A systematic literature review of automated clinical coding and classification systems

Mary H Stanfill,[1] Margaret Williams,[1] Susan H Fenton,[2] Robert A Jenders,[3] William R Hersh[4]

► Additional tables (1, 2, 7, 8), figures (1—5) and appendices (A, B) are published online only. To view these files please visit the journal online (www.jamia.org).

[1]American Health Information Management Association, Chicago, Illinois, USA
[2]Health Information Management, Texas State University, Texas, USA
[3]Department of Medicine, University of California, Los Angeles, California, USA
[4]Department of Medical Informatics & Clinical Epidemiology, Oregon Health & Science University, Portland, Oregon, USA

Correspondence to
Mary H Stanfill, 224 W 406 N, Valparaiso, IN 46385, USA; mary.stanfill@ahima.org

Received 28 August 2009
Accepted 1 September 2010

ABSTRACT

Clinical coding and classification processes transform natural language descriptions in clinical text into data that can subsequently be used for clinical care, research, and other purposes. This systematic literature review examined studies that evaluated all types of automated coding and classification systems to determine the performance of such systems. Studies indexed in Medline or other relevant databases prior to March 2009 were considered. The 113 studies included in this review show that automated tools exist for a variety of coding and classification purposes, focus on various healthcare specialties, and handle a wide variety of clinical document types. Automated coding and classification systems themselves are not generalizable, nor are the results of the studies evaluating them. Published research shows these systems hold promise, but these data must be considered in context, with performance relative to the complexity of the task and the desired outcome.

INTRODUCTION

Automated coding and classification technologies encompass a variety of computer-based approaches that transform narrative text in clinical records into structured text, which may include assignment of codes from standard terminologies, without human interaction. Despite a great amount of research evaluating systems that perform coding and classification, it is not clear whether these automated systems perform as well as manual coding or classification. We want to know if computer applications can code or classify as well as or better than people. To begin to explore this question, we undertook a systematic literature review to identify and analyze the existing evidence on the performance of automated coding and classification systems.

According to Mulrow and Cook,[1] systematic reviews are concise summaries of the best available evidence that address sharply designed clinical questions. Furthermore, systematic reviews use explicit and rigorous methods to identify, critically appraise, and synthesize relevant studies. They seek to assemble and examine all the available high-quality evidence that bears on the question at hand.

To our knowledge, there are no systematic reviews on automated clinical coding and classification systems. Meystre *et al*[2] conducted a narrative review to examine published research on the extraction of information from textual documents in the electronic health record. In that review, natural language processing techniques were examined, but few of the studies dealt specifically with automated coding and classification software. The authors focused on the performance of information extraction systems, a much broader concept than automated clinical coding and classification. Coding and classification studies that Meystre *et al* reviewed were narrowly focused and did not reflect the full range of automated coding and classification systems. Thus we undertook a systematic literature review to identify all published studies evaluating the performance of automated coding and classification systems. This paper presents the results of our systematic review.

BACKGROUND

Automated coding and classification systems are an emerging technology. Researchers are building and evaluating such systems. It is important to explore what is known concerning the performance of automated coding and classification systems to determine how applicable these systems are to the industry-wide coding process currently used to gather healthcare data.

Correct coding and reporting of healthcare diagnoses and services has become increasingly critical as healthcare data needs have evolved. The use of structured data in coded form continues to grow as the healthcare industry explores value-based purchasing and seeks overall improvement in the quality of care. The data used for these purposes are typically encoded via a manual coding process. This process involves human review of clinical documentation to identify applicable codes. When applying a complex coding scheme, the process may be assisted by the use of code books, picking from abbreviated lists, or employing software applications that facilitate alphabetic searches and provide edits and tips. Code assignment may be carried out by physicians, but is often performed by other personnel, such as coding professionals.

