Natural Language Processing in Radiology: A Systematic Review[1]

Ewoud Pons, MD
Loes M. M. Braun, MD, PhD
M. G. Myriam Hunink, MD, PhD
Jan A. Kors, PhD

Radiological reporting has generated large quantities of digital content within the electronic health record, which is potentially a valuable source of information for improving clinical care and supporting research. Although radiology reports are stored for communication and documentation of diagnostic imaging, harnessing their potential requires efficient and automated information extraction: they exist mainly as free-text clinical narrative, from which it is a major challenge to obtain structured data. Natural language processing (NLP) provides techniques that aid the conversion of text into a structured representation, and thus enables computers to derive meaning from human (ie, natural language) input. Used on radiology reports, NLP techniques enable automatic identification and extraction of information. By exploring the various purposes for their use, this review examines how radiology benefits from NLP. A systematic literature search identified 67 relevant publications describing NLP methods that support practical applications in radiology. This review takes a close look at the individual studies in terms of tasks (ie, the extracted information), the NLP methodology and tools used, and their application purpose and performance results. Additionally, limitations, future challenges, and requirements for advancing NLP in radiology will be discussed.

Online supplemental material is available for this article.

[1] From the Departments of Radiology (E.P., L.M.M.B., M.G.M.H.) and Medical Informatics (J.A.K.), Erasmus Medical Center, PO Box 2040, 3000 CA Rotterdam, the Netherlands. Received December 19, 2014; revision requested January 16, 2015; final revision received March 31; accepted April 13; final version accepted April 24; final review December 14. Supported by the Open Technology Programme of Technology Foundation STW. **Address correspondence to** E.P. (e-mail: *e.pons@erasmusmc.nl*).

The rise of the electronic health record (EHR) is generating new challenges and opportunities in the medical domain, and the increasing use of digital content, both structured data and narrative text, is expected to offer many benefits. As well as the original purpose of improving clinical care through high-quality documentation, EHR data can make more extensive contributions to research and clinical workflows (1). Harnessing the potential of clinical narrative in the EHR, including radiologic reporting, will require strategies for efficient and automated information extraction.

As the formal product of a diagnostic imaging referral, the radiology report is used for communication and documentation purposes. There are various guidelines for effective reporting of diagnostic imaging (2,3), although essentially a report consists of free text, organized in a number of standard sections. Due to the unstructured, free-text nature of these reports, their conversion into a computer manageable representation is a major challenge. Techniques for doing so are provided by natural language processing (NLP), which can convert unstructured text into a structured form, and therefore enable automatic identification and extraction of information. For example, such structured output can be the classification of patients in different groups or the codes from a clinical coding system. The terms *text mining* and *information extraction* are also commonly used to denote the task of NLP (4).

In an earlier review, Meystre et al (5) covered text mining of clinical narratives in general. They discussed a wide range of NLP techniques and objectives for extracting information from clinical texts. Demner-Fushman et al (6) described methods and systems for clinical decision support and discussed the contribution that NLP can make. Stanfill et al (7) systematically reviewed automated clinical coding and classification systems. NLP techniques were not the primary focus of their review, and the authors limited their scope to systems with predefined classes as output. None of these reviews on the use of NLP in EHRs is both specific to radiology and systematic. The objective of this review is to provide a systematic, up-to-date overview of NLP applications in radiology, focusing on the performance, benefits, and current limitations.

Essentials

- Natural language processing (NLP) applications are key to obtaining structured information from radiology reports and have been developed for many different purposes.
- Through automation, NLP applications can process large amounts of data and bring new functionality to clinical workflows.
- Performance of NLP systems is generally high, but not many applications are actually being used in routine clinical practice or research.
- Proliferation of NLP applications in radiology may improve by establishing performance requirements, report standardization, and external validation.

Background

To identify and extract information from unstructured text, NLP applications rely on a sequence of steps that produces structured textual features from the radiology report. We consider a broad definition of NLP that includes techniques for both generation and subsequent processing of these features (8). Common components of an NLP pipeline are illustrated in Figure 1, although depending on the application purpose only a subset may be used.

In a first preprocessing step, radiology reports are split into their respective sections. Successive processing steps can use a subset of sections (eg, focus only on the reports' impression) or assign specific weights to the content of different sections. The text is then further divided into sentences (sentence splitting) and words (tokenization). On the word level, additional normalization steps determine the lexical root of words (stemming), fix spelling mistakes, and expand abbreviations to their full form. Subsequent syntactic analysis determines the part of speech of words (eg, noun, verb, adjective), their grammatical structure (eg, noun phrase, verb phrase, prepositional phrase), or dependency relations (subject of or object of) (8). Semantic analysis assigns meaning to the words and phrases by linking them to semantic types (eg, symptom, disease, procedure) and concepts. A lexicon of words with definitions and synonyms may be used for this purpose, for example, the Unified Medical Language System metathesaurus (9) or RadLex® (10), a specialized radiologic lexicon. Semantic relations can be derived from ontologies, which are specialized lexicons containing concepts and relations between them. The negation detection step checks whether concepts or relations in the text are negated.

The combined result of all previous steps in the pipeline produces the NLP features. These features are subsequently used to solve the system's task, for instance, text classification or information extraction. To accomplish this, structured textual features can be processed by an automatically generated classifier (ie, machine-learning approach) or be combined in rules hand-crafted by experts (ie, rule-based approach). A hybrid approach that combines the machine-learning and rule-based approaches can also be used (eg, manually crafted rules are used to correct errors of an automatic classifier).

Published online
10.1148/radiol.16142770 **Content codes:** IN SQ

Radiology 2016; 279:329–343

Abbreviations:
BI-RADS = Breast Imaging-Reporting and Data System
EHR = electronic health record
ICD-9 = International Classification of Diseases, version 9
NLP = natural language processing
PPV = positive predictive value

Conflicts of interest are listed at the end of this article.

