A SYSTEMATIC REVIEW OF AUTOMATED MEDICAL CODING AND CLASSIFICATION SYSTEMS

by

Mary H. Stanfill, RHIA

A CAPSTONE PROJECT

Presented to the Department of Medical informatics & Clinical Epidemiology

and the Oregon Health & Science University School of Medicine

in partial fulfillment of

the requirements for the degree of

Master of Biomedical Informatics

May 2009

APPROVALS

School of Medicine

Oregon Health & Science University

Certificate of Approval

This is to certify that the Master's Capstone Project of

Mary H. Stanfill, RHIA

"A Systematic Review of Automated Medical Coding and Classification Systems"

Has been approved

Capstone Advisor – William Hersh, MD

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr. William Hersh, whose assistance and support made this project possible. I would also like to thank the other professionals who directly contributed to this work. Margaret Williams, MLS assisted with the database searches and is credited with the thoroughness with which these searches were undertaken. Susan Fenton, PhD, RHIA and Robert A. Jenders, MD, MS were instrumental in applying inclusion criteria and otherwise advising on the methodological quality of this systematic review. I feel truly privileged to have worked with a team with such depth of expertise. I am grateful for their assistance and continued support as we persevered to complete this work.

Thank you also to Andrea Ilg and Diane Doctor for their support of the distance learning students and myself in particular.

I would not have been able to participate in this masters program without the support and encouragement of my colleagues at AHIMA, who made it possible for me to divide my attention between work and graduate school.

I am humbled by and grateful for the love and encouragement of my family, who give meaning and purpose to my life and work.

ABSTRACT

Correct coding and reporting of healthcare diagnoses and services has become more critical in recent years as health care data needs have evolved. Computer applications for automating this process are available yet, to date such automated solutions are not widely used. The objective of this systematic review is to assess whether automated coding and classification systems, currently available for administrative coding purposes, perform as well as human coders. Recognizing that a great deal of research has been done on automated medical coding and classification, with only a small portion focused on administrative coding classifications systems, we determined to review all types of automated coding and classification evaluation studies and determine if the available evidence is conclusive for performance in the coding process currently employed industry-wide to gather healthcare data

Methods: The criteria for study inclusion in this systematic review were that the study had to be an original study involving research on the use of a computer application to automatically generate medical codes, from free-text clinical documents. The research had to be done 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 had to evaluate the performance of the computer application for classifying medical codes. The type of classification applied in the study was not constrained. A search strategy was designed to identify all potentially relevant publications about the accuracy of automated coding and classification systems. Searches were last conducted in February, 2009 so this review includes all studies published (or pre-published) and, where applicable, indexed for Medline prior to March 1, 2009.

Results: The 113 studies included in this systematic review show that automated tools are available for a variety of coding and classification purposes, focused on various healthcare specialties, and include a wide variety of clinical document types. Study methodologies varied widely across the included corpus making it difficult to compare performance of the systems. One methodological distinction was the mechanism used to create a reference standard against which the automated systems were evaluated. Another important distinction was the statistical methods employed to evaluate system performance. The complexity of the coding task also varied widely adding to the complexity of comparing study results.

Conclusion: We conclude that automated medical coding and classification 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. The published research in this review shows that automated coding and classification systems hold some promise, but application of automated coding and classification must be considered in context. Further research is needed before a conclusion can be reached on whether or not automated coding and classification systems are fit for use in the complex coding process used for capturing ICD-9-CM and CPT codes and the application of guidelines used for administrative reporting of this data.

A SYSTEMATIC REVIEW OF AUTOMATED MEDICAL CODING AND CLASSIFICATION SYSTEMS

Introduction

This paper summarizes the findings of a systematic literature review on the performance of automated coding and classification systems. The systematic review resulted from a desire to explore whether automated coding systems perform as well as human manual coding. The central research question is: Can computer applications code as well or better than people? To begin to answer this question, we resolved to examine the existing evidence. Thus the purpose of this systematic review was to fully explore the extent to which the performance of automated coding and classification systems has been investigated and determine how relevant these investigations are to the coding process currently employed industry-wide to gather healthcare data.

The author could not locate an existing systematic review on automated coding and classification systems. Meystre et al. (2008) conducted a narrative review to examine published research on the extraction of information from textual documents in the EHR.[1] In the Meystre review, natural language processing techniques were reviewed across the included publications. While a few of the studies included in the Meystre review were automated coding and classification studies, the majority were not. Meystre et al. focused on the performance of information extraction systems, a much broader concept than automated medical coding and classification. The few automated coding and classification studies included in the Meystre review were very narrow so the review did not provide a perspective on the full extent to which automated coding and classification systems have been investigated. Hence this systematic review on automated coding and classification was undertaken.

There are multiple methodologies for automated coding and classification to convert clinical information in free text to some sort of structured or "coded" format. There are also multiple coding and classification schemas, including standardized classifications systems, such as ICD or CPT, and use-case specific, non-standardized schemas, such as the presence or absence of a given condition. Without a doubt the most resource intensive coding process in the US healthcare industry is that employed routinely to capture diagnoses and procedures in the standard classification systems for every healthcare encounter or service. However, research studies do not focus only on coding and classification for that particular purpose. So this systematic review included published research on any computer application designed to automatically generate any type of medical code or classification from free-text clinical documents.

There are well established methodologies for conducting a systematic review[2]. This particular review was made more complex by the very nature of the subject matter. There were no directly related MeSH terms to facilitate retrieval of studies, so multiple search methodologies were employed and a large number of unrelated studies were considered and excluded in order to find those that were indeed related to exploring the question of how well computer applications performed in classifying medical codes. This paper presents the results of this review and begins to explore what can be learned from the corpus of included studies.

Background

Computerized tools are available to automate the assignment of certain medical or surgical codes (ICD-9-CM and HCPCS/CPT) from clinical documentation that are traditionally assigned by coding or health information management professionals as well as clinical providers.

In the manual coding workflow typically employed today, clinical documentation is analyzed, translated into ICD-9-CM or HCPCS/CPT codes, and entered into a database. New automation tools for coding allow the translation process to be assisted by computer software instead of manual review and translation alone. This type of automated coding and classification is an emerging technology. People are conducting studies of such software, but these studies more often focus on use of this technology for structuring limited clinical data. The applicability of these studies to the coding process widely employed to gather data in the healthcare industry is not clear.

It is important to fully explore the extent to which the performance of automated coding and classification systems have been investigated to date and determine how relevant these investigations are to the coding process currently employed industry-wide to gather healthcare data. Correct coding and reporting of healthcare diagnoses and services has become more critical in recent years as health care data needs have evolved. Uses of data, encoded in the ICD-9-CM and CPT administrative code sets for example, continue 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 is typically encoded via a manual coding process. This manual process involves review of clinical documentation with code "look up" and subsequent selection of each applicable code. Code look up may be done using code books, picking from abbreviated lists, or via software applications that facilitate alphabetic searches and provide edits and tips. Code assignment may be done 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, yet, to date, such automated solutions are not widely used. This may be because such systems are still in development and their performance in production is unproven. The objective of this review was to assess whether automated coding and classification systems, currently available for administrative coding purposes, perform as well as human coders. Recognizing that a great deal of research has been done in this area, with only a small portion focused on administrative coding classifications systems, we determined to review all types of automated coding and classification evaluation studies.

This systematic review was undertaken to identify all of the published studies of automated coding and classification and determine if they are conclusive for performance in the coding process currently employed industry-wide to gather healthcare data.

Methods

Inclusion Criteria: The criteria for study inclusion in this systematic review were that the study had to be an original study involving research on the use of a computer application to automatically generate medical codes and/or assign classes from free-text clinical documents. The research had to have been done with documents produced in the process of clinical care where both the documents and the computer application were in the English language. The study must have also evaluated the performance of the computer application for assigning medical codes or some type of classification schema.

The type of coding and classification schema applied in the study was not constrained by inclusion criteria. Recognizing that there are multiple coding and classification schemas, including standardized classification systems, such as ICD or 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 medical codes or classes that might be evaluated in a published study.

Studies were excluded if the automated application was not evaluated for performance of the code assignment. Instances where the study focused on evaluating content coverage of the classification or vocabulary for example 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 there was not some type of defined coding or classification system 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 medical codes since codes are often used for the purpose of indexing or retrieving information. Where indexing was done via some sort of coding or classification schema, for example the application of MeSH terms, the study was included. Where an indexing study involved parsing or tagging documents with no specific code output to evaluate, the study was excluded.

Search Strategy: A search strategy was designed to identify all potentially relevant publications about the performance of automated coding and classification systems and used to search MEDLINE, ACM, CINAHL, the Inspec database, and Science Citation Index Expanded. See Appendix A for search parameters and the details of the search statements employed in searching the various databases. Searches were last conducted in February, 2009 so this review includes all studies published (or pre-published) and, where applicable, indexed for Medline prior to March 1, 2009.

In addition to searching these databases, all articles in the AHIMA's FORE Library: HIM Body of Knowledge indexed to the subject "computer-assisted coding" were added to the list for review. References in the “FasterCures” report, "Think Research: Using Medical records to Bridge Patient Care and Research" were also checked for relevancy. We also employed “snowball” methods (pursuing references of references) and sought advice on sources from experts in the field. A core group of researchers in the field was contacted for input regarding potential gaps in the studies selected for review. Their suggestions were included in the pool of potentially relevant studies that were evaluated for inclusion.

Inclusion/Exclusion: All potentially relevant studies identified were reviewed to apply the inclusion criteria. Each title/abstract retrieved was reviewed by two independent reviewers. When the two reviewers disagreed, the full article was obtained and a third reviewer made the final decision. We retrieved the full-text articles for citations selected for possible inclusion in the systematic review. When an article met the criteria for inclusion, summary information was extracted from the study.

The various mechanisms for searching the literature yielded 2322 possibly relevant references. Applying the inclusion criteria resulted in a total of 113 studies included in this systematic review. The 113 included studies are listed in Appendix B. Meta-analysis of these studies was not possible, given the variety of research purposes and study methodologies across the included corpus. Instead the 113 studies were closely reviewed and key data elements, such as the following, were abstracted.