An American Health Information Management Association (AHIMA) workgroup, convened to explore computer-assisted coding, reported that this manual coding workflow is expensive and inefficient in an industry where data needs have never been greater. 'The industry needs automated solutions to allow the coding process to become more productive, efficient, accurate, and consistent.'[3] Computer applications for automating this process are available but currently not widely used, most likely because the systems are still in development and their performance in production unproven. This systematic literature review was undertaken to identify all published studies of

automated coding and classification systems to determine if any system can perform the coding process currently used industry-wide to gather healthcare data. Recognizing that a great deal of research has been carried out in this area, with only a small portion focused on administrative coding classification systems, we determined to review all types of automated coding and classification evaluation studies. As such, this systematic literature review included published research on any computer application designed to automatically generate any type of clinical code or classification from free-text clinical documents.

METHODS

A search strategy was designed to identify all potentially relevant publications about the performance of automated coding and classification systems. It was used to search PubMed, the Cumulative Index to Nursing and Allied Health Literature, the Association for Computing Machinery and Inspec databases, and Science Citation Index Expanded. See appendix A, available as an online data supplement at www.jamia.org, for search parameters and the details of the search statements used in searching the various databases. This review includes all studies published (or pre-published online) and, where applicable, indexed to MeSH terms prior to March 2009.

In addition to searching these databases, all articles in AHIMA's Body of Knowledge indexed to the subject 'computer-assisted coding' were added for consideration. References in the 'FasterCures' report, 'Think Research: Using Medical Records to Bridge Patient Care and Research' were checked for relevance. We also used the 'snowball' method (pursuing references of references) and sought input from a core group of researchers in the field to identify additional studies.

A principal criterion for inclusion in this systematic literature review was that the article had to address the results of an original study involving research on the use of a computer application to automatically generate clinical codes and/or assign classes from free-text clinical documents. In addition, the research had to have been carried out with documents produced in the process of clinical care where both the documents and the computer application were in the English language. The study also must have evaluated the performance of the computer application in assigning clinical codes or other classification schema.

The type of coding or classification schema applied in the study did not affect inclusion. Recognizing the existence of multiple coding and classification schemas, including standardized classification systems, such as the International Classification of Diseases (ICD) or Current Procedural Terminology (CPT), and use-case-specific, non-standardized schemas, such as the presence or absence of a given condition, this review was left open to include any and all types of clinical codes or classes.

Studies were excluded if the automated application was not evaluated for performance of the code assignment. For example, instances where the study focused on evaluating content coverage of a classification or vocabulary were excluded. The difference is subtle, but significant. Evaluating whether a terminology or classification is suitable or robust enough for a given purpose is different from evaluating whether an automated system is accurate enough to replace humans. The latter was aligned with our research question, the former was not. Thus, studies testing the breadth of SNOMED CT, for example,[4–6] were excluded.

Studies were also excluded if no defined coding or classification system was applied. As a result, some information retrieval, information extraction, and/or indexing studies were included and some were not. It can be difficult to discern the difference between indexing and applying clinical codes, since codes are often used for the purpose of indexing or retrieving information. Where indexing was performed using a coding or classification schema—for example, the application of MeSH terms—the study was included. Where indexing involved parsing or indexing documents with no specific code output to evaluate, the study was excluded.

All potentially relevant studies identified were reviewed for inclusion. Each title and abstract retrieved was reviewed by two independent reviewers. When inclusion could not be determined from a title or abstract, the full text of the article was reviewed. When the two initial reviewers reached different conclusions applying inclusion criteria, a third reviewer adjudicated to produce a final decision. Summary information was extracted from all studies satisfying the inclusion criteria.

The systematic literature search yielded 2322 possibly relevant references. There were 2209 articles eliminated as not meeting all of the inclusion criteria, leaving a total of 113 studies for analysis in this systematic literature review. The 113 included studies are listed in online appendix B (available at www.jamia.org). Meta-analysis of these studies was not possible, given the variety of research purposes and study methodologies. Instead, the 113 studies were closely reviewed, and key data elements, such as the following, were abstracted.

