir Brusic, Dana-Farber Cancer Institute, United States of America

**Received:** August 26, 2011; **Accepted:** December 15, 2011; **Published:** January 19, 2012

**Copyright:** © 2012 Wang et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** The research leading to these results has received funding from the Wellcome Trust (086091/Z/08/Z; <http://www.wellcome.ac.uk/>) and the National Institute of Health Research (RP-PG-0407-10314; <http://www.nihr.ac.uk/>) under the project CALIBER (Cardiovascular Disease Research Using Linked Bespoke Studies). This work was supported in part by the IST Programme of the European Community, under the PASCAL2 Network of Excellence. Anoop Shah is supported by a Wellcome Trust Clinical Research Training Fellowship (0938/30/Z/10/Z). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. No additional external funding was received for this study.

**Competing Interests:** The authors have declared that no competing interests exist.

\* E-mail: [anoop@doctors.org.uk](mailto:anoop@doctors.org.uk)

## Introduction

Although electronic health records are an important source of data for health research, much of the information is stored in an unstructured way and can be difficult to extract. Research to date has predominantly used the coded data because it is readily analysed, but unstructured ‘free’ text in clinical entries may contain important information [1–4]. Manual review of free text is time-consuming and may require anonymisation to protect patient confidentiality. There has therefore been interest in software algorithms to analyse free text; examples include programs to

identify angina diagnoses [2] and acute respiratory infections [4]. Analysis of clinical text is difficult because it can contain a wide range of terminology, complex language structures, context-specific abbreviations, and acronyms. Medical natural language processing systems such as MedLEE [5] rely on a detailed knowledge base and manually programmed linguistic rules. Natural language processors are expensive to develop as they have to be tuned specifically for each task or disease area.

Alternatively, a machine learning approach may be used, in which the computer attempts to ‘learn’ from a collection of training examples and apply this knowledge to classify new texts.

For example, Support Vector Machine (SVM) algorithms have been used for a range of classification tasks based on electronic clinical notes, such as identifying smoking status [6,7] and predicting response to quality of life questionnaires [8]. Hidden Markov Models have been used for paragraph-level topic segmentation and labelling in electronic health records [9,10]. For the task of automatic diagnostic coding, cascade or hybrid systems with machine learning components have been shown to outperform purely rule-based or pattern matching systems [11–13]. The advantage of machine learning approaches is that they do not require manual programming of specific language features or knowledge of the subject area. However their performance can be variable, depending on the particular machine learning algorithm as well as the similarity between the underlying feature distributions in the training and the test sets.

Our aim was to develop a machine learning algorithm to classify whether a free text entry contains information of interest (e.g. a diagnosis or test result). Our novel algorithm, the ‘Semi-supervised Set Covering Machine’ ( $S^3CM$ ) is related to two previous models by Rosales et al. [14]. Firstly they demonstrated a joint framework of semi-supervised active learning based on a Naïve Bayes Network and showed that unlabelled data in addition to the labelled training examples could contribute to the learning process. After this, in a separate work, they introduced an  $L_1$ -regularised SVM-style classifier, which enabled sparse feature representations for the target information to be obtained directly after learning [15].

We tested the  $S^3CM$  algorithm on free text samples from the UK General Practice Research Database (GPRD) which are relevant to our ongoing research studies. GPRD contains anonymised longitudinal medical records from 5 million patients actively registered in 590 contributing primary care centres [16]. It has been widely used for research on drug safety and clinical epidemiology [17]. It contains information on diagnoses, referrals, test results and prescriptions. Diagnoses are coded by general practitioners (GP) using the ‘Read’ coding system [18], and each Read coded entry may contain additional information as free text. This free text can contain clinical notes entered by the GP (e.g. test results, discussion with a patient, referral letters) as well as scanned clinic letters and discharge summaries.

We applied the  $S^3CM$  algorithm to an example of identifying texts containing investigation results (coronary angiograms) and an example of detecting diagnoses (ovarian cancer). Coronary angiograms are performed in hospital but are relevant to the long term management of patients with ischaemic heart disease in primary care. The longitudinal nature of the GPRD record is extremely useful for such studies but the coded record rarely contains angiogram results; only 4.2% of GPRD patients with myocardial infarction have a Read code stating the angiogram result, but a larger proportion have a code stating that an angiogram was performed. It is not possible to obtain angiogram results from hospital records for GPRD patients because they are anonymised to protect confidentiality. However, investigation results may be recorded in the free text in GPRD, either typed by the GP or in scanned letters. The Read codes associated with such texts may be non-specific (e.g. ‘Scanned letter’) so they are difficult to identify by conventional means.

The second case study aimed to detect suspected or definite diagnosis stated in the text prior to the date that it is formally coded. Ovarian cancer is a condition with insidious onset of symptoms, making it difficult to diagnose early, but documentation of suspected cancer may occur in the free text prior to a formal coded diagnosis [1]. This provides insight into the clinical

reasoning of the doctor, and is relevant to research aimed at achieving earlier diagnosis in ovarian cancer.

We compared the performance of  $S^3CM$  against three other algorithms: the original fully-supervised SCM [19], the Transductive Support Vector Machine (TSVM) [20] which is a semi-supervised but non-sparse algorithm, and the Freetext Matching Algorithm (FMA), a natural language processing system we have developed (see Text S1).