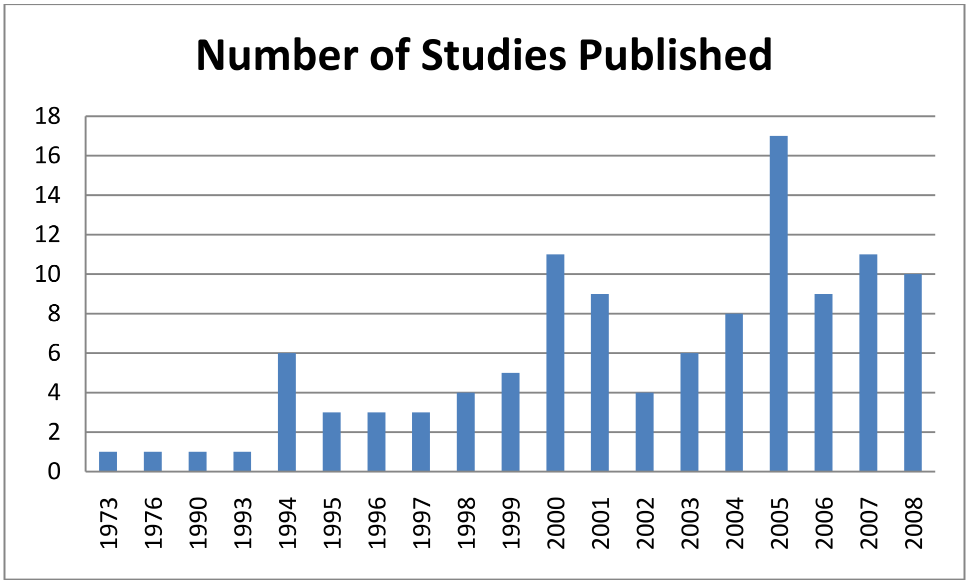
- The classification system applied by the automated system and associated healthcare domain (e.g. SNOMED for diagnoses on chest radiographs)
- Objective of the study (e.g. to determine if an automated system can replace manual chart review to identify cases for a clinical trial)
- The study methodology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

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

Results

The earliest study in the included corpus was published in 1973. There was another study published in 1976 and then nothing more until 1990. All but four of the studies (96%) were published in the 14 years beginning in 1994 to present. Figure 1 shows the distribution of included studies across time.

Figure 1: Distribution of publication of the studies in the included corpus across time



The studies in the included corpus focused on various conditions or healthcare specialties and a wide variety of document types. Table 1 below provides details on the conditions and document types specified in the included studies. Pneumonia was the most frequently studied condition. The studies included community acquired pneumonia as well as acute bacterial pneumonia, while a couple focused on early detection of pneumonia in neonates. Interestingly, 37 of the studies that specify a particular condition were focused on some type of respiratory condition. This correlates with the document types studied. The most frequently studied documents were chest x-ray reports. In general, various diagnostic reports were studied more often than other report types with 54 of the specified document types representing some sort of diagnostic test. Discharge summaries were the next most frequently studied document type.

Table 1: Healthcare Domains Studied

Specific Conditions	**Count**	**Document Types**	**Count**
Pneumonia	18	Chest x-rays	26
Heart failure/CHF	12	Discharge summaries	17
Neoplasms	6	Radiology reports	13
Pleural effusion	6	Various medical reports	9

Specific Conditions	Count	Document Types	Count
Pneumothorax	6	ED records	7
Smoking cessation	6	ED for chief complaints	7
Asthma	5	Encounter notes	6
Cardiovascular problems (mult)	4	EMR notes	5
Syndromic categories	4	Surgical pathology reports	5
COPD	3	Outpatient notes	4
Acute/chronic respiratory illness	2	Neuro-radiology reports	3
Breast cancer	2	Pathology reports	3
Diabetes	2	Admission reports	2
General practice	2	Diagnostic statements	2
Hypertension	2	History &Physicals	2
Medications	2	Inpatient notes	2
Tuberculosis	2	Physician notes	2
Acute GI syndrome	1	Surgical notes	2
Adverse events	1	Anesthesia medication record	1
AMI & CAP guidelines	1	Colonoscopy report	1
Anthrax (inhalation)	1	CT report	1
Appendicitis	1	Death certificate	1
Atelectasis	1	ICU problem list	1
Causes of death	1	Ventilation/perfusion lung scans	1
Central venous catheter	1	Mammogram reports	1
Fever	1	Pelvic ultrasound reports	1
Fracture	1		
Glaucoma	1		
Lung cancer	1		
Ovarian cancer	1		
Prostate cancer	1		
Pulmonary embolism	1		
Rehab diagnoses (most freq)	1		
Rib fracture	1		
Vaccine reactions	1		
Venous thrombo-embolism	1		

Table 1: Healthcare Domains Studied. Some studies addressed more than one condition or document type. Also some studies did not specify a particular condition, but instead were looking for concepts in the document or perhaps negation phrases for example. For these reasons, the counts do not add up to the number of studies (N= 113) in the included corpus.

The studies in the included corpus evaluated the performance of various computer applications, many of which were identified by name. Table 2 below provides details on the systems named in the studies in the included corpus. There were forty-six (46) different systems named and twenty-one (21) un-named systems. Of the named systems, Columbia University's MedLEE was the system utilized most frequently. MedLEE was named in 24 included studies (or 21%). MedLEE was followed in frequency by SymText, MMTx, and NegEx. These four systems together represent 91% of the named systems studied; and they represent 37% of the studies in the included corpus.

Table 2: Systems Tested

System Tested	No. of Studies
Not named	21
MedLEE	16
MedLEE output used with various algorithms/processors	8
SymText	7
MMTx	5
NegEx	5
SAPHIRE	3
A-life Medical's CM Extractor	2
A-Life medical's LifeCode	2
CoCo (the classifier used for RODS)	2
DITTO (Diabetes Identification Through Textual element occurrences)	2
EMT-P (Emergency Medical Text Processor)	2
GATE	2
MediClass	2
NA (challenge: different teams with mult different systems participating)	2
Nuance Leximer	2
REX (Regenstrief extraction tool)	2
RODS system	2
UMLS MetaMap or search engine	2
Three fever detection algorithms: Keyword HP, Keyword CC, CoCo	1
Three: Lucene, BoosTexter, hand crafted rules	1
ACS (Automated Coding System)	1
Automatic medical index classification architecture	1
caTIES	1
CSIS (cancer stage interpretation system)	1
EM and SVM	1
FIGLEAF	1
Fruit Machine	1
HITEx (health information text extraction)	1
INQUERY and FIGLEAF	1
IPS (identify patient sets system)	1
JESS inference engine, ontology enhanced with WSSS	1

System Tested	No. of Studies
LSP-MLP (Linguistic String Project medical language processor)	1
MED and a natural language processor	1
MVP (Mayo vocabulary processor)	1
Negfinder	1
NIH and Northwick Park Pilot encoder	1
NLUSs (natural language understanding systems)	1
Nuance med extraction system	1
NUDIST	1
PostDoc and Pindex	1
RadTRAC	1
RuleQuest's See5 data mining tool	1
SPRUS (syndromic surveillance form textual records)	1
TRANSOFT with context-insensitive and context-sensitive model	1
UIMA (unstructured information management architecture)	1

The types of organizations conducting the research represented in the included studies can be classified as academic institutions, corporate organizations, governing bodies, or provider organizations. Table 3 below shows what types of organizations were responsible for the included studies. As demonstrated in this table, academic institutions were involved in 75% of this work.

Table 3. Types of Organizations conducting this research

Type of Organization	No. of Studies
Academic/ Educational institutions	85
Provider organization	20
Corporate organization	6
Governing body (both Australia)	2

It is notable, though not surprising, that the majority of the studies (68%) indicated that the research was made possible, at least in part, by a grant. Only 36 of the 113 included studies made no mention of a grant. This percentage was even higher within academic institutions. Out of the 85 studies conducted by academic institutions, only 18 made no mention of a grant, indicating that nearly 79% of this work in academic organizations was supported by grants. As one would expect this percent drops markedly in provider organizations, where only 55% indicated that the work was supported by grant funds.

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. In reviewing the included studies, 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; could be majority vote or review/discussion to obtain agreement
- Trained standard: one expert classifies the majority of the training set, however validity of the expert's assignment is verified and training is provided to improve performance/consistency
- Regular practice: one human reviewer, as in the usual manual process; often an existing database reflecting the normal or usual practice was employed

Table 4 applies this schema to the included corpus. Creating a gold standard, with multiple independent reviewers, is costly but generally recognized as a more rigorous approach. Forty-three percent (43%) of the included studies employed a gold standard as it is defined here. Approximately half (51%) of the studies compared the automated process to the usual manual process, employing regular practice as the standard for comparison.

Table 4. 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

Another distinction in the methodologies employed across the corpus was the statistical methods employed. Many studies reported more than one measure, for example sensitivity and specificity, or recall and precision. Some studies reported simple accuracy rates. A handful of studies reported more rigorous statistics, such as Kappa scores, F measures, and receiver operating characteristic (ROC) curve analysis. Table 5 includes the most commonly reported statistics and shows that the most frequent statistical measure was recall (also called sensitivity).

Table 5. Statistical Methods

Statistical Method Reported	**No. of Studies**
Recall/ Sensitivity	78
PPV/Precision	52
Specificity	49
Accuracy	28

Table 5: Statistical methods. 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 5 does not match the number of studies (N = 113) in the included corpus.

The type of coding and classification scheme applied by the system evaluated in the study also varied widely. We found that the types of coding fell into two primary groups: 1) those that employed some sort of classification, vocabulary, or terminology system and 2) those that did not use an existing classification system but rather employed a clinical guideline or clinical coding scheme, often developed specifically for the study. Forty two (42) of the studies were classified in Group 1, while the remaining 71 studies were classified 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 71 studies in Group 2 were further subdivided to reflect the complexity of the coding and classification scheme applied. We subdivided them as follows:

- 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 5 A's of smoking cessation (Ask, Advise, Assess, Assist, Arrange).

Table 6 reflects how the 71 studies in Group 2 were divided according to these subdivisions

Table 6. 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

Discussion

The wide variety of coding and classification schemas and study methodologies among the studies in the included corpus made it difficult to compare and contrast these studies. However, reported results were analyzed to determine what can be learned from this corpus. We could not do a meta-analysis but statistical results were plotted in scatter plots over time to observe for any obvious patterns. Figures 2, 3, 4, and 5 reflect scatter plots for the most commonly reported statistical measures.

