Automatically Estimating the Incidence of Symptoms Recorded in GP Free Text Notes

Rob Koeling
Department of Informatics
University of Sussex
Brighton BN1 9QJ, UK
robk@sussex.ac.uk

A. Rosemary Tate
Department of Informatics
University of Sussex
Brighton BN1 9QJ, UK
rosemary@sussex.ac.uk

John A. Carroll
Department of Informatics
University of Sussex
Brighton BN1 9QJ, UK
j.a.carroll@sussex.ac.uk

ABSTRACT

The UK General Practice Research Database (GPRD) is a valuable source of information for health services research. It contains coded data supplemented by free text (physicians' notes and letters). However, due to the difficulty of extracting useful information and the cost of anonymisation, this text is seldom utilised in epidemiological research. We annotated the records of 344 women in the year prior to a diagnosis of ovarian cancer and developed a method for automatically detecting mentions of symptoms in text. We estimated the incidence of five commonly presenting symptoms using: (1) coded symptoms, (2) codes augmented by symptoms automatically extracted from text, and (3) a 'gold standard' dataset of codes and text tagged by three clinically trained annotators. The estimates of incidence of each symptom increased by at least 40% when coded information was enhanced using the manually tagged free text. Our automatic method extracted a significant proportion of this extra information. Our straightforward approach should be extremely useful for medical researchers who wish to validate studies based on codes, or to accurately assess symptoms, using information that can be automatically extracted from unanonymised free text.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*linguistic processing, thesauruses*; J.3 [**Computer Applications**]: Life and Medical Sciences—*health, medical information systems*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*text analysis*

General Terms

Algorithms, Experimentation

Keywords

Information extraction, primary care health records, clinical data, epidemiology

MIXHS'11, October 28, 2011, Glasgow, Scotland, UK.

1. INTRODUCTION

UK primary care databases provide a valuable source of information for research on disease epidemiology, drug safety and adverse drug reactions. Analyses of existing large-scale electronic patient records held in the form of primary care datasets such as the General Practice Research Database [5] have to date almost exclusively exploited coded data. Such data are readily accessible to the classical methods of epidemiological analysis, once the complexities of defining and selecting a patient cohort have been overcome.

However, since clinicians can choose to what extent they assign codes to a consultation, there is often a considerable amount of clinical information represented outside the coded data, in the free text notes. The text often contains important information on the severity of symptoms or on additional symptoms which have not been coded [8, 6]. It has been suggested that it could be beneficial if clinicians were forced to enter this information in a more structured way. However, others have argued (e.g. [22]) that clinicians need a certain amount of flexibility to catch the nuances of patient variability, hence the need for a free text field in health records in addition to the coded information (see the Section 'Characteristics of a Good EHR' in [9] for a discussion of this issue).

In Figure 1 we give an example of a frequently occurring situation within the database. Even though there is detailed information in the free text field, the Read code[1] associated with the record is very generic ('Home visit'). The information within the free text is not readily accessible for epidemiological research.

1.1 Related Work

Automatic extraction of information from GP text notes, which are often ungrammatical and contain ambiguous terms, misspellings and abbreviations, is a very challenging NLP task [20, 29]. A number of research projects have developed algorithms or tools to extract specific categories of information from free text. Most of these projects aim to extract a particular type of information, for example, diagnoses from discharge summaries [15], or the smoking status

[1]Read codes were originally developed in the 1980s and are currently used for coding clinical events in primary care in the United Kingdom. Currently there are more than 100,000 such codes. Each code has an associated textual description – for example 'abdominal pain', 'right iliac fossa pain', 'constipation', 'diabetes' – which are available on GP desktop computer systems (usually in a drop down menu) to help them record an appropriate code during the consultation.

(a) 8CB..00 Home visit
asked for by daughter to see if OGD could be hurried up. Same Sx, no change. ***Bloating***, ***epigastric discomfort***, ***belching*** / burping, ***constipated*** , no ***blood in stool*** / malaena, saw GI Physicians last week, told that she should have OGD. o/e, ***abdo bl***

(b)

R073400	Bloating
R090500	Epigastric pain
19B..11	Belching symptom
19C2.00	Constipated
J681.11	Blood in stool
19A2.00	Abdomen feels bloated

Figure 1: Some records are assigned generic codes by the GP, but nevertheless contain potentially important information in the free text: (a) an example record, consisting of a Read code and its textual description followed by free text; (b) uncoded symptoms and signs in this record (highlighted in bold in the free text).

of patients [25]. Recently, more general medical text processing tools have been developed and made available [4, 24].