## Methods

### Ethics statement

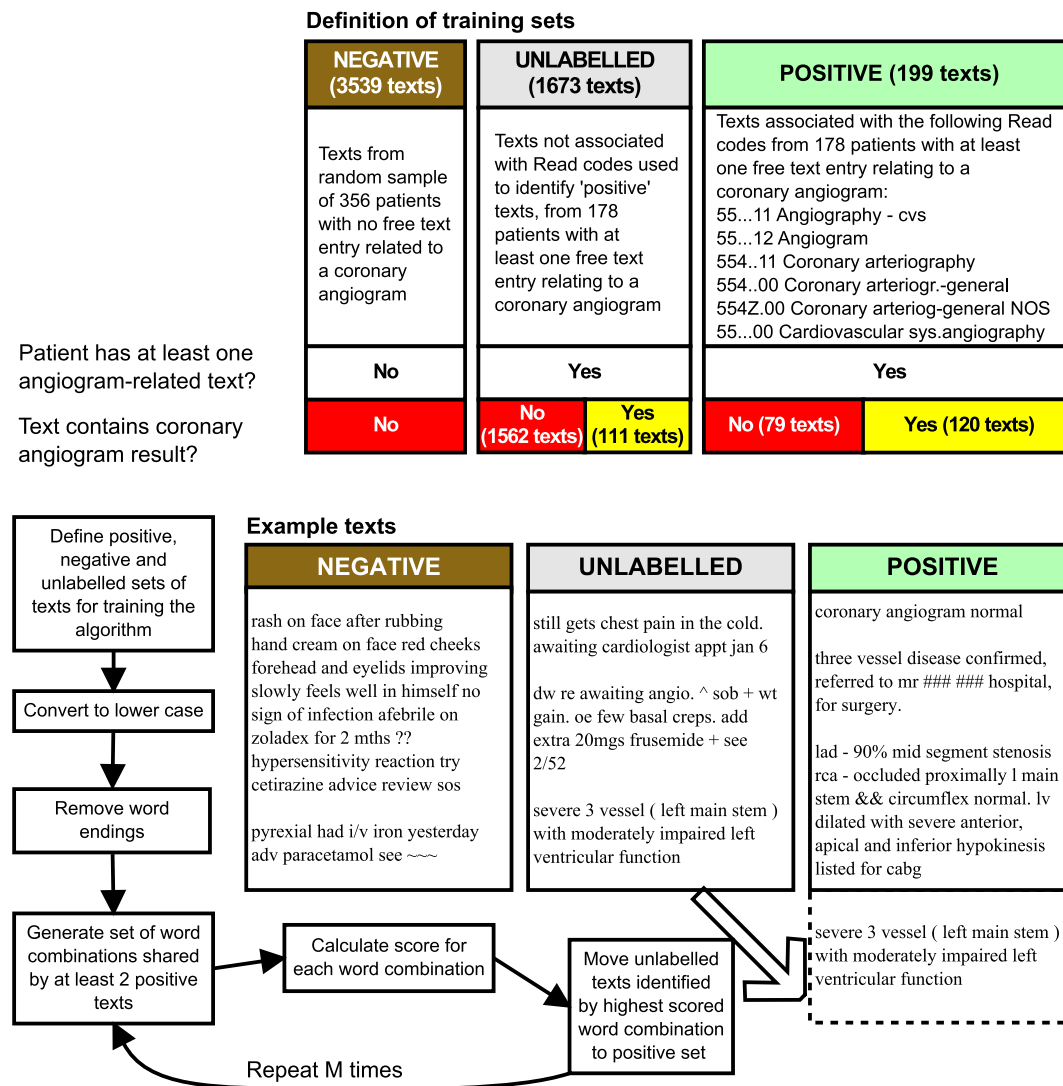
The General Practice Research Database (GPRD) Division of the Medicines and Healthcare products Regulatory Agency has been granted Multi-Centre Research Ethics committee approval for all observational studies using GPRD data. All GPRD study proposals are prospectively reviewed by the GPRD Independent Scientific Advisory Committee, who specifically approved our study (protocols 07\_069 and 09\_123R) and did not require informed patient consent. All data including free text were anonymised by GPRD before being released to researchers.

### Development of machine learning algorithm

We developed a novel machine learning algorithm: the ‘Semi-supervised Set Covering Machine’ ( $S^3CM$ ). This utilised the feature of GPRD data that every free text entry is associated with a Read code. Clinical entries in the GP software are organised into ‘events’ which consist of a Read code denoting the diagnosis or context of the entry, and linked data fields for additional information or free text. GPs encode important diagnoses using Read codes so that they appear in a patient’s summary view and problem list. The text associated with Read codes for diagnoses may contain additional details about the diagnosis (e.g. qualifiers such as severity, or a narrative account), and is presented with the Read term on the doctor’s computer screen. Clinical information may also be entered in free text associated with non-specific Read codes such as ‘Scanned letter’ or ‘History/symptoms’; this can be more difficult to find, often requiring a search of the entire free text.

A set is defined mathematically as a collection of distinct objects in which the order of the objects does not matter. Two sets are considered to be identical if they contain the same objects. In this article we shall refer to sets of words as ‘word combinations’, free text entries associated with Read codes in GPRD as ‘texts’, and a set of texts used for training the algorithm as a ‘training set’. We defined ‘positive’, ‘negative’ and ‘unlabelled’ training sets as follows: the positive training set contained texts associated with the diagnosis of interest (identified by Read codes), the negative training set contained texts not associated with the diagnosis of interest, and the unlabelled set contained texts which the algorithm would try to classify. Figure 1 shows the definition of training sets for the coronary angiogram task, and Figure 2 for the ovarian cancer task. We compared the performance of the  $S^3CM$  with other machine learning algorithms and with our Freetext Matching Algorithm (FMA). FMA uses tables of synonyms and hard-coded semantic information to map words and phrases in free text to Read terms, and assigns attributes for context (e.g. family history, negation or uncertainty). It is described in more detail in Text S1.