Figure 1

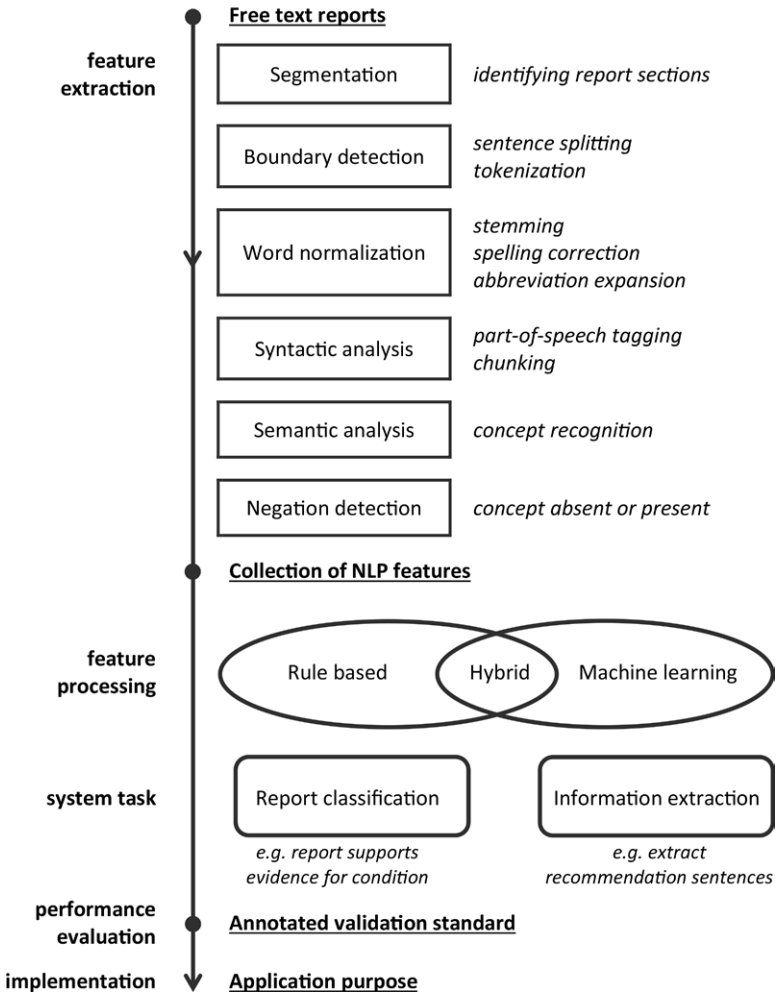

Figure 1: Natural language processing pipeline.

NLP applications are generally trained and validated on a reference set, that is, a set of reports that were manually annotated by one or more experts in clinical radiology for the outcome of interest. Often the annotated outcomes are binary (eg, whether the condition is present, the finding is actionable, or the report contains relevant recommendations), but sometimes the outcomes are spans of text (eg, relevant parts of sentences, specific concepts, or quantitative values). The reference set is often split into a training and validation set. The validation set is withheld during training and only used to assess the performance of the application. Another common approach is the use of cross-validation, where the reference set is split in a number of subsets. The algorithm is iteratively trained, leaving out a different subset for validation in each round. Therefore, all data are used for training, while the validation results are averaged over the multiple rounds of cross-validation.

Search Strategy

We conducted a search to identify all potentially relevant publications about NLP applications in radiology. The MEDLINE (11) and EMBASE (12) databases were queried for articles indexed up to October 7, 2014. The search was performed by using the following free-text keywords: "natural language processing," "natural language understanding," "medical language processing," "NLP," "MLP," "information extraction," or "text mining," in combination with key terms that limit the results to the radiologic domain: "radiology," "radiologic," or "radiological." The MEDLINE query was expanded with the Medical Subject Headings index terms "radiology," "radiology information systems," and "natural language processing." We limited the search to articles in English. The exact query syntax is provided in the Appendix E1 (online).

Study Inclusion and Data Extraction

All publications resulting from the search were independently assessed by two authors (E.P., medical doctor with 6 months of experience in clinical radiology; J.A.K., medical informatician with 15 years of experience in NLP). The main requirement for inclusion was the description and evaluation of an NLP method or tool yielding a practical application in radiology. Publications were excluded if no full text could be retrieved or if they were published in a journal without an assigned impact factor in the 2015 Journal Citation Report science edition (13). Inclusion was based on title or abstract, although the full-text article was assessed when any of the inclusion criteria remained ambiguous. Disagreements in the inclusion process were resolved by consensus discussion (E.P., J.A.K.). We hand-searched citation lists of studies for publications not retrieved by our electronic search and included additional studies fulfilling all the criteria.

In a second stage, two authors (E.P., J.A.K.) independently assessed the full-text articles. For each included article, the following data were extracted: *(a)* task of the NLP system (ie, what information is identified), *(b)* application purpose or use case (ie, why the information is identified), *(c)* used NLP methodology and tools, *(d)* evaluation measures and performance results, *(e)* size of validation dataset and prevalence of identified outcome, and *(f)* operational use (ie, whether the system was actually used after development).

Table 1

Performance Measures

Measure	Also Known As	Formula*
Sensitivity	Recall, true-positive rate	TP/(TP + FN)
Specificity	True-negative rate	TN/(TN + FP)
PPV	Precision	TP/(TP + FP)
F score	F measure, F1 score	2 × (Precision × Recall)/(Precision + Recall)
Accuracy	Not applicable	(TP + TN)/(TP + FP + TN + FN)

* TP = true-positive, FP = false-positive, FN = false-negative, TP = true-negative.

The review was conducted according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses, or PRISMA, statement for systematic reviews (14). Risk of bias in individual studies was not assessed, because relevant quality indicators could not be identified.

Performance Measures

Performance measures that have been used in the studies in this review include sensitivity (also called recall in the field of NLP), specificity, positive predictive value (PPV) (also called precision), *F* score (harmonized average of recall and precision) (15), and accuracy. When describing the performance of a method, we focus on sensitivity and specificity; PPV is referred to if specificity is unavailable. Some studies provide the *F* score, which is frequently used in the field of NLP as a single, overall measure of system performance. Nomenclature and formulas are listed in Table 1. It should be noted that the performance measures pertain to the performance of the systems in identifying and extracting relevant information from radiology reports and do not reflect the accuracy of the diagnostic reporting with respect to underlying pathologic conditions or clinical diagnosis.

Data Representation and Data Analysis

Due to the heterogeneity and multidisciplinary nature of the included studies, a formal meta-analysis was not possible. We did, however, visually determine overall performance by representing the sensitivity and specificity of individual studies in receiver operating characteristic space (16). If sensitivity or specificity were not reported, we tried to compute them based on the available information (eg, a contingency table or by deriving specificity from a given sensitivity, PPV, and prevalence). If a study tested multiple methods for a particular application, we report the performance of the best performing method (and took averages for studies identifying multiple conditions).

The MEDLINE and EMBASE queries yielded 266 records after removal of duplicates. Through eligibility screening, 67 studies were selected for detailed review. Figure 2 illustrates the inclusion process. The first included study was published in 1993, and the publication rate increased exponentially over time (Fig 3).

On the basis of the extracted application purpose, we grouped the 67 studies in five broad categories that represent different relevant purposes: diagnostic surveillance (n = 17), cohort building for epidemiologic studies (n = 18), query-based case retrieval (n = 7), quality assessment of radiologic practice (n = 15), and clinical support services (n = 10). We will discuss the studies by application category and are aware that studies in different categories may overlap in terms of methodology and tools. Table 2 gives a summary of the NLP tools, lexicons, and other resources that were utilized in the included studies.

Diagnostic Surveillance

Seventeen studies described the automated detection of critical observations for surveillance (17–33). Applications in this category raise alerts for the occurrence of predetermined findings or conditions that have been reported but not acted on. Such alert systems add safeguards to clinical practice and potentially reduce the chance of critical observations being overlooked by clinicians. Conditions that have been targeted include appendicitis (17), acute lung injury (19), pneumonia (27–30), thromboembolic diseases (33), or various potentially malignant lesions (20–22).

In an acute medical setting, surveillance can be applied to minimize delay in communication between the radiologist and the referring clinician. Rink et al (17) recently developed a method using all individual statements from a report to identify appendicitis. They exploited a hybrid approach involving a customized lexicon, manually defined patterns and machine learning (support vector machine), and achieved a sensitivity of 91% and PPV of 83%. Lakhani et al (18) used a rule-based system, including various normalization steps and negation detection, to detect a range of acute conditions from the impression section of a report. The system was tuned to each condition, for example, achieving 99% sensitivity and 89% PPV for appendicitis cases. Overall, sensitivity was 96% and PPV was 91%. Solti et al (19) experimented on chest x-ray reports to detect acute lung injury. They compared performance of a rule-based system based on expert-derived keywords and a machine-learning algorithm (maximum entropy with variable-length character combinations) and found machine learning to be superior (91% sensitivity, 90% PPV). The authors observed that the reports rarely mention acute lung injury explicitly and demonstrated that an NLP system can detect an implicit diagnosis.