Other relevant work is being carried out in the area of Computer Assisted Coding [1, 17]. This work is predominantly industry-driven, aiming to produce tools to assist the task of annotating clinical narratives with standard codes (e.g. for insurance purposes). There is a lot of activity in this area, but since there are commercial interests involved, not all research is published. Jagannathan et al. [7] assess several commercial NLP engines for medication information extraction, giving a good and recent overview of work on this particular topic. There is also some related academic work; for example, Lee et al. [11] investigate methods for encoding clinical datasets with SNOMED CT codes, and Fiszman et al. [3] focus on LOINC codes. Wang et al. [28] describe a more generic approach to knowledge acquisition.

1.2 Aims

One of the difficulties in processing clinical text is the fact that text types vary widely. Systems developed on, for example, radiology reports are not necessarily applicable to clinical notes, and vice versa. Generally available toolboxes are a step in the right direction, but the tools provided often need extensive customisation. In particular, modifying symbolic rules or statistical models of language is generally a task that requires NLP expertise. Also, for adaptation or re-training of models, suitable data is required, but confidentiality issues make it difficult for research groups to share data. As a result, although there has been good progress in applying NLP to clinical data, information extraction systems for clinical data have rarely been applied outside the laboratories in which they were developed [16].

It is clear that many applications require sophisticated approaches which are able to use contextual information and make fine-grained distinctions in order to produce results of high enough quality to enable the resulting data to be used for further research. We are working on methods that learn the relationship between lists of (Read) terms and the way clinicians mention the symptoms related to these terms. We are using the annotated data described in Section 2.1 as a learning set. However, it is often difficult to generalise sophisticated methods in such a way that the method can be used in other contexts (e.g. by non-specialist users). We are aiming at methods that make information within free text records available to clinical researchers who have little or no experience with natural language processing.

As a first step towards this we wanted to explore how far we could get with readily available, well understood techniques. We hypothesised that it may be possible to extract a significant amount of useful information about patients' symptoms using relatively straightforward techniques. To do this, we want to stay as close to a medical researcher's comfort zone as possible. Therefore our starting point is a code list of the sort that they would normally draw up at the beginning of a study, and we investigate how we can exploit the code list with the associated code descriptions in order to find occurrences of these events in the unstructured free text parts of the patients' records.

To test this hypothesis we first established how much extra information was in the text, using a team of annotators. We then developed a string matching algorithm which extracts words, or fragments of words associated with the code for each symptom and assigns a presence/absence classification for each. We tested this algorithm on the anonymised free text records of 344 ovarian cancer patients and investigated how this extra information affects the estimates of incidence of symptoms.

This paper has two main contributions:

- We provide evidence for the amount of information that is expressed in the free text part of primary care health records, and how the availability of this extra information would affect the estimates of incidence of symptoms.

- We show that relatively straightforward techniques can be used to extract a significant portion of this information. This is an important finding, since these techniques could in principle be used by medical researchers with little or no experience with natural language processing.

2. METHODS

This study builds on our previous work [27, 26] which used coded records from the General Practice Research Database of 344 patients between 40 and 80 years of age (inclusive) diagnosed with ovarian cancer between 1 June 2002 and 31 May 2007. This dataset provided the basis for the work presented in this paper. For the current study we obtained the anonymised free text records of all 344 patients for the period 12 months prior to the date of coded diagnosis. Tate et al. [27] give details of the study protocol. We used these records to find all occurrences (either coded or in the text) of the five most commonly presenting symptoms of ovarian cancer: abdominal pain, urogenitary problems, abdominal distension or bloating, constipation, and diarrhoea. Three methods were used to obtain these occurrences: (1) extraction of Read codes for the symptom, (2) automated extraction of symptoms matching on occurrences of the Read code description in the text, and (3) manual review of the text.