The  $S^3CM$  algorithm works by exploring combinations of words which are common to the texts of interest. Case, sentences, word endings and sentence structure are not considered. In the first stage, the algorithm compiles a list of all word combinations shared by at least two positive texts. Each word combination is



**Figure 1. Semi-Supervised Set Covering Machine for detecting coronary angiogram results.** Flow diagram showing logic of the S<sup>3</sup>CM algorithm, and definitions of positive, negative and unlabelled training sets for detection of coronary angiogram results. doi:10.1371/journal.pone.0030412.g001

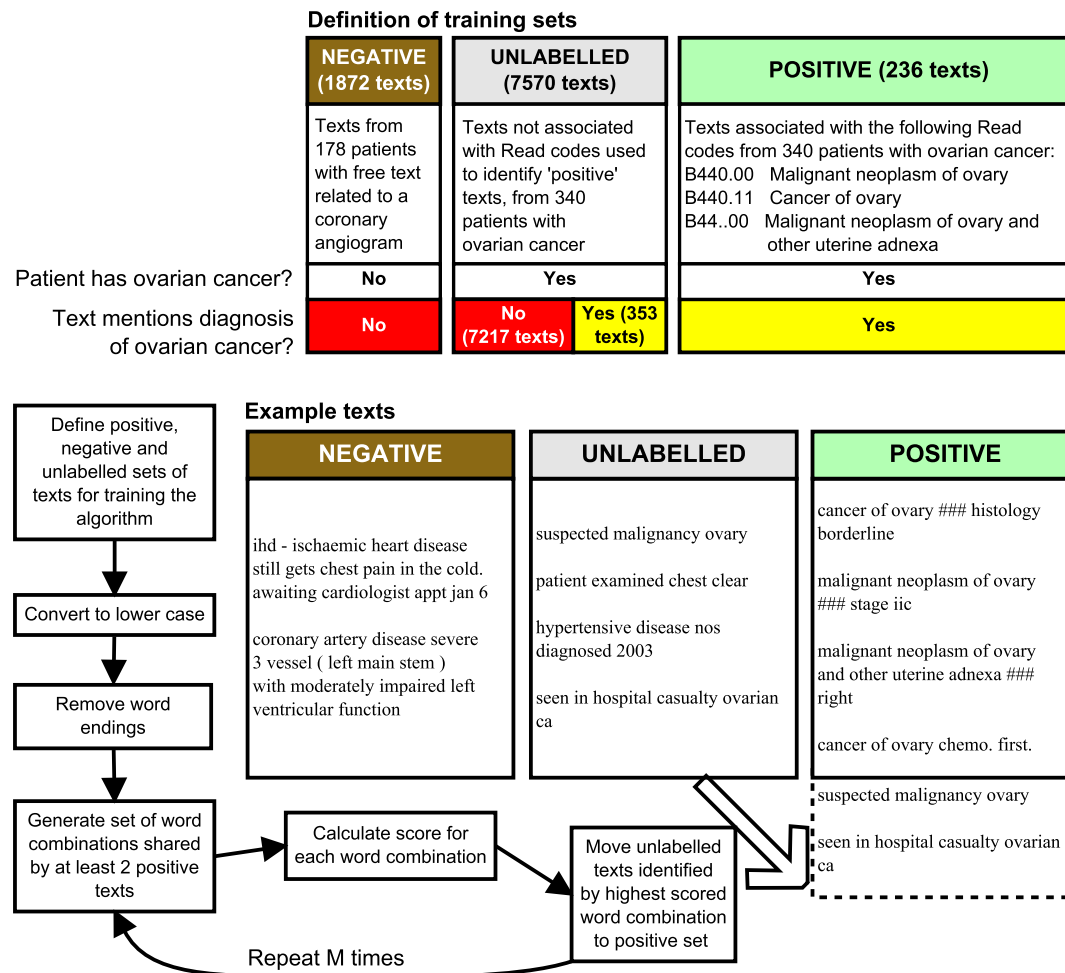
scored on its ability to differentiate positive and negative texts (Figures 1 and 2). The algorithm then enriches its set of word combinations in an iterative manner using the positive and positively classified unlabelled texts. It is a 'semi-supervised' algorithm because it uses unlabelled as well as labelled texts during the training process. The unlabelled texts are used to hone the algorithm by enabling it to find additional word combinations which are associated with the diagnosis of interest, but which may not be included among the original 'positive' texts.

### Detailed technical description of Semi-supervised Set Covering Machine

**Notations and terminology.** We expected that only a small proportion of possible words found in the text would be of use in identifying texts of interest (i.e. the data are sparse) so we chose to use the set covering machine (SCM)[19] as our base algorithm, as it is suitable for sparse data. This algorithm was used in a semi-supervised manner, by training it on labelled and unlabelled texts in a bootstrapping technique.

We denoted each text a data point,  $x$ , and assigned it a label  $y \in \{0,1\}$ . Texts with labels  $y=1$  were called positive texts, and those with  $y=0$  were called negative texts. In the case of semi-supervised learning, there was an unlabelled set with unknown labels  $y$ , which would assist the algorithm during training and be labelled after the training. We used  $\mathcal{P}$ ,  $\mathcal{U}$  and  $\mathcal{N}$  to represent the sets of positive, unlabelled and negative texts respectively. We defined a feature  $h_i$  as a word or word combination (set of words). We expressed each text  $x$  in terms of a feature vector as  $\phi: x \rightarrow [h_1(x), h_2(x), \dots, h_n(x)]$ , where the elements  $h_i(x)$  could have either binary or real values. Given a training set  $\mathcal{S} = \{\mathcal{P}, \mathcal{U}, \mathcal{N}\}$ , the goal of the algorithm was to find a predictive function  $f \in \mathcal{F}_\phi$  such that  $f(x)=y$ . The pseudo-code for these algorithms is given in Figure S1.

**Set Covering Machine (SCM).** The original SCM works in an iterative manner as follows. In each iteration, it greedily selects a feature  $h$  highest-scored by a score function, and removes the examples containing this feature before starting the next iteration, until all prospective (positive or negative) texts have been removed



**Figure 2. Semi-Supervised Set Covering Machine for detecting ovarian cancer diagnoses.** Flow diagram showing logic of the S<sup>3</sup>CM algorithm, and definitions of positive, negative and unlabelled training sets for detection of ovarian cancer diagnoses.  
doi:10.1371/journal.pone.0030412.g002

from the training set, or the size of the learned function  $f$  reaches a predefined value  $K$ . The feature components  $h(x)$  here are binary values, hence the predictive function  $f$  is in the form of a logical conjunction of a set of features. The score function is defined as the number of remaining positive (or negative) examples identified by the algorithm penalized by the number of unexpected examples identified. That is:

$$C(h) : = |\mathcal{P}(h)| - \rho \cdot |\mathcal{N}(h)| \quad (1)$$

where  $\mathcal{P}(h)$  and  $\mathcal{N}(h)$  represent the respective subsets of the positive and negative examples that have feature  $h$ ,  $\rho$  is a weight coefficient, and  $|\cdot|$  denotes the size of a set.