Diagnostic surveillance using NLP can support the management of cases that require follow-up by automatically generating alerts for subsequent examinations or procedures. Zingmond et al (20) as early as 1993 tried to detect reports with actionable findings. A machine-learning classifier, exploiting key

Figure 2

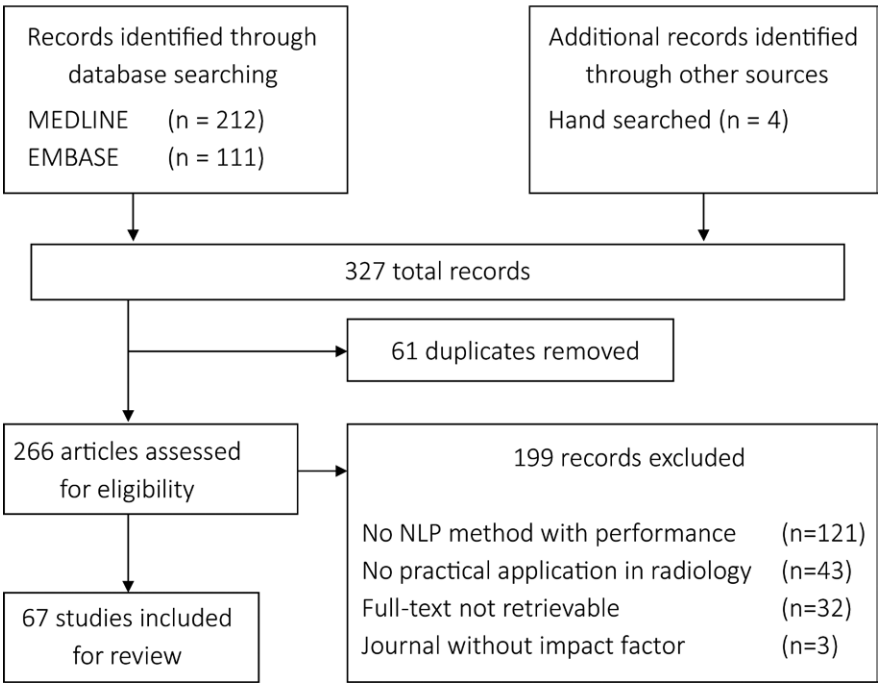

Figure 2: Flow diagram of the literature review process.

Figure 3

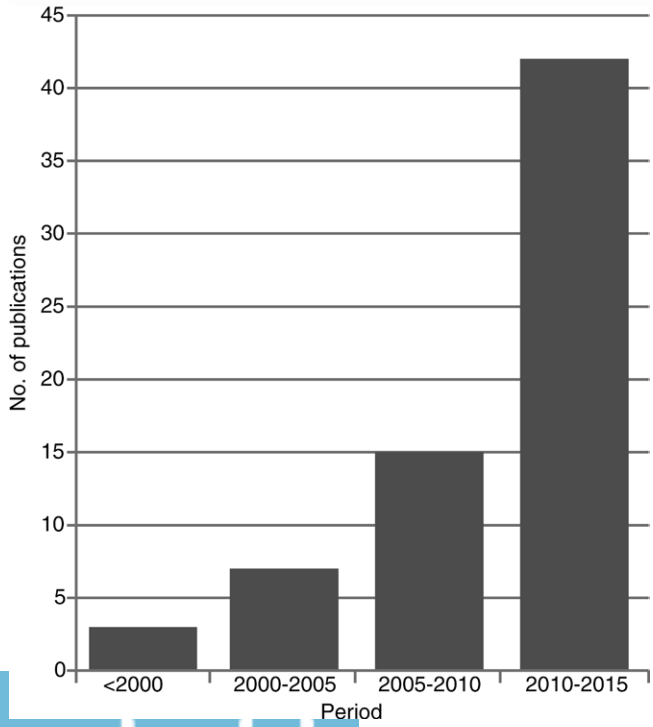

Figure 3: Publication period of included studies.

report phrases and semantic analysis, scored 98% sensitivity and 88% specificity. Garla et al (21) developed a system to identify potentially malignant liver lesions to help ensure the timely and appropriate diagnostic workup of patients suspected of having cancer. Structured output generated by the NLP system cTAKES (clinical Text Analysis and Knowledge Extraction System) was processed with different machine-learning techniques. The highest sensitivity obtained was 98%, which is excellent for surveillance purposes that have to limit the number of missed cases. The corresponding specificity was only 24%; another machine-learning algorithm yielded a more balanced result of 94% sensitivity and 65% specificity.

The importance of temporal context for surveillance was emphasized by Cheng et al (22), who showed the utility of NLP in tracking disease progression. They managed to determine if a tumor was stable or showed regression or progression, with a sensitivity of 81% and specificity of 92%.

Other surveillance applications aim to extract sentences with recommendations for additional imaging from a report. This approach is less specific to monitored condition and imaging modality, because the radiologist's literal advice is detected (eg, "follow-up CT is recommended in three months"), independent of observations. Extracting such recommendations is not a classification task with true-negative results, thus specificity cannot be determined and *F* scores are used as an overall performance measure. The performance of systems that extract recommendations ranged from 87% to 95% *F* score (23–26). Xu et al (26) also extracted the follow-up details (87% *F* score) and the recommended time intervals (98% *F* score). They used exact matching for evaluation (ie, a true-positive was only counted if the extracted span of text exactly matched the reference), whereas the other studies of this kind were limited to detecting whether a sentence contained any recommendation.