2.1 Text Annotation

In order to, on the one hand, determine how much relevant information is actually present in the free text part of the

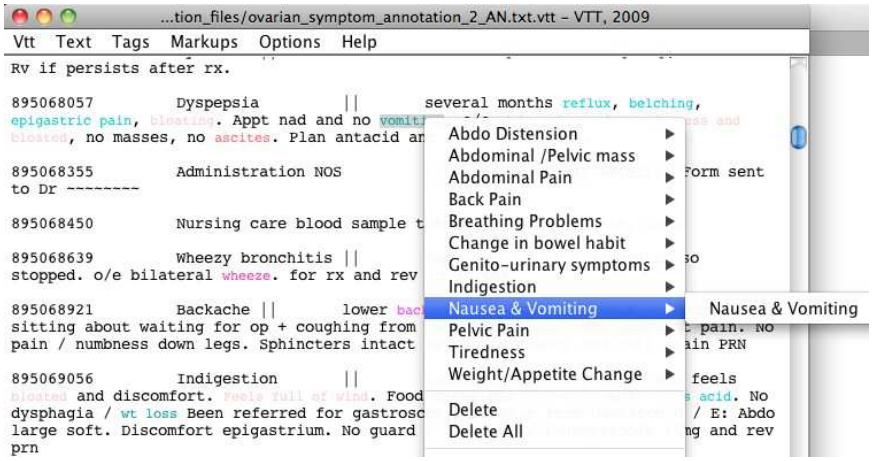

Figure 2: Screenshot of the annotation work bench.

records, and on the other hand, create a dataset that will allow us to learn models that can recognise symptoms in free text, we set up an annotation task to mark up the data. To do this, we recruited three medically trained annotators (one physician and two medical students) and had them read through a random sample of the free text carefully and mark all occurrences of the most commonly presenting symptoms of ovarian cancer listed above.

Setting up the annotation task took a lot of effort, and we went through several iterations in which we refined the annotation guidelines to minimize disagreement between the annotators (learning from others' experience in defining a methodology for annotating clinical data [23]). We used the Visual Tagging Tool (VTT), part of the SPECIALIST NLP Tools [13]. VTT allowed us to create an environment in which the annotators could highlight a phrase in the text and choose the most appropriate tag to describe the symptom they had highlighted (see Figure 2 for a screenshot of the annotation workbench). Each annotator worked independently and annotated all of the text. The inter-annotator agreement was fairly high considering the open nature of the task and the fact that the data was triple-annotated. There was complete three-way agreement (with respect to the start and end position of the phrase and the tag assigned to the phrase) in 62% of the records. The three annotators then met to discuss those cases where there was disagreement. These fell into several categories, ranging from cases where there was a trivial difference (e.g. one of the annotators started the phrase a character earlier then the others), to cases where they agreed on the symptom but not on where the phrase was supposed to start or end, to cases where there was no agreement at all. The latter category was small and all disagreements were resolved in these discussions.

The dataset consists of 6141 records containing approximately 190K words. In [10] we describe the annotation task and the resulting corpus in more detail. Each text fragment was assigned a category relating to one of the five ovarian cancer symptoms. The results were merged with the coded dataset, and for each symptom the percentage of patients recorded as having that symptom at least once in the year prior to diagnosis was calculated using (a) the codes alone, (b) the annotated text, and (c) the merged 'gold standard' dataset containing both codes and text.

2.2 Automatic Recognition of Symptoms

For the first version of the automated symptom recognition algorithm we decided to prioritise *precision* of the algorithm over *recall*. Therefore, while we aimed to maximize the number of found occurrences, we made sure that the number of false positives was kept to a minimum. The matching algorithm uses a list of Read codes corresponding to a set of symptoms. The basic algorithm consists of 3 steps, performed in sequence:

1. Locate an occurrence of textual description of Read code in the text
2. Check whether there is evidence of negation
3. Determine whether the located textual description is within the scope of the negation

In an ideal case, an exact match of the textual description of the Read code is found in the text. However, GPs are not constrained in how they enter text and often use variations of the expression we are looking for. For example, instead of *abdominal pain* the clinician might type *tummy ache*. Furthermore, the nature of these texts complicate processing them. GPs have little time during the consultation to type these notes so there is usually little apparent grammatical structure, many spelling errors, (non-standard) acronyms and abbreviations, and so on. This requires the search to be robust and allow for variations in the surface form of these expressions. We deal with this in two ways. Firstly, we compiled lists of common variations on words used in the Read code descriptions. For example, *abdominal* is often shortened to *abdom*, *abdom.* or *abdo*. For this study we provided these variations manually.