**Modification of SCM for semi-supervised learning.** To adapt the SCM to semi-supervised learning, we first added an additional penalty item to the score function, thus:

$$\tilde{C}(h) : = |\mathcal{P}(h)| - \rho_1 \cdot |\mathcal{U}(h)| - \rho_2 \cdot |\mathcal{N}(h)| \quad (2)$$

where we used the  $\rho_1$ -weighted number of the unlabelled examples that  $h$  identifies ( $|\mathcal{U}(h)|$ ) to give it an extra penalty, since there is a chance of identifying an unlabelled text that could be negative. The feature definition in our task was a combination

of words, so the explicit feature vector of a text  $x$  was the set of all possible word subsets that could be generated from the text.

To avoid dealing with exponentially large explicit vectors, our algorithm was designed as follows. First, it created a set of candidate features from the positive texts by extracting all word combinations shared by at least two texts, thus significantly reducing the feature space. We name this algorithm mSCM for the convenience of future discussion.

We used the algorithm in a semi-supervised manner, with a bootstrapping procedure to gain extra information from the unlabelled examples. In each bootstrap iteration, we moved the unlabelled texts identified by the mSCM in the previous iteration to the positive set, as 'pseudo-positive' texts. We trained a new mSCM based on the updated partitions of the dataset, and repeated this procedure  $M$  times, where  $M$  was a pre-defined number. In each iteration the mSCM compiled the common word combinations among the texts classified as positive, and appended them to the candidate feature set. Thus the algorithm could recall additional features that may not have been included during the initial run (which considered only the labelled positive texts).

The insight behind the bootstrapping procedure was that the unlabelled texts identified by the mSCM in each iteration had the possibility of being positive, and were therefore given a chance to contribute to the score function. Such positive contributions

eliminate the penalty for the remaining unlabelled texts that share common features with them, and increase the possibility of selection of the features shared among them. However, as the pseudo-positive set grows, it increases the chance that the remaining unlabelled examples are identified as positive, and therefore increases the risk of false positives. Therefore we increased the penalty weight for unlabelled texts in each iteration by making it grow linearly with the size of the pseudo-positive set, as shown in Figure S1.

Compared to the work of Rosales et al. [14,15], our  $S^3CM$  has the advantage of synchronously achieving sparse feature representation and contribution of unlabelled data, which were previously realised by two separate models. An advantage of our method compared to semi-supervised active learning is that it can use imperfectly labelled training examples based on diagnostic codes, thus avoiding the need to manually annotate the texts.

**Implementation and complexity analysis.** The algorithm was implemented in C++ and has been tested on Mac OS X and Linux (Ubuntu 11.04). Source code and documentation are available online (<http://sourceforge.net/p/learnehr/home/Home/>).

If we store each text record as a hashtable of words, the time complexity of checking whether a text record contains a word set  $h$  is  $O(|h|)$ . However, the most time-consuming step of the  $S^3CM$  is the procedure for generating common word sets, which is performed in each bootstrapping iteration. Firstly, for two documents  $x_1$  and  $x_2$  the time complexity of finding their largest common subset is  $O(\min(|x_1|, |x_2|))$ . Let  $l$  be the size of the largest common subset obtained. Then the enumeration of all common subsets will require  $\sum_{i=1}^l \binom{l}{i} = 2^l - 1$  unit operations. Although the time complexity is exponential, the running time would still be affordable in practice, as  $l$  tends to be not too big. In practice, one could also restrict the maximum size of common word sets to a threshold  $k$ , which would reduce the time complexity for enumerating all the features to  $O\left(\binom{k}{l}\right)$ .

## Testing

We tested the performance of the  $S^3CM$  in identifying records containing angiogram test results and the diagnosis of ovarian cancer. The 'gold standard' was manual review by a medically qualified researcher who was blinded to the output of the algorithm. The angiogram texts were reviewed by a specialist registrar in internal medicine (AS) and the ovarian cancer texts by a gynaecological oncologist (AM).

We converted the free text examples to lower-case and removed the word endings using the FHC stemmer [21] before applying the  $S^3CM$ . We defined precision (positive predictive value) as the proportion of texts labelled as positive which were true positives, and recall (sensitivity) as the proportion of all positive texts which were correctly labelled as positive by the algorithm.

We compared the performance of  $S^3CM$  with a non-sparse semi-supervised algorithm called the Transductive Support Vector Machine (TSVM) [20], the original fully supervised SCM, and, for the ovarian cancer dataset, our FMA natural language processing system. We did not use FMA on the angiogram dataset because it only detects diagnoses which correspond to Read terms, and there are very few Read terms describing angiogram results. We tuned the parameter settings of the models ( $S^3CM$ , TSVM and SCM) based on the leave-one-out cross-validation (LOO-CV) method. Our adaptation of the LOO-CV method for semi-supervised learning was as follows. We removed the label of one positive text in turn to make it an unlabelled text, trained the algorithm based on this modified data set and tested its classification result for the

pre-selected text. This process was applied to every positive text to obtain an average LOO-CV error rate.

We also evaluated the precision and recall of  $S^3CM$  on classification at the patient level, compared to using Read codes only. We randomly split the case and control patients into a training set and a test set (50 cases and 100 controls for angiogram data; 100 cases and 50 controls for ovarian cancer) and repeated the experiment 10 times. We investigated the timing of the earliest angiogram result or diagnosis of ovarian cancer as detected by the algorithm or Read codes.