In several studies surveillance acted as a fail-safe for the detection of

Table 2

NLP Resources Used in Radiology

Resource*	Description	Web Site
BROK	Java-based information extraction program that determines the BI-RADS final assessment categories	*http://www.brighamandwomens.org/Research/labs/cebi/BROK/default.aspx*
cTAKES	Open-source NLP system for information extraction from electronic medical record clinical free text	*http://ctakes.apache.org/*
DataScout (currently 3M CodeRyte CodeAssist System)	Commercially available product that codes radiology reports to facilitate billing and revenue management	*http://solutions.3m.com/wps/portal/3M/en_US/Health-Information-Systems/HIS/Products-and-Services/Computer-Assisted-Coding/*
dtSearch	Commercially available search engine and index generator	*http://www.dtsearch.com/PLF_Features_2.html*
GATE	Java suite of tools for NLP tasks, including an information extraction system (ANNIE)	*https://gate.ac.uk/*
I2E	Commercially available text-mining software offering NLP-based querying of unstructured text sources	*http://www.linguamatics.com/welcome/software/I2E.html*
iSCOUT	Toolkit that utilizes ontologies to retrieve radiology reports with specific findings	*http://sourceforge.net/projects/iscout/*
LEXIMER	NLP engine that extracts, structures, and classifies unstructured radiology reports, licensed by Nuance Communications	*http://www.nuance.com/index.htm*
LifeCode	Commercially available product that encodes clinical narrative reports for billing purposes	*http://www.optum360.com/hospital/coding-documentation/clinical-documentation-improvement.html*
MALLET	Java-based package for statistical NLP, document classification, clustering, topic modeling, information extraction, and other machine-learning applications to text	*http://mallet.cs.umass.edu/*
MedLEE	NLP tool for medical domain that can extract, structure, and encode clinical information in textual patient reports	*http://dx.doi.org/10.1136/jamia.1994.95236146*
Metamap	NLP tool that maps biomedical text to UMLS concepts	*http://metamap.nlm.nih.gov/*
ONYX	Open-source NLP system that integrates knowledge about syntax and semantics to interpret free text, can be trained on documents from a particular domain	*http://aclweb.org/anthology/W09-1303*
OpenNLP	Machine-learning–based toolkit for the processing of natural language text	*https://opennlp.apache.org/*
RadLex®	Lexicon for standardized indexing and retrieval of radiology information resources	*http://www.radlex.org/*
Render	Online searchable radiology study repository, allows for query-based retrieval of reports and images	Not publicly available
SAPHIRE	An information retrieval system featuring concept matching, automatic indexing, probabilistic retrieval, and hierarchical relationships	*http://dx.doi.org/10.1016/0010-4809(90)90031-7*
SymText	NLP tool for medical domain, integrating syntactic and semantic analysis	*http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2579100/*
UIMA	Component software architecture for the analysis of unstructured information	*https://uima.apache.org/*
UMLS metathesaurus	Compendium of controlled vocabularies and classification systems in the biomedical sciences	*http://www.nlm.nih.gov/research/umls/*
YTEX	Yale cTAKES extensions: clinical NLP, semantic similarity, data mining, feature engineering	*https://code.google.com/p/ytex/*

* BROK = BI-RADS Observation Kit, BI-RADS = Breast Imaging-Reporting and Data System, cTAKES = Apache clinical Text Analysis and Knowledge Extraction System, GATE = General Architecture for Text Engineering, LEXIMER = Lexicon Mediated Entropy Reduction, MedLEE = Medical Language Extraction and Encoding System, UMLS = Unified Medical Language System, UIMA = Unstructured Information Management applications.

incidental findings, that is, findings that are not indicated by the referring clinician and therefore have a higher risk of remaining unattended (21,24,25,33).

In a more traditional biosurveillance application, the detection of hospital-acquired pneumonia in neonates was automated for the purpose of infection management (27,28). The rule-based system, processing features extracted by MedLEE (Medical Language Extraction and Encoding System), identified five out of seven neonates who had been flagged by an infection control professional (71% sensitivity and 95% specificity) (28).

Studies in this application category are generally performed on retrospective data. Only one study evaluated a real-life surveillance system that was deployed during the 2002 Winter

Olympic Games (29). The system monitored various events of public-health interest by applying decision rules, using both NLP features from chest radiographs and other structured EHR data. The system detected one public-health event that exceeded a predefined control limit. This study illustrates that automated systems not only have potential for managing individual patients, but can also monitor public-health–related trends on a hospital or population level. A prerequisite of such applications is the integration of patient data across departments and institutes. Important barriers for such integration are related to the lack of a common data model and privacy issues. Large-scale collection of longitudinal EHR data from heterogeneous sources may also advance proliferation of clinical decision support and personalized medicine, as discussed for instance by Jensen et al (1) and Demner-Fushman et al (6).

Cohort Building for Epidemiologic Studies

Traditionally, building cohorts for epidemiologic studies relies on a time-consuming and laborious manual selection of appropriate cases. By improving the efficiency of epidemiologic research, NLP techniques can contribute to evidence-based radiology. Applications that automatically identify potential cases by NLP processing of the radiologic narrative were reported in 18 studies (34–51). Automatic selection of patients has been studied for various conditions, including renal cysts (34), pneumonia (35), pulmonary nodules (37), pulmonary embolism (40), metastases in general (41), adrenal nodules (42), abdominal aortic aneurysm (43), peripheral arterial disease (44), and liver conditions (45,46). Some studies identified patients with inserted lines or devices (47,48) or with specific BI-RADS assessment categories (38).

Many applications in this category performed a broad screening for potential cases, often followed by manual case validation and data collection. As a consequence, a high sensitivity (while maintaining a reasonable specificity) is desirable to find all potential cases (while considerably reducing the manual validation effort).

O'Connor et al identified renal cysts by NLP to assess the risk of developing renal cancer (34). An algorithm based on the open source text analysis tool GATE (General Architecture for Text Engineering) was used to screen a cohort of 15 695 patients who underwent abdominal computed tomography (CT) and automatically selected patients with findings positive for cysts with 95% sensitivity and 96% specificity. The amount and type of cysts of the selected patients was manually determined from the reports.

Dublin et al used a technique that allows one to balance the tradeoff between manual review effort and accuracy of automatic detection (35). They used the ONYX NLP tool to distinguish between three groups of patients: those with pneumonia, those without pneumonia, and those with equivocal diagnosis requiring manual review. When prioritizing the method for accuracy, 25% of reports were marked for manual review, yielding a sensitivity of 92% and specificity of 87% for the patients who were automatically classified as having positive or negative findings; tailored to reduce manual review to 12% of the reports, sensitivity decreased to 75% and specificity increased to 95%. A lower sensitivity may be acceptable if resources are limited.

Hripcsak et al used MedLEE to detect 24 different diseases and abnormalities from a large database of 889 921 chest radiographs (average sensitivity of 81% and specificity of 99%) (36). The predominant use of this database is clinical research, and the NLP results are used to screen patients for enrollment in studies. The prevalence of the selected conditions ranged from 22% for pleural effusion to 0.04% for tension pneumothorax. When the pneumothorax cases were compared with financial discharge coding, it was noted that only 17% of these cases were correctly associated with the appropriate International Classification of Diseases, version 9 (ICD-9) codes, indicating that ICD-9 codes alone are of limited value for case identification.

Danforth et al combined several ICD-9 codes with a Current Procedural Terminology, or CPT, code for a thorax CT scan within 30 days of the ICD-9 diagnosis to identify patients with pulmonary nodules (37). Although these ICD-9 codes are used to code other condition of the lung, they can also indicate the presence of one or more lung nodules. A rule-based method, using co-occurrence of positive and negative key words and mention of nodule dimensions, achieved 96% sensitivity and 86% specificity for the detection of pulmonary nodules as compared with review by a clinician.

Percha et al (38) used an automatic system to determine the BI-RADS assessment categories from mammography reports. The rule-based system used stemming, negation detection, and a dedicated lexicon that mapped BI-RADS terminologies to the respective composition classes. Its performance was extremely high, with 99% of the cases correctly classified. This result may partly be explained by the consistent use of standardized terminology in describing breast tissue composition, which has become common after the introduction of BI-RADS.

Studies in this application category automate the case identification step in cohort building. Esuli et al (39) showed that NLP can also assist in subsequent data collection (ie, chart review). A total of 500 breast imaging reports were manually annotated with specific phrases or clauses inside reports corresponding to a variety of content, including BI-RADS description, indication, follow-up advice, presence of enhancements, surgery outcomes, and lymph nodes. A series of conditional random field, or CRF, classifiers obtained an *F* score of 86% for extracting the different content types automatically.