Figure 2: Scatter plot of Recall or Sensitivity results reported in the included corpus

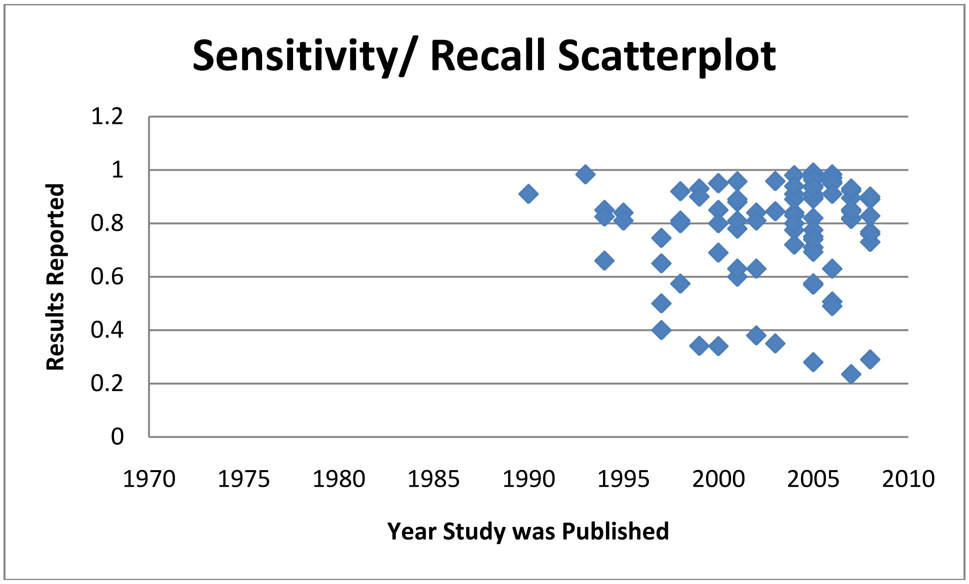


Figure 3: Scatter plot of Specificity results reported in the included corpus

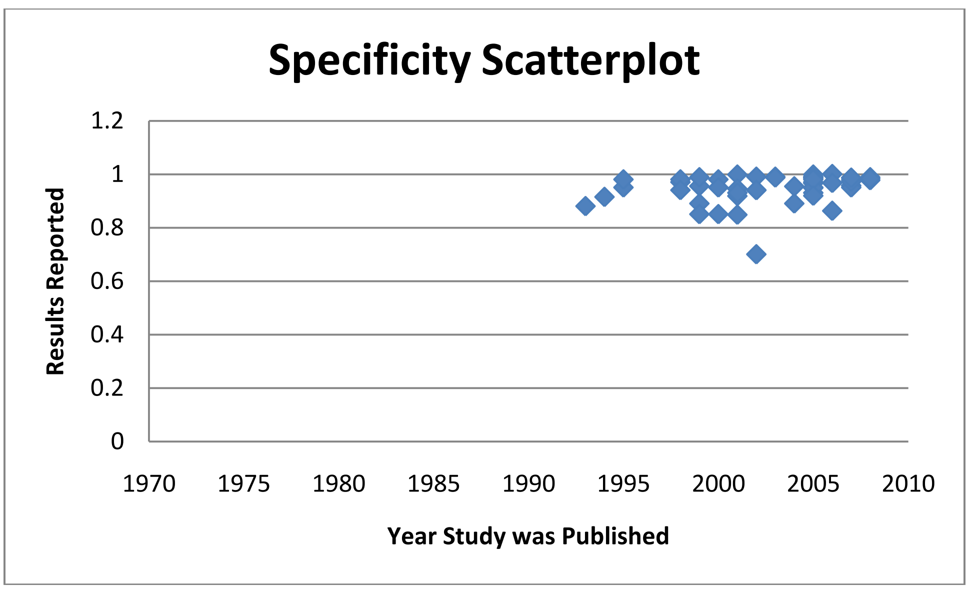


Figure 4: Scatter plot of PPV or Precision results reported in the included corpus

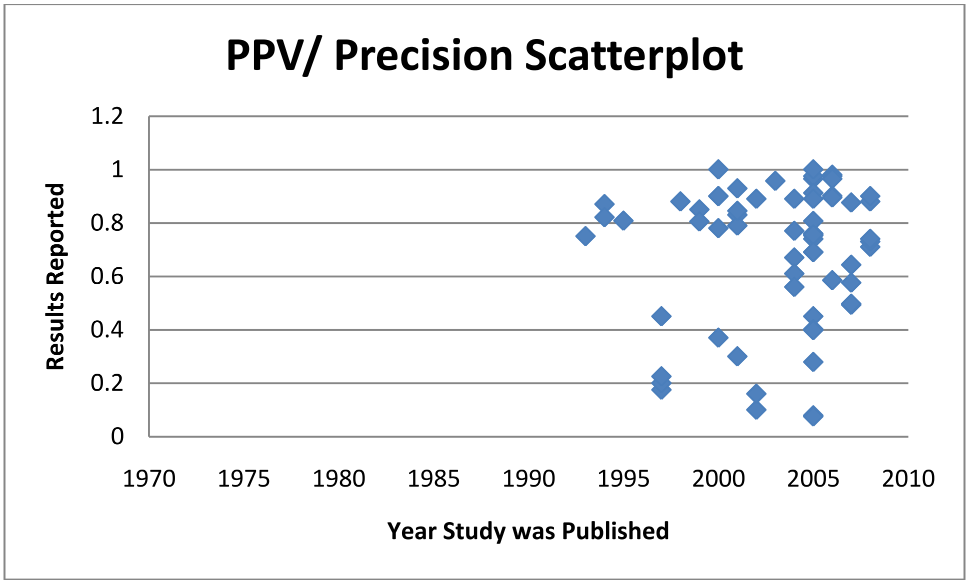
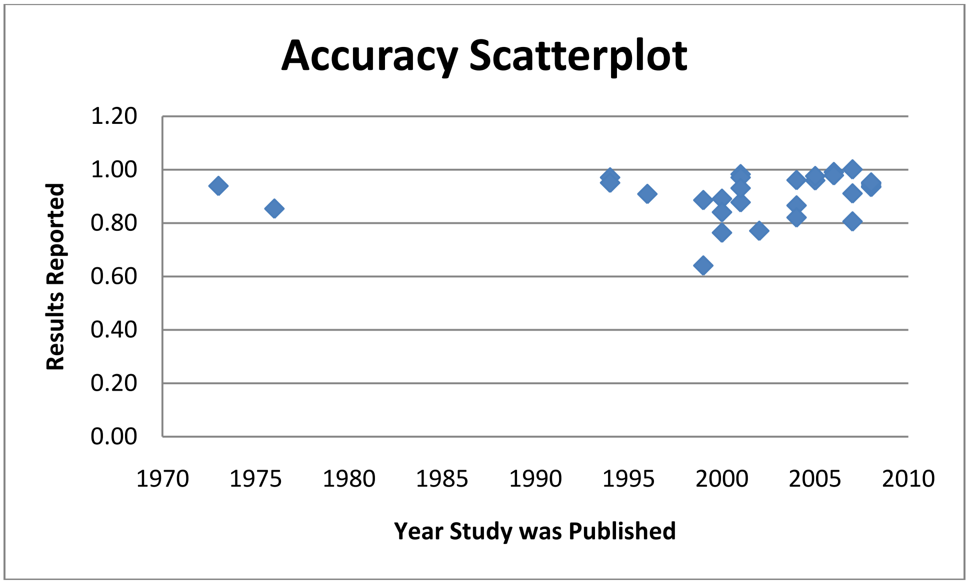


Figure 5: Scatter plot of Accuracy results reported in the included corpus



As shown in Figures 2 through 5, the results were wide and varied but no obvious trends emerged across time. Surprisingly, no obvious improvement in performance emerges across time. Sensitivity scatter plots (in Figures 6 and 7) separating the studies into the Group 1 and Group 2 types we applied also showed little recognizable pattern.

Figure 6: Scatter plot of Sensitivity/Recall results reported for Group 1 studies