## Coronary angiogram dataset

The GPRD Group maintain a library of free text records which have been pooled from previous anonymisation studies. Cases were identified as patients having at least one pre-anonymised freetext record in the library related to a coronary angiogram. Controls were randomly selected from the remaining patients who had at least one entry in the library of pre-anonymised freetext records. Two controls were matched to each case by age within 5 years. The test dataset comprised all pre-anonymised free text entries for the selected patients.

The case data consisted of 2090 free text entries from 178 patients from 122 practices. After removal of blanks and duplicates, 1872 texts remained, of which 199 had a Read code for a coronary angiogram (code list in Figure 1). We reviewed these texts manually and identified 231 records which contained angiogram results, of which 120 were associated with a Read code for angiogram (Table 1). The control data consisted of 3539 records, none of which had a Read code for a coronary angiogram.

Texts associated with Read codes for angiogram ( $n = 199$ ) were taken as positive for the purpose of training the algorithm, whether or not they actually contained angiogram results. Texts from control patients were used as negative examples, and the remaining texts ( $n = 1673$ ) from case patients were taken as unlabelled examples.

We compared the  $S^3CM$  algorithm (with parameter settings:  $\rho_1 = 0.01$ ,  $\rho_2 = 10$ ,  $K = 5$  and  $M = 4$ ) with TSVM (with regularisation coefficient  $\lambda = 1$ , unlabelled data influence parameter  $\lambda' = 0.1$  and positive class fraction of unlabelled data  $r = 0.1$ ), and the fully supervised SCM (with  $\rho = 10$ ).

**Ovarian cancer dataset.** The ovarian cancer dataset was from a study by Tate et al. investigating the dating of diagnosis of ovarian cancer in the GPRD [1]. The case selection criteria have been described previously [22] and are briefly reported here. The target population consisted of women between the ages of 40 and 80 from a random sample of 127 GP practices contributing to GPRD. From this population, we identified women aged 40 to 80 years, who were registered with the practice on 1 June 2002, and who had an incident diagnosis of ovarian cancer between 1 June 2002 and 31 May 2007 (recorded using a Read code in Figure 2). We excluded patients who were registered with the practice for less than 2 years or had a previously recorded Read code for ovarian cancer. We obtained anonymised free text records for all consultations recorded during the 12 months before the date of the earliest Read code indicating a referral for, or suspicion of, ovarian cancer, up to and including the date of definite diagnosis. The initial search yielded 7860 clinical events, from which we excluded blanks and duplicates.

Our test set consisted of 7806 clinical events with non-blank free text entries. The final number of patients was 340 (4 patients met the criteria for inclusion but had no free text recorded). Although all patients had a Read code for ovarian cancer in their electronic patient record, only 236 Read codes (from 234 patients) were associated with non-blank free text and were included in our



**Table 1.** Selection of free text entries for training the algorithm.

	Coronary angiogram dataset	Ovarian cancer dataset
Number of patients	178 patients with at least one text relating to a coronary angiogram	340 patients with new diagnosis of ovarian cancer
Initial number of texts	2090	7860
Number of texts after removal of blanks and duplicates	1872	7806
Text together with Read term for analysis	No	Yes
Number of texts with positive Read code (positive training set)	199 texts with Read code for angiogram	236 texts with Read code for ovarian cancer
Number of texts with positive Read code and positive text on manual review	120 with angiogram results in text and Read code for angiogram	236 (all ovarian cancer Read terms regarded as positive)
Number of unlabelled texts which are positive on manual review	111	353
Number of unlabelled texts which are negative on manual review	1562	7217
Total number of unlabelled texts	1673	7570

doi:10.1371/journal.pone.0030412.t001

sample. We manually reviewed texts containing the fragments ‘ov’, ‘ovar’ or ‘ov.’, and assigned them as ‘positive’ if they stated a suspected or definite diagnosis of ovarian cancer for the current patient. All other texts were assigned as ‘negative’, including those which mentioned ovarian cancer in another context (e.g. negation, family history or patient anxiety). We found 353 texts which referred to ovarian cancer but did not have a Read code for ovarian cancer (Table 1).

We trained the  $S^3CM$  algorithm (with parameters  $\rho_1=0.001$ ,  $\rho_2=10$ ,  $K=5$  and  $M=4$ ) with the following training datasets: texts with Read codes for ovarian cancer ( $n=236$ ) were positive examples, texts without Read codes for ovarian cancer ( $n=7570$ ) were the unlabelled examples, and texts from angiogram case data ( $n=1872$ ) were negative examples (as we did not have access to control data for this study). For this test we appended the free text to the Read term of each record to make it more informative, and appear similar to the way it would be displayed on the GP computer system. We also tested the supervised SCM (with  $\rho=10$ ), TSVM (with  $\lambda=0.01$ ,  $\lambda'=1$  and  $r=0.1$ ) and the Freetext Matching Algorithm. FMA mapped the texts onto Read codes with a context attribute; for this test a Read code in Figure 2 was considered positive as long as it was not associated with an attribute for negation or family history.

## Results

### Coronary angiogram results

Only 60% of texts in the ‘positive’ training set (with read codes for angiogram) actually contained angiogram results in the free text; some contained uninformative text such as ‘hospital admission’. However when tested on unlabelled texts, the  $S^3CM$  algorithm achieved 87% recall with 64% precision. It performed better than the TSVM (precision 3%, recall 2%) and the fully-supervised SCM (precision 60%, recall 78%; see Table 2). The most common word stems associated with positive texts were ‘vessel’, ‘stent’ and ‘lad’ (abbreviation for left anterior descending coronary artery; see Figure 3 A).

In the patient level classification test, we found that the  $S^3CM$  had higher precision than Read codes in identifying patients who had angiogram results (89% versus 71%), but recall was over 90% with both methods. The  $S^3CM$  incorrectly detected angiogram results in 2.7% of control patients (Table 3).