Query-based Case Retrieval

Applications in this category retrieve cases with conditions or outcomes that are not predefined, but specified by the user in a query (52–58). These systems typically allow querying for a broad range of conditions or outcomes, which precludes the use of annotated training

examples for optimization. Instead, NLP is used to identify relevant concepts in radiology reports based on a specialized lexicon, often RadLex®, or customized lexicons. The indexing results are stored in a database, which can subsequently be queried by the user.

Gerstmair et al developed a Web-based system to retrieve radiologic images linked to reports in a picture archiving and communication system (52). The reports were indexed by using RadLex®. Initial performance of the system in capturing important terms was moderate (42% sensitivity and 68% PPV). After enrichment of the lexicon with 1500 terms that had not been captured, performance increased to 95% sensitivity and 93% PPV. These latter performance results are biased because the enrichment was based on the test data; they also suggest that there is much room for improvement of a specialized lexicon such as RadLex®.

Mamlin et al exploited LifeCode (A-Life Medical, San Diego, Calif) to index all findings within cancer-related chest x-ray reports (53). The commercial system, originally designed for billing purposes, mapped all findings to a dedicated lexicon for chest imaging, achieving 85% sensitivity and 96% PPV against all manually identified findings.

Various applications were motivated by image retrieval for educational purposes (52,55–57). For example, Do et al developed RADTF (RadLex®-compatible Teaching File), an NLP system that indexes radiology reports for findings and diagnoses that are relevant for teaching (55). The system was applied to more than 700 000 reports linked to images, and provides a large repository of on-demand teaching material.

Several studies also emphasized the potential of query-based retrieval for research purposes (52,54–56). For example, Dang et al added a research mode to Render, their NLP-driven online searchable radiology repository (56). This mode allows exporting a range of administrative data and patient characteristics of retrieved patients.

Lacson et al developed iSCOUT (54), a query application that enables the retrieval of reports containing synonymous terms from different lexicons. iSCOUT was subsequently used by Warden et al to evaluate the performance of four different lexicons and terminologies (58): RadLex®, National Cancer Institute Thesaurus, Systemized Nomenclature of Medicine (SNOMED-CT), and ICD-9. When the system was evaluated for retrieving reports with three different critical findings, none of the terminologies consistently performed best. Moreover, the authors conclude that retrieval performance was not typically correlated with the number of synonyms in the terminology.

The performance of query-based retrieval systems is highly dependent on the quality and appropriateness of the used lexicon and may not be as high as can be obtained with systems specifically trained for specific conditions. However, the advantage of the query-based systems is their ability to quickly retrieve cases for a broad range of purposes.

Quality Assessment of Radiologic Practice

This category covers applications that identify quality indicators of radiologic practice (59–73). These indicators can be used for internal quality assurance, comparison to established guidelines, or fulfilling legal requirements. Automatic content analysis of large-report databases can give insight in the daily routine and inner workings of the radiology department. NLP systems have been used to generate descriptive statistics on topics such as recommendation behavior (59–63), report completeness (64–66), communication of critical results (67,68), and case management (69,70).

In a series of studies, the Department of Radiology at Massachusetts General Hospital investigated NLP applications that determine recommendation behavior of radiologists (59–62). Dreyer et al developed LEXIMER (Lexicon Mediated Entropy Reduction), an NLP tool that analyzes grammar and terminology from individual sentences of a radiology report, resulting in a collection of phrases (59). These phrases were used as features in decision trees that classify the report as containing clinically important findings (97.5% sensitivity, 96.6% specificity) and recommendations for subsequent action (99.6% sensitivity, 98.2% specificity). Together this information allowed the correlation of diagnostic yield (the rate of clinically important findings at an imaging examination) with recommendation practice in radiology, leading to insights into appropriateness of high-cost and high-volume radiologic procedures. Dang et al used LEXIMER to generate statistics on trends in recommendations for different types of imaging examinations (60,61). A database of over 4 million radiology reports spanning 10 years was investigated. Recommendations were correlated with indications, diseases, patient age groups, sex, subspecialties, referring physicians, and inpatient versus outpatient status. Significant differences in recommendation rates were found, and the authors concluded that such large-scale analysis is relevant to make more uniform recommendations for imaging procedures.

Ip et al used a commercial NLP product to detect recommendations (87.9% sensitivity, 99.5% specificity) after observing pancreatic lesions at CT or magnetic resonance (MR) imaging (63). They found that radiologists with expertise in abdominal imaging were 2.8 times less likely to recommend further imaging than radiologists of other subspecialties. The authors remark that this finding is important, because unwarranted variation in radiologic practice compromises the quality of care. In another study, using General Architecture for Text Engineering, the authors achieved a higher sensitivity of 94.5% and comparable specificity for recommendation detection (69). They showed that 82.2% of repeat imaging procedures are performed in the absence of a follow-up recommendation by a radiologist, suggesting that additional research on the occurrence of unwanted chains of diagnostic events is warranted.

Lacson et al used iSCOUT to automatically select reports with pulmonary nodules. They correlated node management with recommendations from the

Fleischner Society Guidelines (70). The management information was still manually extracted from the report.

Researchers from Brigham and Women's Hospital implemented a validated prediction model as clinical decision support system (CDSS) into their Computerized Physician Order Entry, or CPOE, for imaging procedures (71). The CDSS, which relies on manual input from the referring clinician, advises on the need for CT pulmonary angiography. NLP was used to retrospectively evaluate the impact of the CDSS. Using General Architecture for Text Engineering, pulmonary embolisms were identified to assess the use and diagnostic yield of CT pulmonary angiography. Over a period of 4 years, CT pulmonary angiography usage decreased by 20.1%, which corresponded with a 69% increase in diagnostic yield.

A subgroup of studies in this category exploited NLP to assess quality of content and format of the radiology report itself (64–68). Good radiologic practice is subject to conventions for standardized reporting. Official guidelines assist practitioners in providing appropriate radiologic care for patients, covering minimal format requirements as well (2,3). To limit interuser and interinstitutional reporting variations, it is important to uphold report structure and completeness to these standards.

As an example of report quality assessment, Lakhani et al developed a system to automatically identify whether review of comparison images was properly documented (64). Guidelines of the American College of Radiology advise to clearly document the use of comparison imaging, which is considered to improve diagnostic accuracy. The rule-based system that was developed had an accuracy of 96%. Study results demonstrated that 26% of reports did not mention utilization of comparison imaging, while the images were available.

Gershanik et al (65) observed that referring physicians often focus on the summary information in the impression section of a report. When actionable or other relevant findings are not repeated in the impression section, the risk of disregarding critical information increases. iSCOUT was used to detect mentions of pulmonary nodules in the observation and impression sections of CT scan reports (sensitivity 80%, PPV 96%). It appeared that 36% of the documented pulmonary nodules were not repeated in the impression section.

Duszak et al used DataScout (CodeRyte, Bethesda, Md), a commercially available product, to verify the completeness of abdominal ultrasound reports according to Current Procedural Terminology criteria (66), because deficiencies in documentation can lead to lost revenue. From a billing database of reports of 37 practices, DataScout automatically detected the presence of report statements pertaining to the description of eight specific anatomic features. The authors noted that adherence to structured reporting templates will improve reporting quality and mitigate lost revenue.