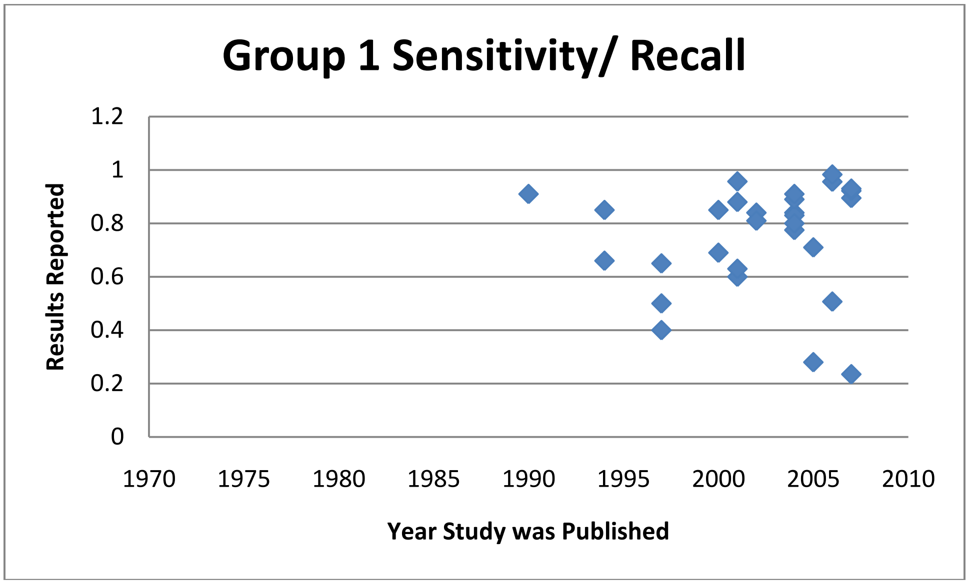
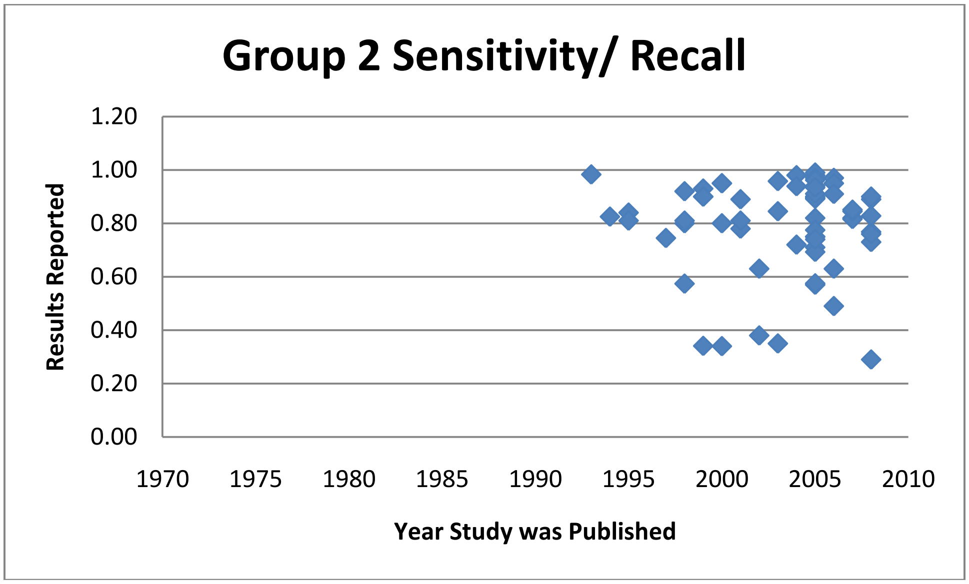
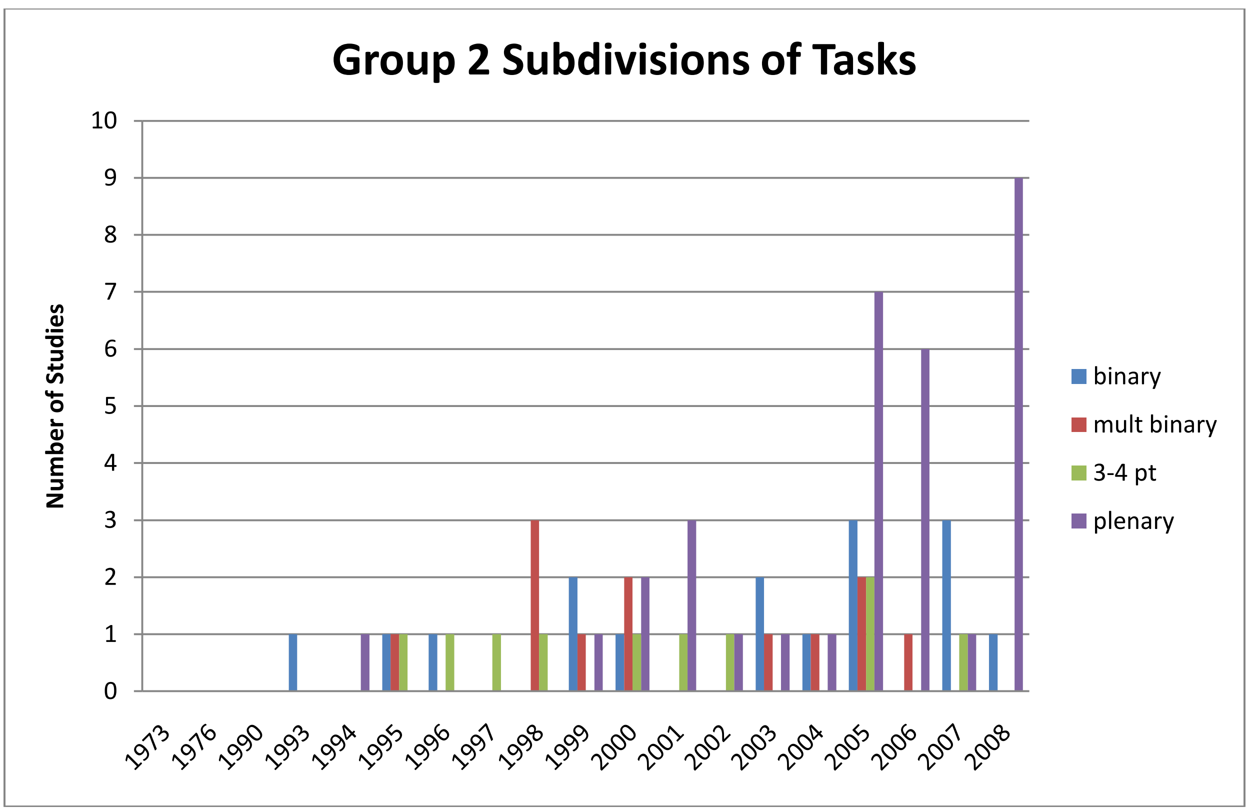


Figure 7: Scatter plot of Sensitivity/Recall results reported for Group 2 studies



Further analysis would be required to determine if the results indeed remain static over the years or if this is a reflection of more and more difficult tasks being attempted by the automated systems under evaluation. The more difficult tasks are those that involve multiple parameters requiring multiple and complex computer algorithms. Therefore the most difficult coding and classification tasks for a computer application among the types in this corpus were those in Group 1 and the Plenary subdivision of Group 2. When we looked at the timing of the coding and classification tasks in Plenary Group 2 there did seem to be an indication that the more difficult Plenary type of coding and classification tasks were studied in more recent years. Figure 8 shows that nearly all of the Group 2 Plenary coding and classification tasks were conducted since 2000 with most in 2005, 2006 and 2008. More work would be needed to correlate the complexity of the task undertaken with the evaluation results. However, there does seem to be an indication that more difficult tasks were undertaken by automated coding and classification systems in more recent years.

Figure 8: Group 2 Subdivisions of Coding and Classification Tasks



We further analyzed the included corpus to determine if automated coding and classification systems were employed to solve practical real-world problems. We found that the studies evaluated systems developed for several different purposes. Table 7 applies a schema of the various purposes to studies in the included corpus.

Table 7. Purposes of the automated systems studied

Purpose of the System	Count	Time Span of Studies
Structured text for clinical decision support/ pt care	35	1996 – 2008
Facilitate retrieval of cases (e.g. for research)	21	1994 – 2005
Testing techniques (e.g. NLP methodologies)	17	1998 – 2008
Bio surveillance	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

It is clear from the time span of the studies for these various purposes that researchers have been trying to solve the problem of time consuming chart review through automated methods for several years. Attempts to automatically identify cases for controlled trials for example, as well as applying clinical guidelines and structuring text for clinical decision support, have been studied since the mid 1990's. The timing of automated techniques for biosurveillance appears to be related to environmental factors. The earliest system studied was piloted at the 1996 Atlanta Olympics. The anthrax exposures of 2001 as well as the Salt Lake Olympics in 2002 appear to have spurred additional work. It is interesting however, that the application of automated systems to address real-world problems such as reporting quality measures and automating problem lists have only recently been studied.

There are varying degrees of complexity associated with each of these purposes. More work could also be done to correlate the purpose and related complexity with evaluation results in the corpus. Clearly computers can automatically assign codes and classes for each of these purposes, but how well do they actually perform? The researchers who conducted the evaluations in the included studies had much to say about this. As early as 1999, Chapman and Haug[7] relayed 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, when multiple parameters were involved (i.e. not a binary task), computers were not yet as accurate as humans. They also noted that manual and automated coding each introduced separate errors. Chapman et al. concluded in a 2003 study that "text processing systems are becoming accurate enough to be applied to real-world medical problems."[9] But as late as 2006, Kakafka et al. 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."[10] Out of the 113 studies included in our corpus, the authors of 26 studies specifically stated that the automated system performed better or equal to humans. Only four explicitly stated that humans outperformed the automated system. The recurring theme throughout the corpus 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 in recognizing what sort of problems automated systems can perform well. This is particularly challenging as medical natural language processing (NLP) tools, commonly used in these tasks, are difficult to adapt, generalize and re-use.[11] Turchin et al. reported 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."[12] Therefore, automated coding and classification systems are not generalizable and thus neither are the evaluation results.

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 for the purpose of automating the administrative coding process. Tables 8 and 9 provide some of the study elements from these seven studies.

Table 8. Study elements on systems to automate the administrative coding process

Study & year published	**System tested**	**Classification system applied**	**Condition & document types**	**Reference standard**
Dinwoodie [13] 1973	Fruit machine	ICD-8	General practice diagnostic statements	Regular practice
Goldstein [14] 2007	Lucene, BoosTexter, & hand crafted rules	ICD-9-CM	Impressions in radiology reports	Gold standard created
Kukafka [10] 2006	MedLEE	ICF	Diagnoses in rehab discharge summaries	Gold standard created
Lussier [15] 2000	MedLEE	ICD-9-CM	diagnoses in discharge summaries	Regular practice
Morris [16] 2000	A-Life's LifeCode	CPT E/M codes	Service levels in ED records	Gold standard created
Pakhomov [17] 2006	automatic medical index classification architecture	HICDA (an adaptation of ICD-8)	Diagnoses in outpatient visit notes	Regular practice
Warner [18] 2000	Not named	ICD-9-CM & CPT	Diagnoses and procedures in ED records	Gold standard created

The study elements relayed in Table 8 (above) underscore the variability in methodology and focus of the seven studies included in this administrative coding subset of the corpus. There were a number of different systems tested, applying various classification systems to various document types. Four of the studies applied the more rigorous methodology of creating a gold standard, while three relied on regular practice as the reference standard.

Table 9. Summary of reported results on systems to automate the administrative coding process

Study & year published	**System performance**	**Human performance**	**Pertinent Notes**	**Researcher's conclusions**
Dinwoodie [13] 1973	92% to 95.6% accurate	82.8% to 87.7% accurate	Accuracy rate for the system does not reflect the machine's "failure to code" some cases	The data suggests that machine coding is on the order of 8-10% more effective than manual coding.
Goldstein [14] 2007	FRB: precision = .876, recall= .895, F-Measure= .885. Lucene: precision = 0.695, recall= 0.646, F-Measure= 0.669. BoosTexter: precision = 0.852, recall= 0.751, F-Measure= 0.804		The 3 systems were compared to a gold standard. Human performance on this task was not assessed.	Semantic information significantly contributes to ICD-9-CM coding with lexical elements. Also, a simple hand-crafted rule based system with lexical elements and semantic information can outperform more complex systems.
Kukafka [10] 2006	NLP Kappa value = 0.593 to 0.160	Expert coders Kappa value= 0.719 to 0.591	The expert coders performed better than either the non-expert coders or the NLP system.	Automated coding can be used to assign ICF codes, with results similar to those obtained by human coders at least for the section of ICF and code assignment considered here.
Lussier [15] 2000	21% false pos, 24% false neg, 69% sensitivity		Human coding was considered the reference standard so human coding performance was not measured	To improve performance, need an external knowledge base to provide taxonomic information and special rules for coding.
Morris [16] 2000	0.71 Inter-rater consensus agreement	0.78 best expert's inter-rater consensus agreement	Consensus agreements ranged from a low of 0.59 to high of 0.78. LifeCode was about average among participants	Accuracy in E/M coding is relative. If an E/M coding consensus can be used to represent a "gold standard" then only moderate agreement was observed between any study participants. In which case,

Study & year published	**System performance**	**Human performance**	**Pertinent Notes**	**Researcher's conclusions**
				LifeCode is an advantage (it's predictable, repeatable, and fast).
Pakhomov [17] 2006	Type A: macro-av precision = 98%, recall = 98.3%, Fscore = 98.2%. Type B: macro-av precision =90.1%, recall= 95.6%, Fscore = 93.1%. Type C: macro-av precision =58.5%, recall= 50.7%, Fscore= 54.4%		Type A= 48% of EMR prob list entries Type B = 34% Type C = 18% Human coding performance not evaluated.	Precision/recall results of Type A exceeded objectives and were deemed appropriate to be left unsupervised. Type B must be supervised and aids the coding process. The study shows that we can reliably achieve our design objective- increase throughput without loss in accuracy on 82% of incoming diagnoses.
Warner [18] 2000	82% agreement with the ER coder.	Only 18% perfect agreement in all participants	67% of cases processed by the NLP system were eliminated from the study as either not meeting format requirements or the NLP system was not sufficiently "confident" to assign a code.	Agreement levels among all study participants were higher for diagnosis code assignment than for E/M level coding. E/M coders, human or automated agreed with each other somewhere between 43% to 78% of the time. Using degree of concordance as the metric, NLP compares favorably with the other participants in the study.