Four patients had angiogram results in the free text earlier than the first angiogram Read code, and 43 patients had angiogram results in the free text but no Read code for angiogram anywhere in their record. Forty of these 47 patients were correctly identified by the algorithm. However, 15 records were incorrectly identified as containing angiogram results, giving precision 73%, recall 85% and F score 79%.

### Ovarian cancer diagnosis

The  $S^3CM$  algorithm performed better than the other machine learning approaches in identifying diagnoses of ovarian cancer in unlabelled texts, detecting 303 of the 353 diagnoses (recall 86%, precision 74%). FMA had greater precision than the  $S^3CM$  (85%) but lower recall (62%; see Table 2). The most common word stem combinations denoting a diagnosis of ovarian cancer were ‘ovari’ with either ‘cancer’, ‘malign’ or ‘carcinoma’ (Figure 3 B).

The algorithm identified 99% of the patients in the test set as having ovarian cancer, even though only 82% of patients had a Read code for ovarian cancer amongst the clinical entries in our dataset (Table 3).

Of the 138 free text records containing a diagnosis of ovarian cancer earlier than the first Read code for ovarian cancer, 123 were correctly identified by the algorithm. However, 81 records were incorrectly identified as denoting an ovarian cancer diagnosis, giving precision 60%, recall 89% and F score 72%.

### Performance

In the unlabelled text classification experiment, running on a Mac computer with an Intel Core i7 2.7 GHz processor and 4GB memory, the  $S^3CM$  took on average 34.3s and 93.6s CPU time in each bootstrapping iteration on the angiogram data and the ovarian cancer data, respectively. Four bootstrapping iterations in total were performed each time to obtain the results in Table 2.

## Discussion

### Summary of main findings

We have developed a novel sparse semi-supervised learning algorithm to classify clinical text records, and have obtained promising results in pilot studies for identification of angiogram results and diagnoses of ovarian cancer in samples of free text from

**Table 2.** Results of testing: classification of unlabelled texts.

Algorithm	Number of texts	True positive	False positive	False negative	Precision, % (95% CI)	Recall, % (95% CI)	F score, %
Presence of coronary angiogram results							
S <sup>3</sup> CM	1673	96	55	15	63.6 (55.3, 71.1)	86.5 (78.4, 92.0)	73.3
SCM	1673	67	19	44	77.9 (67.4, 85.9)	60.4 (50.6, 69.4)	68.0
TSVM	1673	2	64	109	3.0 (0.5, 11.5)	1.8 (0.3, 7.0)	2.3
Ovarian cancer diagnosis							
S <sup>3</sup> CM	7570	303	106	50	74.1 (69.5, 78.2)	85.8 (81.7, 89.2)	79.5
FMA	7570	218	38	134	85.2 (80.1, 89.2)	61.9 (56.6, 67.0)	71.8
SCM	7570	95	53	254	64.2 (55.9, 71.8)	27.2 (22.7, 32.3)	38.2
TSVM	7570	26	534	323	4.6 (3.1, 6.8)	7.4 (5.0, 10.9)	5.7

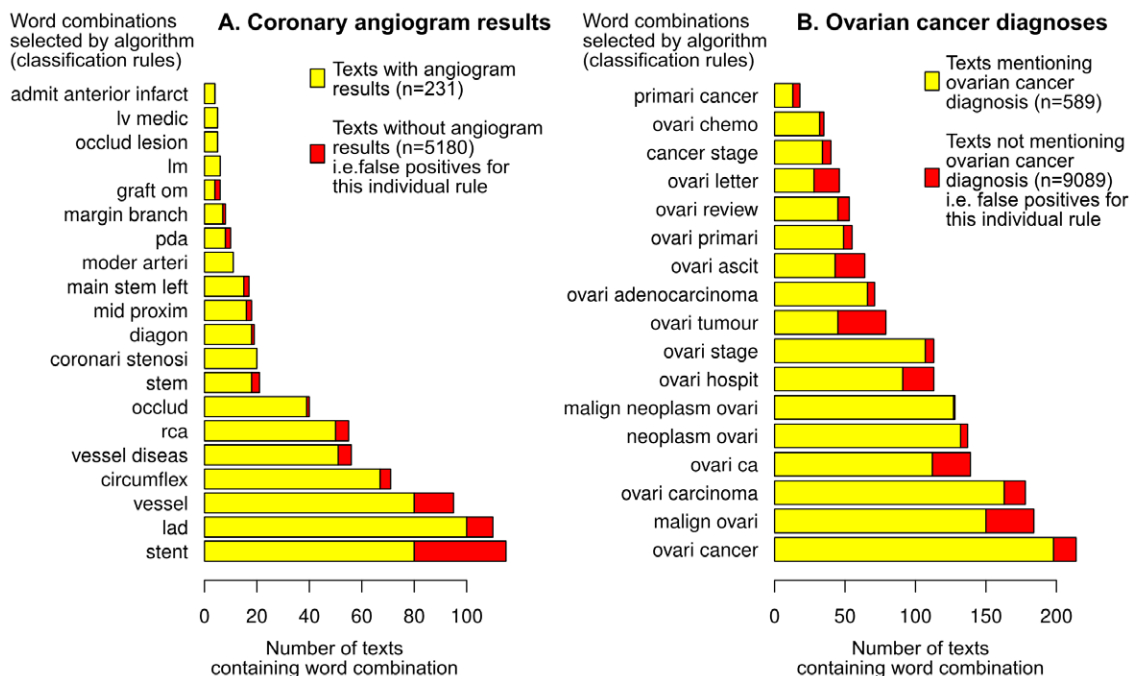
Precision (positive predictive value) is the percentage of texts positively classified by the algorithm that are true positive, and recall (sensitivity) is the percentage of all positive texts correctly classified as positive by the algorithm. F score is the harmonic mean of precision and recall. Figures in parentheses are 95% confidence intervals. doi:10.1371/journal.pone.0030412.t002

the GPRD. The algorithm performed well despite the fact that for the angiogram dataset, the allocation of training examples was imperfect. 'Positive' training examples were denoted by Read codes and not by manual review, and almost 40% of the texts with Read codes for angiogram did not actually contain angiogram results.