For acute findings it is desirable to ensure active delivery by using direct communication channels such as the telephone. An American College of Radiology standard encourages the documentation of nonroutine communications in radiology reports (2). Lakhani et al combined their diagnostic surveillance application for identifying critical results (18) with an algorithm for detecting radiologist-to-referrer communications (98% sensitivity, 97% PPV). The developers applied their algorithm on 9.3 million reports and showed that over a period of 20 years, documentation of critical communication increased from 19% to 72% (67). This study shows the potential of NLP in discovering trends in radiologic reporting over many years.

Clinical Support Services

Applications in this category are integrated in the clinical workflow to provide assistance to radiologists at the time of reporting (74–83).

Do et al used NLP in an application that extracts both the presence of fractures and their anatomic location (74). The rule-based algorithm correctly detected reported fractures with 90% sensitivity and 95% specificity; the bone involved was identified with an accuracy of 79%. The system was tested with a voice dictation workflow and provided real-time feedback to the radiologist in the form of a reminder or advice (based on clinical practice and musculoskeletal textbooks) specific to the identified fracture, for example, to consider a second fracture and follow-up MR imaging. The authors remark that automatic retrieval of information can anticipate the need of the radiologist, allowing for a fluent incorporation into the workflow, in contrast to a search initiated by the radiologist using a text reference or Google. The linking of findings to anatomic locations is an example of relation mining. The same task was explored in two other studies that related anatomic locations to a set of predefined findings (sensitivity and PPV both 86%) (75), or to any finding (average sensitivity of 87%, PPV of 40%) (76).

Pneumonia detection has been a frequent use case for NLP-related research. SymText, an NLP system that analyzes both syntax and semantics of free text, was used by Fizman et al to extract concepts related to infiltrates, aspiration, and pneumonia from reports (77,78). These concepts were intended as input for a program called the antibiotic assistant, which helps physicians to select appropriate antibiotics for infectious diseases. The system scored 84%–100% sensitivity and 90%–99% specificity for the individual concepts, outperforming the original keyword search from the antibiotic assistant, which achieved 56%–94% sensitivity and 88%–100% specificity. SymText also achieved 95% sensitivity and 85% specificity for the automatic detection of acute bacterial pneumonia based on the extracted concepts.

In another application of NLP, an error correction module on top of a speech recognition process was implemented (79). Word co-occurrence statistics were used to predict the probability that a word occurs within a given context and to subsequently make corrections if necessary. Normally, in NLP the syntactic structure of language is used to extract meaning, now this

process is reversed to predict what was probably meant.

Supervised structuring of radiology reports as part of routine clinical practice may produce high-quality structured data that can be used for intelligent indexing, searching, and retrieval of radiology reports. Sinha et al (80) developed a system that produces a structured report in parallel to the traditionally formatted record. A graphic interface allowed for interactive editing of NLP output, resulting in a tabular list of finding and all the associated attributes. Clearly, human correction effort is dependent on the NLP performance; on a data set containing 477 findings, 85% sensitivity and 90% PPV were obtained, before manual adjustments.

Another service is the automatic mapping of radiology reports to a coding system for administrative, financial, and analytical purposes. Farkas et al (81) investigated the automatic assignment of ICD-9 codes to radiology records, as part of a challenge that involved 45 different ICD-9 codes in 1954 records (84). The authors observed that the challenge was dominated by systems that use handcrafted classification rules, but question the scalability of this approach when thousands of different ICD-9 codes need to be assigned. Their system combined automatically generated rules derived from the original ICD-9 coding guidelines with synonym enrichment by statistical analysis of labeled data. The system yielded a sensitivity of 90% and a PPV of 88% on the challenge test set.

Performance Summary

Figure 4 shows the performance of the NLP applications for which we were able to retrieve sensitivity and specificity, while Figure 5 lists the study-specific sensitivity and specificity of the applications. Overall, the performance of applications is very high, with sensitivity and specificity of many systems above 90%. There is no discernible trend in performance over time (Fig 5), nor is there a substantial difference in performance between application categories, although diagnostic surveillance falls slightly behind the other categories. This may be because the conditions detected in this category are sometimes implicit and pertain to tasks of greater complexity.

Figure 4

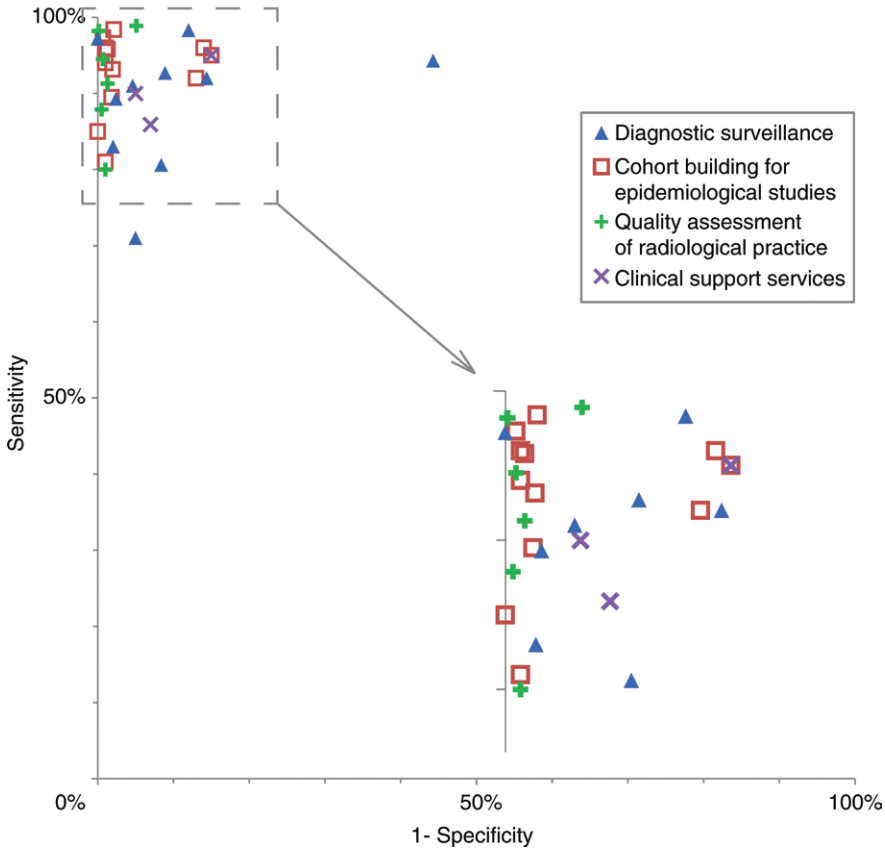

Figure 4: Performance of natural language processing systems in different application categories in radiology. Points indicate the performance (sensitivity, specificity) of individual systems. The magnified area corresponds to the upper left quadrant of the receiver operating characteristic space.