Table 9 provides summary level information on the reported results from each study. Two studies, Dinwoodie and Howell [13] and Warner [18], only evaluated the system on cases the system was able to code with confidence. Eliminating cases the system was unable or uncertain how to code introduced significant bias into these studies. Conclusions in the Morris et al.[16] study were promising, but rather than serving to show how well computer systems performed, it merely underscored how difficult it was to apply evaluation and management (E/M) code levels with any consistency (a particularly difficult subset of codes). Results in the

Lussier et al.[15] study do not appear sufficient for production. Lussier et al. did point out opportunities for improvement, yet results in the later Kukafka et al.[10] and Goldstein et al.[14] studies do not necessarily show the improvement one would hope to see and merely evoke cautious optimism. The Pakhomov et al.[17] study was the most encouraging with Type A results reaching 98% and Type B results from 90% to 95%. Pakhomov et al. also presented a tantalizing possibility for partially employing automated coding systems in conjunction with human oversight via tiered system outputs.

This published research shows that automated coding and classification systems hold some promise, but further research is needed before a conclusion can be reached on whether or not automated systems are fit for use in the specific coding process used for capturing ICD-9-CM and CPT codes and the application of guidelines used for administrative reporting of this data.

Conclusion

The 113 studies evaluating automated coding and classification systems included in this systematic review show that automated tools are available for a variety of purposes, are focused on various healthcare specialties, and include a wide variety of clinical document types. Study methodologies varied widely across the included corpus making it difficult to compare performance of the systems. One methodological distinction was the mechanism used to create a reference standard against which the automated systems were evaluated. Another important distinction was the statistical methods employed to evaluate system performance. The complexity of the coding and classification schema applied also varied widely, adding to the difficulty in comparing study results.

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

It is clear from the time span of the studies, for the various purposes found in this review, that researchers have been trying to solve the problem of time consuming chart reviews with automated methods for several years. Inconsistency in code assignment was a problem even as early as 1964, according to Smith and Melton who stated "…different coders are always at liberty to code the same diagnoses in different ways. This causes irregularities of the coded information and precludes uniform searching…"[19] Similarly, almost ten years later in 1973, Dinwoodie and Howell[13] stated "Coding is a repetitive, voluminous, tedious, error-prone clerical chore which seems, at least in theory, ripe for automation." Today, over 35 years later, we still find manual coding is expensive and error prone and automated tools are still needed.

This systematic review of automated coding and classification systems underscores that automating medical coding is proving to be a very difficult task, made even more difficult by the medical texts that must be processed. Barrows et al. stated, "As if NLU 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."[20]

We conclude from this systematic review that automated medical 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. More work could be done however to correlate the purpose and related complexity of these studies with evaluation results. Further analysis would also be needed to determine if results in the included studies infer that the performance of automated systems has remained static over time or if the lack of obvious statistical improvement over time is a reflection of more and more difficult tasks being attempted by the automated systems under evaluation.

The published studies in this review show that automated coding and classification systems hold some promise, but the application of automated coding must be considered in context. Further research is needed before a conclusion can be reached on whether or not automated coding and classification systems are fit for use in the complex coding process used for capturing ICD-9-CM and CPT codes and the application of guidelines used for administrative reporting of this data.

References

1. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF: **Extracting information from textual documents in the electronic health record: a review of recent research**. *Yearb Med Inform* 2008:128-144.

2. Glasziou P: **Systematic reviews in health care : a practical guide**. Cambridge, U.K.; New York: Cambridge University Press; 2001.

3. AHIMA computer-assisted coding e-HIM work group: **Delving into computer-assisted coding**. *J AHIMA* 2004, **75**(10):48A-48H.

4. Campbell JR, Carpenter P, Sneiderman C, Cohn S, Chute CG, Warren J: **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**(3):238-251.

5. Chute CG, Cohn SP, Campbell KE, Oliver DE, Campbell JR: **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**(3):224-233.

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**. *Journal of the American Medical Informatics Association* 1999:216-220.

8. Elkins JS, Friedman C, Boden-Albala B, Sacco RL, Hripcsak G: **Coding neuroradiology reports for the Northern Manhattan Stroke Study: a comparison of natural language processing and manual review**. *Comput Biomed Res* 2000, **33**(1):1-10.

9. Chapman WW, Cooper GF, Hanbury P, Chapman BE, Harrison LH, Wagner MM: **Creating a text classifier to detect radiology reports describing mediastinal findings associated with inhalational anthrax and other disorders**. *Journal of the American Medical Informatics Association* 2003, **10**(5):494-503.

10. Kukafka R, Bales ME, Burkhardt A, Friedman C: **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**(5):508-515.

11. Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R: **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, Makhni EC, Pendergrass ML, Einbinder JS: **Using regular expressions to abstract blood pressure and treatment intensification information from the text of physician notes**. *J Am Med Inform Assoc* 2006, **13**(6):691-695.

13. Dinwoodie HP, Howell RW: **Automatic disease coding: the 'fruit-machine' method in general practice**. *British Journal of Preventive & Social Medicine* 1973, **27**(1):59.

14. Goldstein I, Arzrumtsyan A, Uzuner O: **Three approaches to automatic assignment of ICD-9-CM codes to radiology reports**. *AMIA Annu Symp Proc* 2007:279-283.

15. Lussier YA, Shagina L, Friedman C: **Automating ICD-9-CM encoding using medical language processing: A feasibility study**. *Journal of the American Medical Informatics Association* 2000:1072-1072.

16. Morris WC, Heinze DT, Warner Jr HR, Primack A, Morsch AE, Sheffer RE, Jennings MA, Morsch ML, Jimmink MA: **Assessing the accuracy of an automated coding system in emergency medicine**. *Proc AMIA Symp* 2000:595-599.

17. 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**(5):516-525.

18. Warner HR, Jr.: **Can natural language processing aid outpatient coders?** *J AHIMA* 2000, **71**(8):78-81; quiz 83-74.

19. Smith JC, Melton J: **Manipulation of Autopsy Diagnoses by Computer Technique**. *JAMA: The Journal of the American Medical Association* 1964, **188**:958.

20. Barrows Jr RC, Busuioc M, Friedman C: **Limited parsing of notational text visit notes: ad-hoc vs. NLP approaches**. *Proc AMIA Symp* 2000:51-55.

Appendix A: Search Parameters

The primary strategy for identifying all potentially relevant publications about the performance of automated coding and classification systems was developed for searching the PubMed interface to the National Library of Medicine's Medline database.[i] Since there is no MeSH term corresponding to "automated coding", the initial search was for related text words appearing in any field of the Medline record, including title and abstract. We searched the words or phrases "autocoding", "(automated or automatic) coding", "computer assisted coding", "automatic concept indexing", "computer coding", "(automated or automatic) extraction", "(automated or automatic) text mining", and limited the results to studies published in English.

> Search details:
> autocoding[All Fields] OR “automated coding”[All Fields] OR “automatic coding"[All Fields] OR "computer assisted coding"[All Fields] OR (automatic[All Fields] AND concept[All Fields] AND (("abstracting and indexing"[TIAB] NOT Medline[SB]) OR "abstracting and indexing"[MeSH Terms] OR indexing[Text Word])) OR "computer coding"[All Fields] OR "automated indexing"[All Fields] OR (automatic[All Fields] AND text[All Fields] AND ("mining"[MeSH Terms] OR mining[Text Word])) OR (automated[All Fields] AND text[All Fields] AND ("mining"[MeSH Terms] OR mining[Text Word])) AND English[Lang

Based on the results of the text search, we identified MeSH terms commonly found in the citations retrieved. Two appearing to be of particular relevance were "Medical Record Systems, Computerized" and "Natural Language Processing". These MeSH terms were combined in a number of ways with other terms, for example:

- Medical Record Systems (or Medical Records) combined with Natural Language Processing as a major indexing term.

> Search details: ("Medical Records Systems, Computerized"[MeSH] OR "Medical Records"[MeSH]) AND "Natural Language Processing"[MAJR] AND English[Lang]

- Natural Language Processing (in all fields, including MeSH terms) combined with articles with "Medical Record Systems, Computerized" or "Medical Records" or "Disease, Classification" or "Abstracting and Indexing", or "Medical

Records/Classification" as major indexing terms. News stories indexed to these terms were not included.

Search details: ("Natural Language Processing"[All Fields] AND ("Disease/classification"[MAJR] OR "Abstracting and Indexing"[MAJR] OR "International Classification of Diseases"[MAJR] OR "Medical Records/classification"[MAJR] OR "Medical Records Systems, Computerized"[MAJR])) OR ("Medical Records Systems, Computerized"[All Fields] AND ("Natural Language Processing"[MAJR] OR "International Classification of Diseases"[MAJR] OR "Disease/classification"[MAJR] OR "Diagnosis-Related Groups"[MAJR])) NOT News[ptyp] AND English[Lang]

- Natural Language Processing as a MeSH term, limiting the results to content where NLP was a major topic in the article

Search details: "Natural Language Processing"[MAJR] AND English[Lang]

- Medical Record Systems with MeSH terms for International Classification of Diseases (ICD), Healthcare Common Procedure Coding System (HCPCS), Current Procedural Terminology (CPT), Diagnosis-related Groups, and Automatic Data Processing.