To facilitate this process, we produced lists of potential variations automatically by constructing a 'distributional thesaurus' [14] from a corpus of GP notes that had previously been anonymised by the GPRD for other research projects; we then filtered these lists manually. A distributional thesaurus groups together words that are used in similar contexts. For example this technique identified more than ten variations of the word *patient* in our corpus. These variations consist of common abbreviations and misspellings, including *Pat.*, *pt*, *pateint* and *patien*. Distributional similarity is a powerful technique for inducing information about

words and phrases from raw text, but it does need large amounts of representative data in order to produce accurate results. The amount of data we have available at the moment is limited (millions of words rather then tens of millions of words) and is not a random sample of the database[2]. As a consequence, there is quite a bit of noise in the results returned. We are working on methods to overcome this.

Secondly, we allow for spelling mistakes by matching words that are a small edit distance (we use the Levenshtein distance) from the original word [19, 12]. In these experiments the maximum edit distance allowed is two (i.e. deleting two letters, inserting two letters, or replacing one letter). So *abdominql* is accepted as a variant of *abdominal*, but, for example, *abdoml* is not. The restriction to a maximum edit distance of two means that a considerable number of potential matches are missed, but in preliminary experiments we found that allowing a greater edit distance introduced too many false positives. We are currently working on an approach that allows for a greater edit distance in certain circumstances, by taking the context of the word into account.

2.3 Negation Detection

A mere mention of a symptom is not enough evidence to conclude that the GP is attributing the symptom to the patient. Therefore, after automatically locating potential symptoms in the free text, our method checks whether there is any evidence to the contrary. There are several reasons why a symptom might not be included. For example, we have come across records where the patient was warned that a certain medication might have side effects, or where there was a mention in the text that a person related to the patient suffered from the mentioned symptom. However, by far the most frequent reason for excluding a symptom is negation. Negation has long been recognised as an important issue when mining health record text [18, 2]. Patrick et al. [21] give an overview of approaches to negation identification.

NegEx [2] is a well established tool for identifying negations in medical record text and determining their scope. NegEx produces good results for a variety of different types of text and is customisable to some extent. Unfortunately, it seems to have been developed on more standard English than the data we have. In initial experiments it performed poorly. The problem with our data seems to be mostly that there is very little consistency in marking sentence boundaries – or, more accurately, most text records are not sequences of sentences, but rather sequences of compressed messages. Below are a few examples where NegEx gives incorrect results. (Potentially identifying information is indicated by strings of dashes, and mentions of symptoms are italicised).

(1) flat mood not hungry has had *stress* from gynae probs and daughter flight to ------- discussed reassure no sig wt loss

(2) has constipation no f/h of bowel trouble worried as flying to ----- on saturday no fever has tried mebeverine with no improvement o/e sl *distended abdo* ?? ascites no masses

[2]This data was selected for past research projects at the GPRD and is therefore biased towards certain diseases. The GPRD was able to make this data available to us because it was anonymised. Even though it is not a random sample of the database, it still enables us to learn certain aspects of the type of language used in these notes.

Table 1: Accuracy of NegEx and our negation detection algorithm.

	Negation incorrectly detected	Negation missed
our algorithm	1	2
NegEx	14	3

(3) Urgent, 08. 10hrs came out of hosp yesterday am; started vomiting last night, also S/B --- last night 18.30, BUcastem no good; BO mild *constipation* , nil else;

The first example does not show any evidence of sentence boundaries. In the second example the boundary between two the parts of the record is given by the expression *o/e* ('on examination'). In the third example, messages are separated by semicolons.

We decided to address these problems by simplifying the NegEx approach, and managed to achieve much better accuracy on our type of data. Our algorithm only considers a window of five words preceding the found Read code description. If any evidence for a negating expression is found in this window, it is flagged as a possibly negated term. In order to determine the scope of the negating expression, the algorithm checks whether there is a sentence boundary within this window. If there is evidence for a sentence boundary, the found term is considered to be negated if (and only if) the sentence boundary precedes the negation.

We compared our approach with NegEx by randomly sampling 100 records from our corpus of 6141 records and feeding these both to our negation detection algorithm and NegEx. The outputs were judged by one of the medically trained annotators involved in creating the gold standard. In 12 out of the 100 records there was negation present, with no negation in the remaining 88. The results are summarised in Table 1.