Operational Use of NLP Systems

For each study we determined whether information was provided about the operational use of the described NLP system, that is, if the system was actually used after development. We distinguished between three levels: operational use was not discussed, operational use was anticipated but not yet realized, and operational use was realized. Table 3 shows the number of systems for the different levels of operational use per application category. Systems that are intended for integration into a clinical workflow (ie, those in the diagnostic surveillance or clinical support services categories) are seldom reported to be in operational use (only one of the 27 systems that were considered). Systems that fall in the other application categories are much more prone to be operational (19 of 40), but mostly in a single institution. Only six of these systems were applied to data from more than one institution (38,45–47,66,76).

Discussion

Our review shows that NLP in radiology is used for many different purposes. The largest application categories contain systems that perform diagnostic

Figure 5

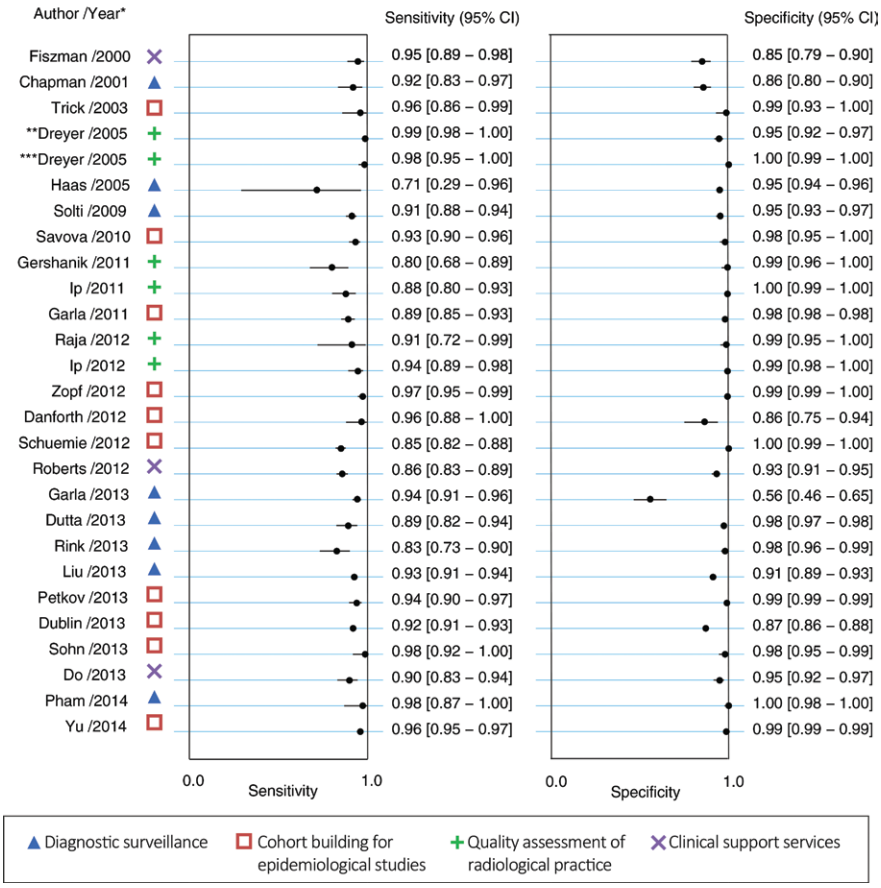

Figure 5: Study-specific sensitivity and specificity of natural language processing systems in radiology. *Studies are ordered by year of publication, **performance for detection of findings, ***performance for detection of recommendations.

Table 3

Levels of Operational Use of NLP Systems for Different Application Categories in Radiology

Application Category	Level 1*	Level 2	Level 3
Diagnostic surveillance	9	7	1
Cohort building for epidemiologic studies	9	2	7
Query-based case retrieval	4	0	3
Quality assessment of radiologic practice	5	1	9
Clinical support services	7	3	0
Total	34	13	20

* Level 1 = system development and validation, operational use not discussed; level 2 = operational use discussed and anticipated but not yet realized; level 3 = system in operational use.

surveillance, identify cases for research studies, and assess the quality of radiologic practice. Although the applications differ, most systems aim to identify the same type of information, in particular whether a radiology report contains a specific radiologic outcome (ie, a condition or individual finding). The major benefit of NLP applications is automation: They reduce–or even obviate–manual review effort, enabling the assessment of large amounts of data. As a result, tasks that previously were not contemplated become feasible. Second, NLP can bring new functionality to clinical workflows by background monitoring of reporting and advising the radiologist or referring clinician.

Applications vary in their implementation of the different possible NLP steps. All systems do tokenization and most perform stemming. For many applications it is also important to identify report sections, and nearly half of the systems include a segmentation step. Few systems perform syntactic analysis, possibly–as some studies illustrate–because little performance is gained (23,25,46). In contrast, systems often improve their performance by using features that derive from semantic analysis, which commonly employs a specialized lexicon to identify relevant terms and synonyms. Such lexicons are typically manually created by domain experts, but may also be combined with existing lexicons (33,38,43,47,52,74). Systems frequently implement negation detection, which is essential in a discipline where diagnostic imaging is often used to rule out a condition. While the simplest methods for predicting radiologic outcomes combine manually crafted decision rules with key terms and negation detection, advanced machine learning techniques can combine numerous semantic and syntactic features. Rule-based algorithms and machine learning classifiers can achieve similar performance when tested on the same dataset (19,30). A few hybrid approaches were shown to improve on the performance of a machine-learning classifier by the addition of manually constructed rules that capture exceptions (17,25,46). Machine-learning algorithms have become more popular in recent years, possibly because of their improved scalability and ease of use.

Although many NLP applications in radiology show excellent performance, overall 30% (20 of 67) of the systems

are described to be in operational use, but there are large differences between application categories. In the diagnostic surveillance and clinical support services categories, only one of 20 applications has been implemented and validated in a clinical workflow. On the other hand, in the quality assessment category, the majority of NLP applications were put to the test on large radiology databases (60–62,66,68,69). Rather than focusing on design and performance of the system, these studies aim to derive practical conclusions from application output.

There may be several reasons why many NLP applications in radiology remain in a proof-of-concept stage. First, uncertainty about minimal performance requirements may hamper system implementation, especially when a task is fully automated. Evidence-based practice insists on justification not only of therapeutic and diagnostic procedures, but also of computerized support of the workflow. However, there is no guidance on minimal performance requirements of computer systems. Such requirements should be related to human performance on the task of interest, for example, in the form of interobserver agreement scores. Unfortunately, interobserver agreement for most tasks is unknown. Studies should therefore not only evaluate system performance on a validation set, but also measure and report interobserver agreement on the same set. Performance requirements should also be task-specific: Applications that influence the clinical workflow cannot afford many errors, while less stringent criteria may be applicable to quality assessment applications. Errors may also be less costly if the system requires user validation, for example, when the output is a suggestion or recommendation that is at the physician's' or researchers' discretion to accept.

The lack of implementation of NLP systems in routine clinical practice and research may also be attributed to issues that are not performance related. Clinicians and researchers may be reluctant to accept output from automatic algorithms because it is difficult or impossible to trace how the output was generated. In that respect, rule-based classifiers and decision trees are generally more easily to comprehend than statistical classifiers.