Search details: (("International Classification of Diseases"[MAJR] OR "Healthcare Common Procedure Coding System"[MAJR] OR "Current Procedural Terminology"[MAJR] OR "Diagnosis-Related Groups"[MAJR] OR "Automatic Data Processing"[MAJR]) AND "Medical Records Systems, Computerized"[MAJR]) NOT News[ptyp] AND English[Lang]

PubMed's "related" functionality was used for articles that appeared to be particularly relevant as part of our snowball methodology.

The PubMed search was adapted to fit the search capabilities of other databases. All of the terms searched in PubMed were searched as free text in the Association for Computing Machinery (ACM) Digital Library[ii]. Articles indexed to the terms "medical information systems" and “natural language processing" were also searched. CINAHL[iii] (Cumulative Index to Nursing & Allied Health) was searched using the PubMed methodology, replacing "Medical Record Systems, Computerized" with the CINAHL subject heading "Patient Record Systems." The Inspec database[iv] was searched by combining the controlled index term "medical computing" with the classification code for "natural language processing".

Citations in articles selected from the original searches were reviewed for relevancy, and Science Citation Index Expanded[v] was searched to identify additional articles as part of our snowball methodology of surveying for anything relevant which might not appear in the standard sources.

---

[i] http://www.ncbi.nlm.nih.gov/sites/entrez?db=PubMed
[ii] http://portal.acm.org/dl.cfm?coll
iii http://www.cinahl.com/prodsvcs/cinahldb.htm
iv http://www.iee.org/publish/inspec/about/
v http://scientific.thomson.com/products/scie/

Appendix B: Table of Included Citations

Included Citation	Classification Schema	Purpose	System Tested
Aronow DB, J. R. Cooley and S. Soderland. (1995) "Automated identification of episodes of asthma exacerbation for quality measurement in a computer-based medical record." *Proc Annu Symp Comput Appl Med Care.* 309-13.	3-4 point scale	Clinical decision support	FIGLEAF
Aronow DB, S. Soderland, J. M. Ponte, F. Feng, W. B. Croft and W. G. Lehnert. (1995) "Automated classification of encounter notes in a computer based medical record." *Medinfo.* 8 Pt 1 8-12.	Binary	Facilitate retrieval of cases	INQUERY and FIGLEAF
Baldwin, K. B. (2008). "Evaluating healthcare quality using natural language processing." *J Healthc Qual* 30(4): 24-9	Plenary	Collect specific data	NUDIST
Barrows RC Jr, M. Busuioc and C. Friedman. (2000) "Limited parsing of notational text visit notes: ad-hoc vs. NLP approaches." *Proc AMIA Symp.* 51-5.	Plenary	Clinical decision support	MedLee, GDP (Glaucoma Dedicated Parser)
Bashyam, V. and R. K. Taira (2005). "Indexing anatomical phrases in neuro-radiology reports to the UMLS 2005AA." *AMIA Annu Symp Proc*: 26-30.	Classification	Testing NLP techniques	Not named
Bashyam, V., G. Divita, et al. (2007). "A normalized lexical lookup approach to identifying UMLS concepts in free text." *Stud Health Technol Inform* 129(Pt 1): 545-9.	Classification	Testing NLP techniques	Not named
Carrell, D., D. Miglioretti, et al. (2007). "Coding free text radiology reports using the Cancer Text Information Extraction System (caTIES)." *AMIA Annu Symp Proc*: 889.	3-4 point scale	Facilitate retrieval of cases	caTIES
Chapman WW and P. J. Haug. (1999) "Comparing expert systems for identifying chest x-ray reports that support pneumonia." *Proc AMIA Symp.* 216-220.	Binary	Clinical decision support	SymText
Chapman WW, G. F. Cooper, P. Hanbury, B. E. Chapman, L. H. Harrison and M. M. Wagner. (2003) "Creating a text classifier to detect radiology reports describing mediastinal findings associated with inhalational anthrax and other disorders." *Journal of the American Medical Informatics Association.* 10 (5): 494-503.	Binary	Bio-surveillance	IPS - identify patient sets system
Chapman WW, J. N. Dowling and M. M. Wagner. (2004) "Fever detection from free-text clinical records for biosurveillance." *J Biomed Inform.* 37 (2): 120-7.	Binary	Bio-surveillance	3 algorithms: Keyword HP, Keyword CC, CoCo (+NegEx)
Chapman WW, L. M. Christensen, M. M. Wagner, P. J. Haug, O. Ivanov, J. N. Dowling and R. T. Olszewski. (2005) "Classifying free-text triage chief complaints into syndromic categories with natural language processing." *Artif Intell Med.* 33 (1): 31-40.	Plenary	Bio-surveillance	RODS system

Included Citation	Classification Schema	Purpose	System Tested
Chapman WW, M. Fiszman, J. N. Dowling, B. E. Chapman and T. C. Rindflesch. (2004) "Identifying respiratory findings in emergency department reports for biosurveillance using MetaMap." *Medinfo.* 11 (Pt 1): 487-91.	Multiple binary	Bio-surveillance	UMLS MetaMap
Chapman WW, M. Fizman, B. E. Chapman and P. J. Huag. (2001) "A comparison of classification algorithms to automatically identify chest X-ray reports that support pneumonia." *Journal of Biomedical Informatics.* 34 (1): 4-14.	3-4 point scale	Clinical decision support	SymText with 3 different classification algorithms
Chapman WW, W. Bridewell, P. Hanbury, G. F. Cooper and B. G. Buchanan. (2001) "A simple algorithm for identifying negated findings and diseases in discharge summaries." *J Biomed Inform.* 34 (5): 301-10.	Plenary	Clinical decision support	NegEx
Chapman WW, W. Bridewell, P. Hanbury, G. F. Cooper and B. G. Buchanan. (2001) "Evaluation of negation phrases in narrative clinical reports." *Proc AMIA Symp.* 105-9.	Plenary	Clinical decision support	NegEx
Chapman, W. W., J. N. Dowling, et al. (2005). "Classification of emergency department chief complaints into 7 syndromes: a retrospective analysis of 527,228 patients." Annals of Emergency Medicine 46(5): 445	Plenary	Bio-surveillance	CoCo
Chapman, W. W., M. Fiszman, et al. (1999). Correct vs. Parsed Data for Inferring Pneumonia in Chest X-ray Reports. AMIA proceedings (?)	Binary	Clinical decision support	SymText
Chu, D., J. N. Dowling, et al. (2006). "Evaluating the effectiveness of four contextual features in classifying annotated clinical conditions in emergency department reports." *AMIA Annu Symp Proc*: 141-5.	Plenary	Bio-surveillance	SySTR (syndromic surveillance form textual records)
Chuang JH, C. Friedman and G. Hripcsak. (2002) "A comparison of the Charlson comorbidities derived from medical language processing and administrative data." *Proc AMIA Symp.* 160-4.	Classification	Facilitate retrieval of cases	MedLEE
Clark, C., K. Good, et al. (2008). "Identifying smokers with a medical extraction system." *J Am Med Inform Assoc* 15(1): 36-9.	Plenary	Testing NLP techniques	Nuance med extraction system
Cohen, A. M. (2008). "Five-way smoking status classification using text hot-spot identification and error-correcting output codes." *J Am Med Inform Assoc* 15(1): 32-5	Plenary	Testing NLP techniques	Not named
Coles EC and G. Slavin. (1976) "An evaluation of automatic coding of surgical pathology reports." *J Clin Pathol.* 29 (7): 621-5.	Classification	Clinical decision support	NIH and, Northwick Park pilot encoder
Cooper GF and R. A. Miller. (1998) "An experiment comparing lexical and statistical methods for extracting MeSH terms from clinical free text." *J Am Med Inform Assoc.* 5 (1): 62-75.	Classification	Facilitate retrieval of cases	PostDoc, and Pindex

Included Citation	Classification Schema	Purpose	System Tested
Dang, P. A., M. K. Kalra, et al. (2008). "Extraction of recommendation features in radiology with natural language processing: exploratory study." *AJR Am J Roentgenol* 191(2): 313-20.	Plenary	Clinical decision support	Nuance Leximer
Dara, J., J. N. Dowling, et al. (2008). "Evaluation of preprocessing techniques for chief complaint classification." *J Biomed Inform* 41(4): 613-23.	Plenary	Bio-surveillance	RODS (CoCo and KC)
Dinwoodie, H P, and R W Howell. "Automatic disease coding: the 'fruit-machine' method in general practice." *British Journal of Preventive & Social Medicine* 27, no. 1 (February 1973): 59-62	Classification	Admin coding process	Fruit machine
Dreyer KJ, M. K. Kalra, M. M. Maher, A. M. Hurier, B. A. Asfaw, T. Schultz, E. F. Halpern and J. H. Thrall. (2005) "Application of recently developed computer algorithm for automatic classification of unstructured radiology reports: Validation study." *Radiology.* 234 (2): 323-329.	Multiple binary	Clinical decision support	Nuance leximer
Eardley, D. D. and et al. (2000). Using Decision Tree Classifiers to Confirm Pneumonia Diagnosis. AMIA proceedings (?)	Binary	Facilitate retrieval of cases	RuleQuest's See5 data mining tool
Elkin PL, A. P. Ruggieri, S. H. Brown, J. Buntrock, B. A. Bauer, D. Wahner-Roedler, S. C. Litin, J. Beinborn, K. R. Bailey and L. Bergstrom. (2001) "A randomized controlled trial of the accuracy of clinical record retrieval using SNOMED-RT as compared with ICD9-CM." *Proc AMIA Symp.* 159-63.	Classification	Facilitate retrieval of cases	Mayo Vocabulary processor (MVP)
Elkin PL, S. H. Brown, B. A. Bauer, C. S. Husser, W. Carruth, L. R. Bergstrom and D. L. Wahner-Roedler. (2005) "A controlled trial of automated classification of negation from clinical notes." *BMC Med Inform Decis Mak.* 5 (1): 13.	3-4 point scale	Testing NLP techniques	Not named
Elkins JS, C. Friedman, B. Boden-Albala, R. L. Sacco and G. Hripcsak. (2000) "Coding neuroradiology reports for the Northern Manhattan Stroke Study: a comparison of natural language processing and manual review." *Comput Biomed Res.* 33 (1): 1-10.	Plenary	Collect specific data	MedLEE
Ertle AR, E. M. Campbell and W. R. Hersh. (1996) "Automated application of clinical practice guidelines for asthma management." *Proc AMIA Annu Fall Symp.* 552-6.	3-4 point scale	Apply clinical guidelines	Not named
Fiszman M and P. J. Haug. (2000) "Using medical language processing to support real-time evaluation of pneumonia guidelines." *Proc AMIA Symp.* 235-9.	3-4 point scale	Apply clinical guidelines	SymText
Fiszman M, P. J. Haug and P. R. Frederick. (1998) "Automatic extraction of PIOPED interpretations from ventilation/perfusion lung scan reports." *Proc AMIA Symp.* 860-4.	3-4 point scale	Clinical decision support	SymText
Fiszman M, W. W. Chapman, D. Aronsky, R. S. Evans and P. J. Haug. (2000) "Automatic detection of acute bacterial pneumonia from chest X-ray reports." *J Am Med Inform Assoc.* 7 (6): 593-604.	Multiple binary	Clinical decision support	SymText