- The classification system applied by the automated system and associated healthcare domain (eg, SNOMED for diagnoses on chest radiographs)
- Objective of the study (eg, to determine if an automated system can replace manual chart review to identify cases for a clinical trial)
- The study methodology (including sample size, sample selection, statistical analysis used, and who built the system versus who conducted the evaluation)
- The reference standard for performance
- System performance
- The purpose or use of the automated system
- Conclusions from the study

This abstracted information was examined and key observations are reported here.

RESULTS

The earliest study in the included corpus was published in 1973. Another was published in 1976, and then none until 1990. All but four of the studies (96%) were published after 1994. Online figure 1 (available at www.jamia.org) shows the distribution of the studies over time.

The studies in this review focused on various conditions or healthcare specialties and a wide variety of document types. Online table 1 (available at www.jamia.org) provides details on the conditions and document types specified in the included studies. Pneumonia was the condition most often addressed by these systems, including community-acquired pneumonia, acute bacterial pneumonia, and early detection of pneumonia in neonates. Interestingly, 37 of the studies that specified a particular condition focused on a respiratory condition, which correlates with the most frequently studied documents, chest radiology reports. In general, diagnostic reports were studied more often than other report types, with 54 of the specified document types representing a diagnostic test.

The studies evaluated the performance of various computer applications, many of which were identified by name. Online table 2 (available at www.jamia.org) provides details on these systems. There were 46 different systems named and 21 not named. Of the named systems, Columbia University's MedLEE

was the system studied most often, followed in frequency by SymText, MMTx, and NegEx. These four systems together represent 91% of the named systems studied and 37% of the total corpus.

Study methodologies varied widely across the included corpus. One distinction was the mechanism used to create a reference standard against which the automated systems were evaluated. We found that reference standards fell into one of the following general methodologies.

- Gold standard: multiple, two or more, independent reviewers with adjudication of disagreements to establish consensus in some manner—for example, by majority vote or review/discussion to obtain agreement
- Trained standard: one expert reviewer classifies the majority of the training set, but validity of the reviewer's assignment is verified and training is provided to improve the reviewer's performance/consistency
- Regular practice: one human reviewer, as in the usual manual process; often an existing database reflecting the normal or usual practice was used

Table 3 applies this schema to the included corpus. About 43% of the studies used a gold standard as defined above, a more rigorous, but costly approach. Approximately 51% of the studies compared the automated process with the usual manual process, using regular practice as the standard for comparison.

Statistical methods also varied in system evaluation. Some studies reported simple accuracy rates. A handful of studies utilized more rigorous statistics, such as κ scores, F measures, and receiver operating characteristic curve analysis. Many studies reported more than one measure—for example, sensitivity and specificity, or recall and precision. Table 4 shows the most commonly reported statistics, with the most common measure being recall (or 'sensitivity').

The type of coding or classification scheme applied by the system also varied widely. We found the types of coding fell into two primary groups: (1) those that used an existing classification, vocabulary, or terminology system; (2) those that used a clinical guideline or clinical coding scheme, often developed specifically for the study. A total of 42 studies fell into group 1, with the remaining 71 studies in group 2. Examples of coding classification systems applied by studies in group 1 include:

- CPT
- ICD-8
- ICD-9-CM
- ICF
- UMLS
- MeSH terms
- MedLEE's controlled vocabulary (MED)
- HICDA (Mayo modification of ICD-8)
- RxNorm
- SNOMED (multiple versions: 3.5, RT, III, CT)
- SNOP

The studies in group 2 were subdivided as follows (table 5), reflecting the complexity of the coding and classification scheme applied:

Table 3 Reference standards

Reference standard methodology	No of studies
Regular practice	58
Gold standard	49
Trained standard	5
Unknown (process for determining correctness not specified in the paper)	1

Table 4 Statistical methods

Statistical method reported	No of studies
Recall or sensitivity	78
PPV or precision	52
Specificity	49
Accuracy	28

Many of the studies reported more than one statistical method—for example, both recall and precision. In these instances the study is reflected in more than one statistical method in the table so the number of studies in table 4 does not match the number of studies (N=113) in the included corpus.