Table 1 confirms the problems we suspected NegEx would have when dealing with our data. Unconventional uses of sentence separators confuse the determination of the scope of a negation. As a result it is often the case that an expression is inaccurately considered to be within the scope of a negation that was found elsewhere in the record. These results say little about NegEx but a lot about our data. There is a small percentage of standard, grammatical text (mostly OCRed letters), and for this it appears that there is room for improvement. We are investigating if we can incorporate a classifier that distinguishes between GP-typed notes and letters. We would then apply our simplified negation algorithm to the former class of text, and NegEx to the latter.

3. RESULTS

3.1 Annotated Dataset

The annotators marked 1669 text fragments as corresponding to one or more of the five symptoms in 944 text records for 285 patients. Of these fragments, 1442 (86%) were classified by the negation algorithm as indicating the presence of the symptom and 227 (14%) its absence. Manual checking of a random sample (of approximately one third of the corpus) of the output of the negation algorithm revealed a misclassification rate of 2% (which is in line with the results of the experiment in which we compared the performance of

Table 2: Number and percentage of patients recorded with each of the five symptoms in the 12 months before diagnosis based on: Read codes alone, symptom mentions in the annotated text, and both codes and text.

Symptom	Codes	Annotated text	Both
Abdominal pain	147 (43%)	159 (46 %)	208 (60%)
Urogenitary problems	87 (25%)	109 (32 %)	140 (41%)
Abdominal distension/bloating	86 (25%)	168 (49 %)	190 (55%)
Constipation	57 (17%)	104 (30%)	127 (37%)
Diarrhoea	28 (8%)	74 (22 %)	87 (25%)

our negation algorithm with NegEx described in the previous section).

There is a significant difference between the percentages obtained from the coded and from the enhanced dataset for each of the five symptoms (Table 2). For example, the detected incidence of abdominal bloating and constipation doubled when information from free text was combined with the codes.

3.2 Keyword Extracted Dataset

The automatic symptom recognition algorithm found 777 matching strings for one or more of the five symptoms in 594 text records for 226 patients. Of the strings matching symptom descriptions, 677 were classified by the algorithm as indicating the presence of the symptom and 100 its absence. Manual checking (by the second author, who was not involved in developing the algorithm) revealed 29 mis-classifications. In 21 cases the algorithm inaccurately detected a positive mention of a symptom and in 8 cases a negated mention of a symptom. This results in a precision score of 96.3% with a recall of 45.6%. Approximately 25% of the mis-classifications were due to the symptom being mentioned as a possible side effect of medication.

The results show a significant difference between the percentages obtained from the coded and enhanced dataset for each of the five symptoms (Table 3), with increases of between 24% and 150%, depending on the type of symptom. Table 4 demonstrates that the automatic algorithm is managing to extract a significant proportion of the relevant symptom information in the text, and when combined with the coded information achieves from 77% to 98% of the upper bound.

4. CONCLUSIONS

We have described a set of experiments into automatic estimation of the incidence of symptoms using coded and free text information in primary care patient records.

For some symptoms, we were able to double the amount of information extracted from the records compared to just considering coded information. In some cases, we were not far off the upper bound for this task (based on manual annotation of the data) using a relatively straightforward natural language processing technique based on matching Read code textual descriptions. The methods are applicable to unanonymised text, and we would expect them also to be applicable to other symptoms and diseases. Our results also show that considering only coded data considerably underestimates the incidence of commonly presenting ovarian cancer symptoms, and we expect that this will be true for many other diseases.

One of the bottlenecks in using text extraction tools outside the labs in which they are developed is the amount of NLP expertise that is required to use them effectively. Our strategy for making these tools more accessible to the medical researcher is to start from code lists, with which they are already familiar. In the experiments reported in this paper we used Read codes, but our approach should be generalizable to any kind of code list where the codes are paired with a term or short description.

In this paper, we have shown that there is a lot of information in primary care free text records that is relatively easy to extract. Next, we intend to validate the generalizability of our approach by applying it to a different disease. Our goal is to develop a tool that can produce these results on the basis of a list of codes given by a medical researcher.

From a natural language processing perspective, GP consultation notes are very challenging. The text is in general very different from standard English, and approaches developed for edited text often show unsatisfactory performance on this kind of data. Even though we have shown that relatively a straightforward string matching technique is able to uncover a lot of information, we expect to obtain better results using more sophisticated methods to improve the recall of the algorithm. In the next phase of this work we are devising techniques for automatically deriving variations of surface realisations of words, overcoming the need to compile lists of common variations, and improving coverage of non-obvious variations. We are doing this by improving the distributional thesaurus construction process. The main problem we have had so far is the limited amount of available anonymised data. The amount of textual data in the GPRD is many times larger then the sample we have. The obvious solution to this problem is to run the thesaurus software on a large random sample of un-anonymised data. The task of removing identifiable information from the resulting thesaurus is much smaller than anonymising the input data, and indeed might not be necessary at all.