Included Citation	Classification Schema	Purpose	System Tested
Friedlin, J. and C. J. McDonald (2006). "A natural language processing system to extract and code concepts relating to congestive heart failure from chest radiology reports." *AMIA Annu Symp Proc*: 269-73	Multiple binary	Clinical decision support	REX (Regenstrief extraction tool)
Friedlin, J. and C. J. McDonald (2006). "Using a natural language processing system to extract and code family history data from admission reports." *AMIA Annu Symp Proc*: 925.	Plenary	Clinical decision support	REX
Friedman C, C. Knirsch, L. Shagina and G. Hripcsak. (1999) "Automating a severity score guideline for community-acquired pneumonia employing medical language processing of discharge summaries." *Proc AMIA Symp.* 256-60.	Plenary	Clinical decision support	MedLEE
Friedman C, G. Hripcsak and I. Shablinsky. (1998) "An evaluation of natural language processing methodologies." *Proc AMIA Symp.* 855-9.	Multiple binary	Testing NLP techniques	MedLEE
Friedman C, H. Liu, L. Shagina, S. Johnson and G. Hripcsak. (2001) "Evaluating the UMLS as a source of lexical knowledge for medical language processing." *Proc AMIA Symp.* 189-93.	Plenary	Apply clinical guidelines	MedLEE and UMLS
Friedman C, J. J. Cimino and S. B. Johnson. (1994) "A schema for representing medical language applied to clinical radiology." *J Am Med Inform Assoc.* 1 (3): 233-48.	Classification	Clinical decision support	MED and a Natural language processor
Friedman C, L. Shagina, Y. Lussier and G. Hripcsak. (2004) "Automated encoding of clinical documents based on natural language processing." *J Am Med Inform Assoc.* 11 (5): 392-402.	Classification	Testing NLP techniques	MedLEE
Friedman C, P. O. Alderson, J. H. Austin, J. J. Cimino and S. B. Johnson. (1994) "A general natural-language text processor for clinical radiology." *J Am Med Inform Assoc.* 1 (2): 161-74.	Classification	Clinical decision support	Not named
Goldstein, I., A. Arzrumtsyan, et al. (2007). "Three approaches to automatic assignment of ICD-9-CM codes to radiology reports." *AMIA Annu Symp Proc*: 279-83.	Classification	Admin coding process	Lucene, BoosTexter, & hand crafted rules
Gundersen ML, P. J. Haug, T. A. Pryor, R. vanBree, S. Koehler, K. Bauer and B. Clemons. (1996) "Development and evaluation of a computerized admission diagnoses encoding system." *Computers and Biomedical Research.* 29 (5): 351-372.	Classification	Clinical decision support	NLUSs(Natural Language Understanding Systems)
Haas JP, E. A. Mendonca, B. Ross, C. Friedman and E. Larson. (2005) "Use of computerized surveillance to detect nosocomial pneumonia in neonatal intensive care unit patients." *American Journal of Infection Control.* 33 (8): 439-443.	Classification	Clinical decision support	MedLEE
Haug PJ, D. L. Ranum and P. R. Frederick. (1990) "Computerized extraction of coded findings from free-text radiologic reports. Work in progress." *Radiology.* 174 (2): 543-8.	Classification	Clinical decision support	SPRUS (Special Purpose Radiology Understanding System)

Included Citation	Classification Schema	Purpose	System Tested
Hazlehurst, B., D. F. Sittig, et al. (2005). "Natural language processing in the electronic medical record: assessing clinician adherence to tobacco treatment guidelines." *Am J Prev Med* 29(5): 434-9	Plenary	Collect specific data	MediClass
Hazlehurst, B., J. Mullooly, et al. (2005). "Detecting possible vaccination reactions in clinical notes." *AMIA Annu Symp Proc*: 306-10	Binary	Bio-surveillance	MediClass
Heinze, D. T., M. L. Morsch, et al. (2008). "Medical i2b2 NLP smoking challenge: the A-Life system architecture and methodology." *J Am Med Inform Assoc* 15(1): 40-3.	Plenary	Testing NLP techniques	A-Life medical's CM-Extractor
Hersh W, M. Mailhot, C. Arnott-Smith and H. Lowe. (2001) "Selective automated indexing of findings and diagnoses in radiology reports." *J Biomed Inform.* 34 (4): 262-73.	Classification	Facilitate retrieval of cases	SAPHIRE
Hripcsak G, C. A. Knirsch, N. L. Jain and A. Pablos-Mendez. (1997) "Automated tuberculosis detection." *J Am Med Inform Assoc.* 4 (5): 376-81.	3-4 point scale	Bio-surveillance	MedLEE
Hripcsak G, C. Friedman, P. O. Alderson, W. DuMouchel, S. B. Johnson and P. D. Clayton. (1995) "Unlocking clinical data from narrative reports: a study of natural language processing." *Ann Intern Med.* 122 (9): 681-8.	Multiple binary	Clinical decision support	MedLEE
Hripcsak G, G. J. Kuperman and C. Friedman. (1998) "Extracting findings from narrative reports: software transferability and sources of physician disagreement." *Methods Inf Med.* 37 (1): 1-7.	Multiple binary	Testing NLP techniques	MedLEE
Hripcsak G, J. H. Austin, P. O. Alderson and C. Friedman. (2002) "Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports." *Radiology.* 224 (1): 157-63.	Classification	Facilitate retrieval of cases	MedLEE
Huang Y, H. J. Lowe and W. R. Hersh. (2003) "A pilot study of contextual UMLS indexing to improve the precision of concept-based representation in XML-structured clinical radiology reports." *J Am Med Inform Assoc.* 10 (6): 580-7.	Classification	Testing NLP techniques	SAPHIRE
Ivanov, O., M. M. Wagner, et al. (2002). "Accuracy of three classifiers of acute gastrointestinal syndrome for syndromic surveillance." Proceedings / AMIA ... Annual Symposium. AMIA Symposium: 345	Plenary	Bio-surveillance	Not named
Jain NL and C. Friedman. (1997) "Identification of findings suspicious for breast cancer based on natural language processing of mammogram reports." *Proc AMIA Annu Fall Symp.* 829-33.	Classification	Clinical decision support	MedLEE
Jain NL, C. A. Knirsch, C. Friedman and G. Hripcsak. (1996) "Identification of suspected tuberculosis patients based on natural language processing of chest radiograph reports." *Proc AMIA Annu Fall Symp.* 542-6.	Binary	Apply clinical guidelines	MedLEE
Kukafka, R., M. E. Bales, et al. (2006). "Human and automated coding of rehabilitation discharge summaries according to the International Classification of Functioning, Disability, and Health." *J Am Med Inform Assoc* 13(5): 508-15	Classification	Admin coding process	MedLEE

Included Citation	Classification Schema	Purpose	System Tested
Levin, M. A., M. Krol, et al. (2007). "Extraction and mapping of drug names from free text to a standardized nomenclature." *AMIA Annu Symp Proc*: 438-42.	Classification	Clinical decision support	Not named
Liu, K., K. J. Mitchell, et al. (2005). "Automating tissue bank annotation from pathology reports - comparison to a gold standard expert annotation set." *AMIA Annu Symp Proc*: 460-4	Plenary	Collect specific data	GATE
Long, W. (2007). "Lessons extracting diseases from discharge summaries." *AMIA Annu Symp Proc*: 478-82	Classification	Automate problem lists	Not named
Lowe HJ, I. Antipov, W. Hersh, C. A. Smith and M. Mailhot. (1999) "Automated semantic indexing of imaging reports to support retrieval of medical images in the multimedia electronic medical record." *Methods Inf Med.* 38 (4-5): 303-7.	Classification	Facilitate retrieval of cases	SAPHIRE
Lu, H. M., D. Zeng, et al. (2008). (Lu, Zeng, Trujillo, Komatsu, & Chen, 2008) "Ontology-enhanced automatic chief complaint classification for syndromic surveillance." *J Biomed Inform* 41(2): 340-56	Plenary	Bio-surveillance	JESS inference engine, with WSSS
Lussier YA, L. Shagina and C. Friedman. (2001) "Automating SNOMED coding using medical language understanding: a feasibility study." *Proc AMIA Symp.* 418-22.	Classification	Testing NLP techniques	MedLEE
Lussier, YA L. Shagina and C. Friedman. (2000) "Automating ICD-9-CM encoding using medical language processing: A feasibility study." *Journal of the American Medical Informatics Association.* 1072-1072.	Classification	Admin coding process	MedLEE
Mamlin BW, D. T. Heinze and C. J. McDonald. (2003) "Automated extraction and normalization of findings from cancer-related free-text radiology reports." *AMIA Annu Symp Proc.* 420-4.	Plenary	Collect specific data	LIfeCode
McCowan, I. A., D. C. Moore, et al. (2007). "Collection of cancer stage data by classifying free-text medical reports." *J Am Med Inform Assoc* 14(6): 736-45	Plenary	Collect specific data	CSIS (ca stage interpretation system)
McKenzie K, S. Walker and S. Tong. (2001) "Assessment of the impact of the change from manual to automated coding on mortality statistics in Australia." *Health Information Management Journal.* 30 (3):	Classification	Collect specific data	ACS (Automated Coding System)
Melton GB and G. Hripcsak. (2005) "Automated detection of adverse events using natural language processing of discharge summaries." *J Am Med Inform Assoc.* 12 (4): 448-57.	Classification	Clinical decision support	MedLEE
Mendonca EA, J. Haas, L. Shagina, E. Larson and C. Friedman. (2005) "Extracting information on pneumonia in infants using natural language processing of radiology reports." *Journal of Biomedical Informatics.* 38 (4): 314-321.	Binary	Bio-surveillance	MedLEE
Meystre S and P. J. Haug. (2005) "Evaluation of Medical Problem Extraction from Electronic Clinical Documents Using MetaMap Transfer (MMTx)." *Stud Health Technol Inform.* 116 823-8.	Plenary	Automate problem lists	MMTX + NegEx