- Binary: a two-factor scheme, such as follow-up or no follow-up, presence or absence of a particular condition, or positive/negative finding
- Multiple binary: application of multiple two-factor schemes, such as the presence/absence of more than one condition
- 3-4 point scale: application of a limited set of factors, such as yes/no/maybe, present/absent/uncertain, or three to four different elements identified
- Plenary: application of a much more complex coding and classification scheme with multiple conditions or codes. Some examples include: asthma management checklist, 1–5 risk classes for severity, 56 respiratory conditions, Gleason tumor score, and the five A's of smoking cessation (Ask, Advise, Assess, Assist, Arrange).

The wide variety of coding and classification schemas and study methodologies among the studies in this review made them difficult to compare and contrast. This heterogeneity prevented us from performing a meta-analysis. In addition, sensitivity without specificity cannot be interpreted as a statistical measure. Therefore, no statistical analysis was performed. Instead, we examined the study results as reported, and we observed the study results over time for obvious patterns. Online figures 2, 3, 4 and 5 (available at www.jamia.org) reflect scatter plots for the most commonly reported results.

As shown in online figures 1 to 5 (available at www.jamia.org), the results were wide and varied with no obvious trends and, surprisingly, no obvious improvement in performance over time. Sensitivity scatter plots, shown in figures 6 and 7, dividing the studies by type as identified above, also showed little significant pattern.

Further analysis is required to determine if the results did indeed remain static over time, or if this simply reflects attempts at more and more difficult tasks by the automated systems being evaluated. The more difficult tasks are those involving multiple parameters requiring multiple and complex computer algorithms. Thus, the most difficult coding and classification tasks for the computer applications studied here were those that fell into either group 1 or the plenary subdivision of group 2. Figure 8 shows that nearly all the group 2 plenary coding and classification studies were conducted since 2000, with most in 2005, 2006 and 2008. We did not attempt to correlate the complexity of the tasks undertaken with the evaluation results, but our review indicates that more difficult tasks have been undertaken by automated coding and classification systems in recent years.

Table 5 Subdivision of coding and classification schemes in the studies in group 2

Subdivision of group 2 studies	No of studies
Plenary	33
Binary	16
Multiple binary	12
3–4 point scale	10

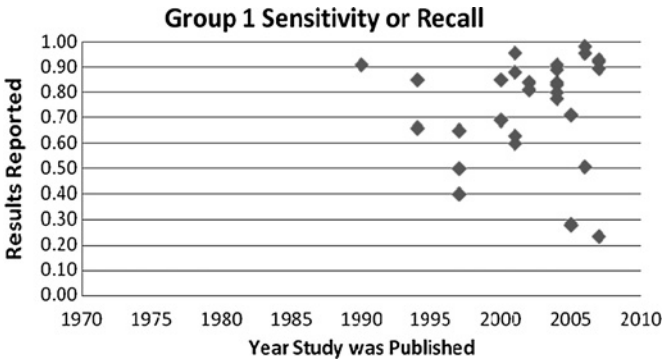

Figure 6 Scatter plot of sensitivity or recall results reported for group 1 studies. Note: group 1 studies included those that used an existing classification, vocabulary, or terminology system.

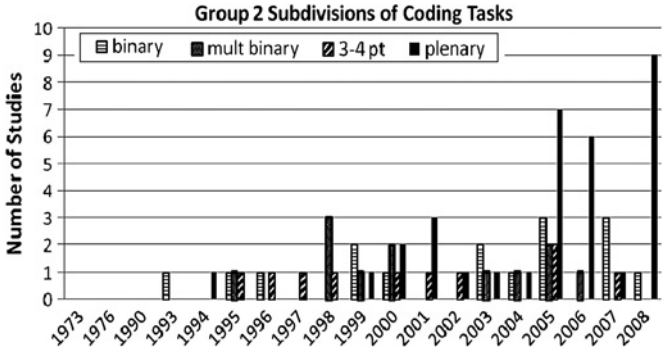

Figure 8 Group 2 subdivisions of coding and classification tasks. Note: group 2 included studies that used a clinical guideline or clinical coding scheme, often developed specifically for the study.