In the work described in this paper, we use the outcome of the annotation task described in Section 2.1 to quantify the amount of information 'hidden' in the free text. However, we are currently building a system which also uses this data to learn to recognise surface variations of terms covered by the code lists.

One limitation of this study is that our processing did not distinguish GP notes from letters. In future work we intend to do this. Our definitions of the symptoms were based on the code list for our original study, and may differ from those of other studies.

We believe our straightforward approach for automatic extraction of symptoms will be extremely useful for medical researchers who wish to validate studies based on codes, or who need to get a more accurate picture of the incidence of

Table 3: Number and percentage of patients with each of the five symptoms in the 12 months before diagnosis based on: Read codes plus manually annotated text (Gold standard), Read codes alone (Codes), and Read codes plus automatic text matching.

Symptom	Gold standard	Codes	Codes + text matching
Abdominal pain	208 (60%)	147 (43%)	191 (56%)
Urogenitary problems	140 (41%)	87 (25%)	108 (31%)
Abdominal distension/bloating	190 (55%)	86 (25%)	147 (43%)
Constipation	127 (37%)	57 (17%)	125 (36%)
Diarrhoea	87 (25%)	28 (8%)	72 (21%)

Table 4: Percentages of patients experiencing a symptom at least once in the 12 months preceding diagnosis according to the gold standard, and the percentages of these that are captured using Read codes alone, and Read codes plus automatic text matching.

Symptom	Gold standard	% Captured by codes	% Captured by codes + text matching
Abdominal pain	208 (60%)	70	92
Urogenitary problems	140 (41%)	61	77
Abdominal distension/bloating	177 (51%)	45	83
Constipation	127 (37%)	46	98
Diarrhoea	87 (25%)	35	83

symptoms. Although we would recommend that, for each study the algorithm should first be checked on anonymised data, the low error rate for negation detection may remove the need for manual checking of subsequent data and thus no further pre-anonymisation may be required.

5. ACKNOWLEDGMENTS

We are grateful to Jackie Cassell, Nancy Loades, Amanda Nicholson and Clare Laxton for contributions to the free text annotation aspect of this research. The work was supported by the Wellcome Trust [086105/Z/08/Z]. Access to the GPRD database was funded through the Medical Research Council's licence agreement with MHRA. The authors were independent from the funder and sponsor, who had no role in conduct, analysis or the decision to publish. This study is based in part on data from the Full Feature General Practice Research Database obtained under licence from the UK Medicines and Healthcare Products Regulatory Agency. However, the interpretation and conclusions contained in this study are those of the authors alone.

6. REFERENCES

[1] AHIMA e-HIM Work Group on Computer-Assisted Coding. Delving into computer-assisted coding. *Journal of AHIMA*, 75(10):48A–H, 2004.

[2] W. Chapman, W. Bridewell, P. Hanbury, G. Cooper, and B. Buchanan. Evaluation of negation phrases in narrative clinical reports. In *Proceedings of AMIA Annual Symposium*, pages 105–109, 2001.

[3] M. Fiszman, D. Shin, C. A. Sneiderman, H. Jin, and T. C. Rindflesch. A knowledge intensive approach to mapping clinical narrative to LOINC. In *Proceedings of AMIA Annual Symposium*, pages 227–231, 2010.

[4] C. Friedman, L. Shagina, Y. Lussier, and G. Hripcsak. Automated encoding of clinical documents based on natural language processing. *JAMIA*, 11(5), 2004.

[5] GPRD. Excellence in public health research. http://www.gprd.com, 2009.

[6] W. Hamilton, T. J. Peters, C. Bankhead, and D. Sharp. Risk of ovarian cancer in women with symptoms in primary care: population based case-control study. *British Medical Journal*, 339, 2009.

[7] V. Jagannathan, C. J. Mullett, J. G. Arbogast, K. A. Halbritter, D. Yellapragada, S. Regulapati, and P. Bandaru. Assessment of commercial NLP engines for medication information extraction from dictated clinical notes. *International Journal of Medical Informatics*, 78(4):284–291, 2009.