A strength of our algorithm is that the training examples can be provided by a diagnostic code search rather than requiring manual review. The algorithm does not rely on a pre-programmed knowledge base or linguistic rule set, and is easy to adapt to other subject areas or languages. It explores the unlabelled data as well as using the positive and negative sets, and compiles a

comprehensive list of word combinations pertaining to the condition of interest, which may be used to feed further research.

The trade-off between recall and precision depends on the task; for example if the algorithm is used to select texts for anonymisation and manual review, good recall is more important than precision. Our Freetext Matching Algorithm achieved better precision than S<sup>3</sup>CM in detecting ovarian cancer diagnoses, but at the cost of only 62% recall. This is because FMA looked for phrases representing diagnoses which could be converted to Read terms, and might miss a diagnosis if the words 'ovary' and 'cancer' were widely separated. However, such texts might be recognised by S<sup>3</sup>CM, which ignores word order.



**Figure 3. Word stem combinations extracted from free text records.** List of word stem combinations selected as classification rules by S<sup>3</sup>CM for (A) coronary angiogram and (B) ovarian cancer test sets. The bars show the frequency of each rule among the combined positive, negative and unlabelled training sets. Words were stemmed in order to aid the grouping of similar words; for example 'ovarian', 'ovary' and 'ovaries' were all converted to the common stem 'ovari'. doi:10.1371/journal.pone.0030412.g003



**Table 3.** Results of testing: detection rate by patient of presence of angiogram results or ovarian cancer diagnosis in the free text.

Method	Precision (%)	Recall (%)	F score	Control error rate (%)
Presence of coronary angiogram results				
S <sup>3</sup> CM	89.3 ± 10.6	93.1 ± 7.5	91.1 ± 7.5	2.7 ± 3.8
Read code	70.5 ± 9.8	95.9 ± 6.3	81.1 ± 6.2	0
Ovarian cancer diagnosis				
S <sup>3</sup> CM	96.4 ± 6.2	98.8 ± 2.6	97.5 ± 3.9	0
Read code	100	82.4 ± 9.7	90.3 ± 5.7	0

Mean ± two standard deviations from 10 experiments testing S<sup>3</sup>CM on classification at the patient level by splitting patients randomly into a training set and a test set.

doi:10.1371/journal.pone.0030412.t003

### Limitations of the S<sup>3</sup>CM algorithm

The main limitation of our algorithm is that it does not use any language knowledge to aid interpretation of texts. As with other machine learning approaches, errors may occur through failure to recall texts containing rare or complex language expressions. Our algorithm attempts classification based only on commonly occurring sets of words, and its precision may be limited by incorrect inclusion of negated phrases. Punctuation, case and the order of words are also ignored; thus it does not utilise all the information that would be available to a human reviewer. Misspellings and abbreviations can also cause errors.

### Limitation of development and testing methodology

Although our testing methodology had strengths – use of two different tasks (detection of diagnosis and detection of a test result) on two different diseases (coronary artery disease and ovarian cancer) – the calculated precision and recall must be used with caution when extrapolating to other datasets. The performance of the algorithm will depend on the disease, the selection procedure for the training datasets, and the size of these datasets. Another limitation is that we only tested the algorithm on data from the GPRD. We recommend that if this algorithm is used for future studies, a sample of the results for each study should be reviewed manually to quantify precision and recall.

A general limitation of using free text is that patients with similar medical histories may have different amounts of information in the free text, influenced by the doctor's documentation habits and whether the GP practice routinely scans all correspondence. Researchers should assess the completeness of recording for a particular study and consider limiting the analysis to practices with more complete recording, or use statistical methods to account for missing data. However this limitation may diminish in the future as information technology becomes more widely adopted.

### Clinical and research application

Our approach may facilitate research using electronic health records where diagnoses or other information of interest (e.g. angiogram results) are recorded in free text rather than in coded form. The algorithm is semi-automatic and therefore cheap to run,

and is fairly sensitive at identifying relevant texts. Although it is not accurate enough for definitive classification, it may be useful for filtering large databases to extract a smaller subset of texts for further analysis.

Although our test sets were from GPRD, this approach can be used on other sources of electronic health information such as discharge letters and electronic hospital notes. S<sup>3</sup>CM is not disease-specific and requires only a small amount of labelled data for training, because it gains additional information from unlabelled data. The only aspect of the algorithm that is language-specific is the 'stemmer' program which standardises word endings prior to analysis. S<sup>3</sup>CM processes sets of words without regard to language features such as grammar or word order, so in principle it should work with many languages, including other Indo-European languages. However, for languages in which long compound words convey a complex meaning it may be necessary to split the words into individual morphemes (the smallest part of a language which has meaning on its own) and allow soft matching of those morphological variants when generating rules in S<sup>3</sup>CM [23].

Future work will involve tuning the algorithm to be able to return more detailed information rather than the merely the absence or presence of a condition. We are working on a system to extract the number of diseased vessels from angiogram reports. We also aim to optimise the code and run it on larger datasets.

Future clinical uses of this algorithm in electronic health record systems may include assisting the coding process and auditing the quality of coding. Such improvements in electronic documentation may benefit the quality of patient care, by ensuring that important clinical information is easily recalled.

### Conclusions

We developed a new algorithm, the Semi-Supervised Set Covering Machine, to identify clinical free text entries of interest. Our preliminary testing found that it worked effectively on free texts in the GPRD associated with two different medical conditions, and it may be of use in future research using electronic health records.

### Supporting Information

**Figure S1** Pseudocode for S3CM algorithm. (PDF)

**Text S1** Description of Freetext Matching Algorithm. (PDF)

### Acknowledgments

GPRD data were obtained under license from the Medicines and Healthcare Products Regulatory Agency, and research protocols were approved by the Independent Scientific Advisory Committee. We would like to thank Julie Sanders for helpful discussions and Alexander Martin for assistance with manual annotation of the ovarian cancer texts.