Given that these studies did not lend themselves to a meta-analysis, we focused on examining the study elements and results themselves for evidence on how the systems performed the tasks of coding or classification. We examined the corpus to determine if automated coding and classification systems were being used to solve practical real-world problems, and found they have been developed for a number of different purposes, from clinical support to biosurveillance to reporting quality measures. Table 6 applies a schema to these purposes.

DISCUSSION

It is clear from the time span these studies cover that researchers have been trying for years to solve the problem of time-consuming chart review using automated methods. For example, attempts to identify subjects automatically for controlled trials, or applying clinical guidelines and structuring text for clinical decision support, have been studied since the mid-1990s. The timing of the development of automated techniques for biosurveillance appears to be related to environmental factors, given that the earliest system studied was piloted at the 1996 Atlanta Olympics, with the anthrax exposures of 2001 and Salt Lake City Olympics in 2002 spurring additional work. The application of automated systems to reporting quality measures and automating problem lists has only recently been studied, perhaps reflecting the current dual priorities of improving healthcare quality while reducing healthcare expenditures.

There are varying degrees of complexity associated with the coding or classification tasks studied, and more work is needed to correlate purpose and related complexity with evaluation results. Clearly, computers can automatically assign codes and classes to unstructured data, but how well do they actually perform? The researchers who conducted the evaluations had much to say about this. Chapman and Haug[7] asserted as early as 1999 that the five algorithms tested in their evaluation performed better than lay persons and at least equal to physicians in a simple binary task of identifying acute bacterial pneumonia on chest x-ray reports. They observed that computerized techniques were more consistent than humans, but that human intuition applied to the task made it difficult to compare humans and computers. In 2000, Elkins *et al*[8] found that, when multiple parameters were involved (ie, not a binary task), computers were not as accurate as humans, but also noted that manual and automated coding each introduced separate errors. Chapman *et al*[9] concluded in 2003 that 'text processing systems are becoming accurate enough to be applied to real-world medical problems.' However, as late as 2006, Kukafka *et al*[10] observed that 'coding tasks involving complex reasoning, such as those in which disparate pieces of information must be connected, are a difficult challenge for current NLP systems.' Of the 113 studies included in our review, 26 specifically asserted that the automated system performed better than, or as well as, humans, while only four explicitly stated that humans outperformed the automated system. A recurring theme was that automated coding and classification system performance was relative to the complexity of the task and the desired outcome.

Clearly, some systems perform well on specific tasks. The difficulty is recognizing what sort of problems automated

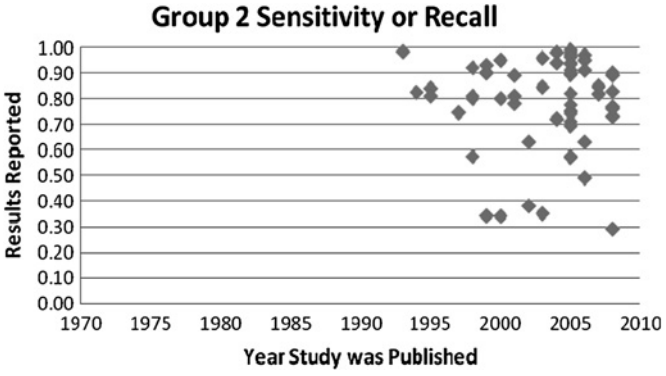

Figure 7 Scatter plot of sensitivity or recall results reported for group 2 studies. Note: group 2 included studies that used a clinical guideline or clinical coding scheme, often developed specifically for the study.