[8] M. A. Johansen, J. Scholl, P. Hasvold, G. Ellingsen, and J. G. Bellika. "Garbage in, garbage out" – extracting disease surveillance data from EPR systems in primary care. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, pages 525–534, San Diego, CA, 2008.

[9] D. Kalra and D. Ingram. Electronic health records. In K. Zielinski, M. Duplaga, and D. Ingram, editors, *Information Technology Solutions for Healthcare*. Springer-Verlag, http://eprints.ucl.ac.uk/1598/, 2006.

[10] R. Koeling, J. Carroll, A. R. Tate, and A. Nicholson. Annotating a corpus of clinical text records for learning to recognize symptoms automatically. In *Proceedings of the 3rd Louhi Workshop on Text and Data Mining of Health Documents*, pages 43–50, Bled, Slovenia, 2011.

[11] D. H. Lee, F. Y. Lau, and H. Quan. A method for encoding clinical datasets with SNOMED CT. *BMC Medical Informatics and Decision Making*, 10:53, 2010.

[12] V. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.

[13] Lexical Systems Group. http://lexsrv3.nlm.nih.gov/LexSysGroup/Home/, 2010.

[14] D. Lin. Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL*, pages 768–774, Montreal, Canada, 1998.

[15] W. Long. Extracting diagnoses from discharge summaries. In *Proceedings of AMIA Annual Symposium*, pages 470–474, 2005.

[16] S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, and J. F. Hurdle. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of Medical Informatics*, pages 128–144, 2008.

[17] M. Morsch. Computer assisted coding with standard document types – advancing best practice in health information management. In *Proceedings of the AHIMA Convention*. http://www.alifemedical.com/Libraries/PDF/CAC_with_SDT.sflb.ashx, 2009.

[18] P. G. Mutalik, A. Deshpande, and P. M. Nadkarni. Use of general-purpose negation detection to augment concept indexing of medical documents. *JAMIA*, 8(6):598–609, 2001.

[19] G. Navarro. A guided tour to approximate string matching. *ACM Computing Surveys*, 33(1):31–88, 2001.

[20] J. Patrick and P. Asgari. Analysing clinical notes for translational research: back to the future. In V. Prince and M. Roche, editors, *Information Retrieval in Biomedicine: Natural Language Processing for Knowledge Integration*, pages 1954–1975. IGI Global, 2009.

[21] J. Patrick, P. Asgari, and N. Motamedi. Improving accuracy of identifying clinical concepts in noisy unstructured clinical notes using existing internal redundancy. In *Proceedings of the 4th Workshop on Analytics for Noisy Unstructured Text Data*, pages 35–42, Toronto, Canada, 2010.

[22] P. Resnik, M. Niv, M. Nossal, A. Kapit, and R. Toren. Communication of clinically relevant information in electronic health records: a comparison between structured data and unrestricted physician language. *Perspectives in Health Information Management*, 2008.

[23] A. Roberts, R. Gaizauskas, M. Hepple, G. Demetriou, Y. Guo, I. Roberts, and A. Setzer. Building a semantically annotated corpus of clinical texts. *Journal of Biomedical Informatics*, 42(5):950–966, 2009.

[24] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. *JAMIA*, 17(5):507–513, 2010.

[25] G. K. Savova, P. V. Ogren, P. H. Duffy, J. D. Buntrock, and C. G. Chute. Mayo clinic NLP system for patient smoking status identification. *JAMIA*, 15(1):25–28, 2008.

[26] A. R. Tate, A. G. R. Martin, A. Ali, and J. A. Cassell. 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*, doi:10.1136/bmjopen-2010-000025, 2011.

[27] A. R. Tate, A. G. R. Martin, T. Murray-Thomas, S. R. Anderson, and J. A. Cassell. 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 Medical Research Methodology*, 9:42, 2009.

[28] X. Wang, A. Chused, N. Elhadad, C. Friedman, and M. Markatou. Automated knowledge acquisition from clinical narrative reports. In *Proceedings of AMIA Annual Symposium*, pages 783–787, 2008.

[29] H. Xu, S. P. Stenner, S. Doan, K. B. Johnson, L. R. Waitman, and J. C. Denny. MedEx: a medication information extraction system for clinical narratives. *JAMIA*, 17(1):19–24, 2010.