### Author Contributions

Conceived and designed the experiments: ZW. Performed the experiments: ZW. Analyzed the data: ZW ADS ART. Wrote the paper: ADS. Reviewed and contributed to the manuscript: ZW ART SD JST HH. Obtained anonymised free text from GPRD for testing the algorithm: SD ART. Study supervision: JST HH.

### References

1. Tate AR, Martin AGR, Ali A, Cassell JA (2011) Using free text information to explore how and when GPs code a diagnosis of ovarian cancer: an observational

study using primary care records of patients with ovarian cancer. *BMJ Open* 1: e000025. doi:10.1136/bmjopen-2010-000025.

2. Pakhomov S, Hemingway H, Weston S, Jacobsen S, Rodeheffer R, et al. (2007) Epidemiology of angina pectoris: Role of natural language processing of the medical record. *Am Heart J* 153: 666–673. doi:10.1016/j.ahj.2006.12.022.
3. Pakhomov S, Buntrock J, Chute CG (2005) Prospective recruitment of patients with congestive heart failure using an ad-hoc binary classifier. *J Biomed Inform* 38: 145–153. doi:10.1016/j.jbi.2004.11.016.
4. DeLisle S, South B, Anthony JA, Kalp E, Gundlapalli A, et al. (2010) Combining free text and structured electronic medical record entries to detect acute respiratory infections. *PLoS One* 5: e13377. doi:10.1371/journal.pone.0013377.
5. Friedman C, Shagina L, Lussier Y, Hripesak G (2004) Automated encoding of clinical documents based on natural language processing. *J Am Med Inform Assoc* 11: 392–402. doi:10.1197/jamia.M1552.
6. Savova GK, Ogren PV, Duffy PH, Buntrock JD, Chute CG (2008) Mayo clinic NLP system for patient smoking status identification. *J Am Med Inform Assoc* 15: 25–28. doi:10.1197/jamia.M2437.
7. Clark C, Good K, Jezierny L, Macpherson M, Wilson B, et al. (2008) Identifying smokers with a medical extraction system. *J Am Med Inform Assoc* 15: 36–39. doi:10.1197/jamia.M2442.
8. Pakhomov S, Shah N, Hanson P, Balasubramaniam S, Smith SA, et al. (2008) Automatic quality of life prediction using electronic medical records. *AMIA Annu Symp Proc*. pp 545–549.
9. Ginter F, Suominen H, Pyysalo S, Salakoski T (2009) Combining hidden Markov models and latent semantic analysis for topic segmentation and labeling: Method and clinical application. *Int J Med Inform* 78: 1–6. doi:10.1016/j.ijmedinf.2009.02.003.
10. Li Y, Lipsky Gorman S, Elhadad N (2010) Section classification in clinical notes using supervised hidden Markov model. In: *Proceedings of the 1st ACM International Health Informatics Symposium*. pp 744–750.
11. Aronson AR, Bodenreider O, Demner-Fushman D, Fung KW, Lee VK, et al. (2007) From indexing the biomedical literature to coding clinical text: experience with mti and machine learning approaches. In: *Proceedings of the Workshop on Biological, Translational, and Clinical Language Processing*. pp 105–112.
12. Crammer K, Dredze M, Ganchev K, Talukdar PP, Carroll S (2007) Automatic code assignment to medical text. In: *Proceedings of the Workshop on Biological, Translational, and Clinical Language Processing*. pp 129–136.
13. Suominen H, Ginter F, Pyysalo S, Airola A, Pahikkala T, et al. (2008) Machine learning to automate the assignment of diagnosis codes to free-text radiology reports: a method description. In: *Proceedings of the ICML/UAJ/COLT Workshop on Machine Learning for Health-Care Applications*.
14. Rosales R, Krishnamurthy P, Rao RB (2007) Semi-supervised active learning for modeling medical concepts from free text. In: *Proceedings of the Sixth International Conference on Machine Learning and Applications*. pp 530–536.
15. Rosales R, Farooq F, Krishnapuram B, Yu S, Fung G (2010) Automated identification of medical concepts and assertions in medical text. In: *Proceedings of the American Medical Informatics Association Annual Symposium*. pp 682–686.
16. General Practice Research Database (2011) The General Practice Research Database. URL <http://www.gprd.com/home/>. Accessed 2011 Dec 22.
17. Herrett E, Thomas SL, Schoonen WM, Smeeth L, Hall AJ (2010) Validation and validity of diagnoses in the General Practice Research Database: a systematic review. *Br J Clin Pharmacol* 69: 4–14. doi:10.1111/j.1365-2125.2009.03537.x.
18. NHS Information Centre (2011) The Read Codes. URL <http://www.connectingforhealth.nhs.uk/systemsandservices/data/uktc/readcodes>. Accessed 2011 Dec 22.
19. Marchand M, Shawe-Taylor J (2002) The set covering machine. *J Mach Learn Res* 3: 723–746.
20. Sindhvani V, Keerthi SS (2006) Large scale semi-supervised linear SVMs. In: *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* 477–484. URL <http://vikas.sindhvani.org/svmlin.html>. Accessed 2011 Dec 22.
21. Fung BCM, Wang K, Ester M (2003) Hierarchical document clustering using frequent itemsets. In: *Proceedings of SIAM International Conference on Data Mining*. pp 59–70.
22. Tate AR, Martin AGR, Murray-Thomas T, Anderson SR, Cassell JA (2009) Determining the date of diagnosis - is it a simple matter? The impact of different approaches to dating diagnosis on estimates of delayed care for ovarian cancer in UK primary care. *BMC Med Res Methodol* 9: 42. doi:10.1186/1471-2288-9-42.
23. Schulz S, Honeck M, Hahn U (2002) Biomedical text retrieval in languages with a complex morphology. In: *Proceedings of the ACL Workshop on Natural Language Processing in the Biomedical Domain*. pp 61–68.