Table 6 Purposes of the automated systems studied

Purpose of the system	Count	Time span of studies
Structured text for clinical decision support/patient care	35	1996–2008
Facilitate retrieval of cases (eg, for research)	21	1994–2005
Testing techniques (eg, NLP methodologies)	17	1998–2008
Biosurveillance	13	1997–2008
Collect specific data	8	2000–2008
Administrative coding process	7	1973–2007
Automate problem lists	5	2005–2007
Apply clinical guidelines	4	1996–2003
Reporting quality measures	3	2007–2008

NLP, natural language processing.

systems tackle well. This is particularly challenging as medical natural language processing tools, commonly used in these tasks, are difficult to adapt, generalize, and re-use.[11] Turchin *et al*[12] reported that an obvious limitation in these tools was the lack of generalizability, '…a new set of regular expressions has to be developed and validated for each particular task.'

To assess whether automated systems currently available for administrative coding purposes perform as well as human coders, we looked more closely at the seven studies conducted to evaluate automation of the administrative coding process. The study elements outlined in online table 7 (available at www.jamia.org) underscore the variability in methodology and focus of the studies included in this administrative coding subset. A number of different systems were tested, applying various classification schemes to various document types. Four studies created a gold standard for comparison, while three relied on regular practice as the reference standard.

Online table 8 (available at www.jamia.org) provides summary level information of the results of the studies in the administrative coding subset. Dinwoodie and Howell[13] and Warner[14] evaluated the systems only on cases where the system was able to code with confidence. Eliminating cases that the system was unable, or uncertain how, to code introduced significant bias into their results. Findings by Morris *et al*[15] were promising, but rather than showing how well computer systems performed, they merely underscored how difficult it was to apply evaluation and management (E/M) code levels (a particularly difficult subset of codes) with any consistency. Results from Lussier *et al*,[16] while pointing to opportunities for improvement, do not appear sufficient for production, while subsequent results from Kukafka *et al*[10] and Goldstein *et al*[17] do not necessarily show the improvement one would hope to see and merely evoke cautious optimism. Findings by Pakhomov *et al*[18] were the most encouraging, with Type A results reaching 98% and Type B results from 90% to 95%. These authors also presented a possibility for partially using automated coding systems in conjunction with human oversight via tiered system outputs.

The 113 studies evaluating automated coding and classification systems included in this systematic literature review show that automated tools are available for a variety of purposes, are focused on various healthcare specialties, and are applicable to a wide variety of clinical document types. Differing research methodologies made it difficult to compare system performance. Two important distinctions that made it particularly difficult to evaluate performance were the mechanism used to create a reference standard against which the automated systems were evaluated and the statistical methods used to evaluate system performance. The complexity of the coding and classification schema used also varied widely, adding to this difficulty.

The types of coding and classification schemas applied by the systems studied fell into two primary groups, those that applied an existing classification system and those that applied a clinical coding scheme, perhaps developed specifically for the study. Further analysis is needed to correlate the complexity of the coding and classification task undertaken with the study results achieved.

This systematic literature review of automated coding and classification systems underscores that automating clinical coding is a difficult task, made even more difficult by the clinical texts that must be processed. Barrows *et al*[19] stated, 'As if NLU (natural language understanding) of narrative text documents by computer systems is not difficult enough, the understanding of notational text documents is perhaps even more difficult due to lack of punctuation and grammar, and frequent use of terse abbreviations and symbols.'

CONCLUSION

We conclude from this systematic literature review that automated clinical coding and classification system performance is relative to the complexity of the task and the desired outcome. Automated coding and classification systems themselves are not generalizable, and neither are the evaluation results in the studies. More work to correlate the purpose and related complexity of these studies with evaluation results could be informative, as would further analysis to determine if performance of automated systems has remained static over time or if the lack of obvious statistical improvement is a reflection of more and more difficult tasks being attempted by the automated systems under evaluation.