Included Citation	Classification Schema	Purpose	System Tested
Meystre, S. and P. Haug (2006). "Improving the sensitivity of the problem list in an intensive care unit by using natural language processing." *AMIA Annu Symp Proc*: 554-8	Plenary	Automate problem lists	MMTx
Meystre, S. and P. J. Haug (2005). "Automation of a problem list using natural language processing." *BMC Med Inform Decis Mak* 5: 30.	Plenary	Automate problem lists	MMTx; MPLUS2; keyword srch
Meystre, S. and P. J. Haug (2006). "Natural language processing to extract medical problems from electronic clinical documents: performance evaluation." *J Biomed Inform* 39(6): 589-99	Plenary	Automate problem lists	MMTx. NegEx
Mitchell KJ, M. J. Becich, J. J. Berman, W. W. Chapman, J. Gilbertson, D. Gupta, J. Harrison, E. Legowski and R. S. Crowley. (2004) "Implementation and evaluation of a negation tagger in a pipeline-based system for information extract from pathology reports." *Medinfo.* 11 (Pt 1): 663-7.	Classification	Facilitate retrieval of cases	NegEx
Moore GW and J. J. Berman. (1994) "Performance analysis of manual and automated systemized nomenclature of medicine (SNOMED) coding." *Am.J.Clin.Pathol.* 101 (3): 253-256.	Classification	Facilitate retrieval of cases	TRANSOFT with context-insensitive and context-sensitive model
Morris WC, D. T. Heinze, H. R. Warner Jr, A. Primack, A. E. Morsch, R. E. Sheffer, M. A. Jennings, M. L. Morsch and M. A. Jimmink. (2000) "Assessing the accuracy of an automated coding system in emergency medicine." *Proc.AMIA.Symp.* 595-599.	Classification	Admin coding process	A-Life's LifeCode
Morsch, M. L., J. L. Vengco, et al. (2006). (Morsch, Vengco, Sheffer, & Heinze, 2006) "CM-Extractor: An Application for Automating Medical Quality Measures Abstraction in a Hospital Setting".	Plenary	Reporting quality measures	A-Life's CM-extractor
Mutalik PG, A. Deshpande and P. M. Nadkarni. (2001) "Use of general-purpose negation detection to augment concept indexing of medical documents: a quantitative study using the UMLS." *J Am Med Inform Assoc.* 8 (6): 598-609.	Classification	Facilitate retrieval of cases	Negfinder
Nadkarni P, R. Chen and C. Brandt. (2001) "UMLS concept indexing for production databases: a feasibility study." *J Am Med Inform Assoc.* 8 (1): 80-91.	Classification	Facilitate retrieval of cases	Not named
Oliver DE and R. B. Altman. (1994) "Extraction of SNOMED concepts from medical record texts." *Proc Annu Symp Comput Appl Med Care.* 179-83.	Classification	Clinical decision support	Not named
Pakhomov SV, A. Ruggieri and C. G. Chute. (2002) "Maximum entropy modeling for mining patient medication status from free text." *AMIA 2002 Symposium. Bio medical Informatics: One Discipline. Annual Symposium of the American Medical Informatics Association. Proceedings.* 587-591.	3-4 point scale	Testing NLP techniques	Not named

Included Citation	Classification Schema	Purpose	System Tested
Pakhomov SV, J. Buntrock and C. G. Chute. (2005) "Prospective recruitment of patients with congestive heart failure using an ad-hoc binary classifier." *J Biomed Inform.* 38 (2): 145-53.	Classification	Facilitate retrieval of cases	2 algorithms: Naïve Bayes & Perceptron Neural Network
Pakhomov SV, J. D. Buntrock and C. G. Chute. (2004) "Using compound codes for automatic classification of clinical diagnoses." *Medinfo.* 11 (Pt 1): 411-415.	Classification	Testing NLP techniques	Not named
Pakhomov, S. V., J. D. Buntrock, et al. (2006). "Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques." *J Am Med Inform Assoc* 13(5): 516-25	Classification	Admin coding process	automatic medical index classification architecture
Pakhomov, S., S. A. Weston, et al. (2007). "Electronic medical records for clinical research: application to the identification of heart failure." *Am J Manage Care* 13(6 Part 1): 281-8	Binary	Facilitate retrieval of cases	Not named
Pestian, J. P., C. Brew, et al. (2007). (Pestian et al., 2007) "A shared task involving multi-label classification of clinical free text." *ACL 2007*: 97	Classification	Testing NLP techniques	NA (50 diff systems)
Pyrros, A., P. Nikolaidis, et al. (2007). "A Bayesian approach for the categorization of radiology reports." *Acad Radiol* 14(4): 426-30	Binary	Facilitate retrieval of cases	Not named
Reichley, R. M., K. E. Henderson, et al. (2007). "Natural language processing to identify venous thromboembolic events." *AMIA Annu Symp Proc*: 1089.	Binary	Reporting quality measures	Not named
Rosenberg KM and D. B. Coultas. (1994) "Acceptability of Unified Medical Language System terms as substitute for natural language general medicine clinic diagnoses." *Proc Annu Symp Comput Appl Med Care.* 193-7.	Classification	Clinical decision support	UMLS Search Engine (USE)
Saad, F. H., G. D. Bell, et al. (2008). "Classification techniques with minimal labelling effort and application to medical reports." *Int J Data Min Bioinform* 2(3): 268-87	Binary	Testing NLP techniques	EM and SVM
Sager N, M. Lyman, C. Bucknall, N. Nhan and L. J. Tick. (1994) "Natural language processing and the representation of clinical data." *J Am Med Inform Assoc.* 1 (2): 142-60.	Plenary	Clinical decision support	LSP-MLP
Savova, G. K., P. V. Ogren, et al. (2008). "Mayo clinic NLP system for patient smoking status identification." *J Am Med Inform Assoc* 15(1): 25-8	Plenary	Testing NLP techniques	unstructured information management architecture (UIMA)
Schadow G and C. J. McDonald. (2003) "Extracting structured information from free text pathology reports." *AMIA Annu Symp Proc.* 584-8.	Classification	Facilitate retrieval of cases	MMTx (NLM's)
Sinha U, B. Dai, D. B. Johnson, R. Taira, J. Dionisio, G. Tashima, M. Golamco and H. Kangarloo. (2000) "Interactive software for generation and visualization of structured findings in radiology reports." *AJR Am J Roentgenol.* 175 (3): 609-12.	Classification	Clinical decision support	Not named

Included Citation	Classification Schema	Purpose	System Tested
Thomas BJ, H. Ouellette, E. F. Halpern and D. I. Rosenthal. (2005) "Automated computer-assisted categorization of radiology reports." *American Journal of Roentgenology.* 184 (2): 687-690.	3-4 point scale	Clinical decision support	Not named - Commercially available software
Travers DA and S. W. Haas. (2004) "Evaluation of emergency medical text processor, a system for cleaning chief complaint text data." *Acad Emerg Med.* 11 (11): 1170-6.	Classification	Testing NLP techniques	EMT-P (Emergency Medical Text Processor)
Travers, D., S. Wu, et al. (2007). "Evaluation of a chief complaint pre-processor for biosurveillance." *AMIA Annu Symp Proc*: 736-40	Classification	Bio-surveillance	EMT-P
Trick WE, W. W. Chapman, M. F. Wisniewski, B. J. Peterson, S. L. Solomon and R. A. Weinstein. (2003) "Electronic interpretation of chest radiograph reports to detect central venous catheters." *Infection Control and Hospital Epidemiology.* 24 (12): 950-4.	Binary	Clinical decision support	SymText
Turchin A, I. S. Kohane and M. L. Pendergrass. (2005) "Identification of Patients With Diabetes From the Text of Physician Notes in the Electronic Medical Record." *Diabetes Care.* 28 (7): 1794-1795.	Binary	Facilitate retrieval of cases	DITTO
Turchin, A., M. L. Pendergrass, et al. (2005). DITTO: a Tool for Identification of Patient Cohorts from the Text of Physician Notes in the Electronic Medical Record. AMIA Annual Symposium Proceedings, Massachusetts Institute of Technology. 2005: 744-748	Multiple binary	Facilitate retrieval of cases	DITTO
Turchin, A., N. S. Kolatkar, et al. (2006). "Using regular expressions to abstract blood pressure and treatment intensification information from the text of physician notes." *J Am Med Inform Assoc* 13(6): 691-5	Plenary	Facilitate retrieval of cases	Not named
Uzuner, O., I. Goldstein, et al. (2008). "Identifying patient smoking status from medical discharge records." *J Am Med Inform Assoc* 15(1): 14-24	Plenary	Reporting quality measures	NA (11 teams with different systems participated)
Warner HR, Jr. (2000) "Can natural language processing aid outpatient coders?" *J AHIMA.* 71 (8): 78-81; quiz 83-4.	Classification	Admin coding process	Not named
Wilcox A and G. Hripcsak. (1998) "Knowledge discovery and data mining to assist natural language understanding." *Proc AMIA Symp.* 835-9.	Multiple binary	Clinical decision support	MedLEE (C5.0 using MedLEE output)
Wilcox A and G. Hripcsak. (1999) "Classification algorithms applied to narrative reports." *Proc.AMIA.Symp.* 455-459.	Multiple binary	Clinical decision support	MedLEE (various algorithms applied to MedLEE output)
Wilcox A and G. Hripcsak. (2000) "Medical text representations for inductive learning." *Proc AMIA Symp.* 923-7.	Multiple binary	Clinical decision support	MedLEE (various algorithms applied to MedLEE output)

Included Citation	Classification Schema	Purpose	System Tested
Wilcox AB and G. Hripcsak. (2003) "The role of domain knowledge in automating medical text report classification." *J Am Med Inform Assoc.* 10 (4): 330-8.	Multiple binary	Clinical decision support	MedLEE (various algorithms applied to MedLEE output)
Xu H, K. Anderson, V. R. Grann and C. Friedman. (2004) "Facilitating cancer research using natural language processing of pathology reports." *Medinfo.* 11 (Pt 1): 565-572.	Plenary	Collect specific data	MedLEE
Zeng, Q. T., S. Goryachev, et al. (2006). "Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system." *BMC Med Inform Decis Mak* 6: 30	Plenary	Facilitate retrieval of cases	Health information text Extraction (HITEx)
Zingmond D and L. A. Lenert. (1993) "Monitoring free-text data using medical language processing." *Comput Biomed Res.* 26 (5): 467-81.	Binary	Clinical decision support	RadTRAC