The published research examined in this review shows that automated coding and classification systems hold promise, but the application of automated coding must be considered in context. An additional issue requiring further study is what level of performance is required in order for these systems to perform useful real-world clinical tasks, such as providing input to an automated decision-support system, a clinical research study, or a quality-measurement analysis.[20] Further development of these systems and a better understanding of the tasks for which they will be used are needed before we can conclude that automated coding and classification systems meet performance standards adequate for use in complex clinical coding processes and are capable of applying appropriate guidelines for reporting these data.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. **Mulrow C,** Cook D, eds. *Systematic Reviews: Synthesis of Best Evidence for Health Care Decisions*. Philadelphia, PA: American College of Physicians, 1998.
2. **Meystre SM,** Savova GK, Kipper-Schuler KC, *et al*. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform* 2008:128—44.
3. **AHIMA computer-assisted coding e-HIM work group.** Delving into computer-assisted coding. *J AHIMA* 2004;**75**:48A—48H.
4. **Campbell JR,** Carpenter P, Sneiderman C, *et al*. Phase II evaluation of clinical coding schemes: completeness, taxonomy, mapping, definitions, and clarity. CPRI Work Group on Codes and Structures. *J Am Med Inform Assoc* 1997;**4**:238—51.
5. **Chute CG,** Cohn SP, Campbell KE, *et al*. The content coverage of clinical classifications. For The Computer-Based Patient Record Institute's Work Group on Codes & Structures. *J Am Med Inform Assoc* 1996;**3**:224—33.
6. **Wasserman H,** Wang J. An applied evaluation of SNOMED CT as a clinical vocabulary for the computerized diagnosis and problem list. *AMIA Annu Symp Proc* 2003:699—703.
7. **Chapman WW,** Haug PJ. Comparing expert systems for identifying chest x-ray reports that support pneumonia. *Proc AMIA Symp* 1999:216—20.
8. **Elkins JS,** Friedman C, Boden-Albala B, *et al*. Coding neuroradiology reports for the Northern Manhattan Stroke Study: a comparison of natural language processing and manual review. *Comput Biomed Res* 2000;**33**:1—10.
9. **Chapman WW,** Cooper GF, Hanbury P, *et al*. Creating a text classifier to detect radiology reports describing mediastinal findings associated with inhalational anthrax and other disorders. *J Am Med Inform Assoc* 2003;**10**:494—503.
10. **Kukafka R,** Bales ME, Burkhardt A, *et al*. Human and automated coding of rehabilitation discharge summaries according to the International Classification of Functioning, Disability, and Health. *J Am Med Inform Assoc* 2006;**13**:508—15.
11. **Zeng QT,** Goryachev S, Weiss S, *et al*. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. *BMC Med Inform Decis Mak* 2006;**6**:30.
12. **Turchin A,** Kolatkar NS, Grant RW, *et al*. Using regular expressions to abstract blood pressure and treatment intensification information from the text of physician notes. *J Am Med Inform Assoc* 2006;**13**:691—5.
13. **Dinwoodie HP,** Howell RW. Automatic disease coding: the 'fruit-machine' method in general practice. *Br J Prev Soc Med* 1973;**27**:59.
14. **Warner HR Jr.** Can natural language processing aid outpatient coders? *J AHIMA* 2000;**71**:78—81; quiz 83—74.

15. **Morris WC,** Heinze DT, Warner HR Jr, *et al*. Assessing the accuracy of an automated coding system in emergency medicine. *Proc AMIA Symp* 2000:595—9.
16. **Lussier YA,** Shagina L, Friedman C. Automating ICD-9-CM encoding using medical language processing: a feasibility study. *J Am Med Inform Assoc* 2000:1072—2.
17. **Goldstein I,** Arzrumtsyan A, Uzuner O. Three approaches to automatic assignment of ICD-9-CM codes to radiology reports. *AMIA Annu Symp Proc* 2007:279—83.
18. **Pakhomov SV,** Buntrock JD, Chute CG. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. *J Am Med Inform Assoc* 2006;**13**:516—25.
19. **Barrows RC Jr,** Busuioc M, Friedman C. Limited parsing of notational text visit notes: ad-hoc vs. NLP approaches. *Proc AMIA Symp* 2000:51—5.
20. **Hersh W.** Evaluation of biomedical text-mining systems: lessons learned from information retrieval. *Brief Bioinform* 2005;**6**:344—56.