Another issue is that applications are often tuned to data from the institution in which they were developed. In the absence of external validation, it is unclear whether the results are generalizable. The need for system tuning is partly explained by a lack of standardization of radiologic reports. Some standards that promote more uniform reporting are available, for example, Integrating the Health care Enterprise (IHE) publishes various interoperability standards for radiology (85), including indication-specific reporting templates. One institution mentioned the use of standardized reports (23,25), although the templates used did not correspond to any official guideline. Improvement of and adherence to universally recognized standards requires radiologists' acceptance, while the endorsement of professional radiology organizations would also contribute considerably. For instance, The Radiological Society of North America has embraced the IHE standards for reporting templates in the Radiology Reporting Initiative (86), uniting experts in the field to create consistent report templates for all indications based on the Management of Radiology Report Templates, or MMRT, format (87). The proliferation of NLP applications would ultimately benefit from adherence to these interoperability standards, by making applications more generalizable and perform better.

Standardization of the terminology that is used in radiologic reporting would also be beneficial. Controlled vocabularies can guide the use of uniform language at the time of reporting and help NLP applications to extract relevant features. However, existing lexicons often do not have enough coverage (ie, miss concepts or synonyms). The improvement of lexicons and their integration in reporting software in an intuitive way can enhance the performance and generalizability of NLP applications, as well as the overall quality of reporting (88).

Our review has a number of limitations. First, the heterogeneity of the applications did not allow us to perform a meta-analysis of the studies. A second limitation is that the effect of individual NLP components on application performance could often not be assessed because most studies did not provide this information. Also, there were only few studies that compared different NLP techniques, and thus it is difficult to draw general conclusions on the techniques that work best. Finally, our overview of the extent to which NLP applications are actually used in daily practice may be incomplete, as local applications may have been implemented after the study was done.

NLP has strong roots in radiology, whose reports are the most studied type of clinical narrative (5,7), and our review illustrates the variety of tasks that have been addressed. However, some more advanced tasks are scarcely studied and deserve further exploration, such as the detection of disease progression using temporal reasoning, mining relations between anatomic locations and findings, administrative coding of radiology reports, and automating chart review in the cohort-building process. We also see potential for interactive systems that aid the radiologist at the time of reporting, for example, by suggesting differential diagnoses, interactive structuring or coding of reports, or linking observations in the report to guidelines and other literature.

In this review we have only considered applications of NLP in radiology, but NLP methods have been applied in many other fields of medicine (5–7). Radiological applications may benefit from NLP applications that operate on EHR data from other fields. For instance, the NLP task of identifying a clinical diagnosis from broader EHR content can be helpful in providing a more definitive reference standard to diagnostic imaging.

In conclusion, NLP enables the automation of a diversity of tasks in radiology. The performance of NLP applications in radiology is generally high, but not many systems have been reported as being actually used in routine clinical

care or research. Establishment of minimal performance requirements, further standardization of reporting format and terminology, and external validation is likely to increase proliferation of NLP applications in this field.

Disclosures of Conflicts of Interest: E.P. Activities related to the present article: received grants from Technology Foundation STW. Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. **L.M.M.B.** Activities related to the present article: received grants from Technology Foundation STW. Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. **M.G.M.H.** Activities related to the present article: received grants from Technology Foundation STW. Activities not related to the present article: personal fees from Cambridge University Press, grants and non-financial support from European Society of Radiology (ESR), non-financial support from European Institute for Biomedical Imaging Research. Other relationships: disclosed no relevant relationships. **J.A.K.** Activities related to the present article: received grants from Technology Foundation STW. Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships.

References

1. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet 2012;13(6):395–405.
2. ACR Appropriateness Criteria. American College of Radiology. http://www.acr.org/ac. Accessed November 26, 2014.
3. European Society of Radiology (ESR). ESR communication guidelines for radiologists. Insights Imaging 2013;4(2):143–146.
4. Hearst MA. Untangling text data mining. In: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. College Park, Md: Association for Computational Linguistics, 1999.
5. 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.
6. Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? J Biomed Inform 2009;42(5):760–772.
7. Stanfill MH, Williams M, Fenton SH, Jenders RA, Hersh WR. A systematic literature review of automated clinical coding and classification systems. J Am Med Inform Assoc 2010;17(6):646–651.
8. Kao A, Poteet S. Overview. In: Kao A, Poteet S, eds. Natural language processing and text mining. New York, NY: Springer, 2007; 1–7.
9. Unified Medical Language System (UMLS). U.S. National Library of Medicine. http://www.nlm.nih.gov/research/umls/. Accessed November 26, 2014.
10. RadLex. Radiological Society of North America. http://www.rsna.org/RadLex.aspx. Accessed November 26, 2014.
11. PubMed.gov. US National Library of Medicine. http://www.ncbi.nlm.nih.gov/pubmed/. Accessed March 24, 2015.
12. EMBASE. Elsevier B.V. http://www.embase.com/. Accessed March 24, 2015.
13. Journal Citation Reports. Thomson Reuters. http://thomsonreuters.com/journal-citation-reports/. Accessed November 26, 2014.
14. Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med 2009;151(4):264–269, W64.
15. Van Rijsbergen CJ. Information retrieval. 2nd ed. London, England: Butterworths, 1979.
16. Obuchowski NA. Receiver operating characteristic curves and their use in radiology. Radiology 2003;229(1):3–8.
17. Rink B, Roberts K, Harabagiu S, et al. Extracting actionable findings of appendicitis from radiology reports using natural language processing. AMIA Jt Summits Transl Sci Proc 2013;2013:221.
18. Lakhani P, Kim W, Langlotz CP. Automated detection of critical results in radiology reports. J Digit Imaging 2012;25(1):30–36.
19. Solti I, Cooke CR, Xia F, Wurfel MM. Automated classification of radiology reports for acute lung injury: comparison of keyword and machine learning based natural language processing approaches. Proceedings (IEEE Int Conf Bioinformatics Biomed) 2009;2009:314–319.
20. Zingmond D, Lenert LA. Monitoring free-text data using medical language processing. Comput Biomed Res 1993;26(5):467–481.
21. Garla V, Taylor C, Brandt C. Semi-supervised clinical text classification with Laplacian SVMs: an application to cancer case management. J Biomed Inform 2013;46(5):869–875.
22. Cheng LT, Zheng J, Savova GK, Erickson BJ. Discerning tumor status from unstructured MRI reports: completeness of information in existing reports and utility of automated natural language processing. J Digit Imaging 2010;23(2):119–132.
23. Yetisgen-Yildiz M, Gunn ML, Xia F, Payne TH. A text processing pipeline to extract recommendations from radiology reports. J Biomed Inform 2013;46(2):354–362.
24. Dutta S, Long WJ, Brown DF, Reisner AT. Automated detection using natural language processing of radiologists recommendations for additional imaging of incidental findings. Ann Emerg Med 2013;62(2):162–169.
25. Yetisgen-Yildiz M, Gunn ML, Xia F, Payne TH. Automatic identification of critical follow-up recommendation sentences in radiology reports. AMIA Annu Symp Proc 2011;2011:1593–1602.
26. Xu Y, Tsujii J, Chang EI. Named entity recognition of follow-up and time information in 20,000 radiology reports. J Am Med Inform Assoc 2012;19(5):792–799.
27. Mendonça EA, Haas J, Shagina L, Larson E, Friedman C. Extracting information on pneumonia in infants using natural language processing of radiology reports. J Biomed Inform 2005;38(4):314–321.
28. Haas JP, Mendonça EA, Ross B, Friedman C, Larson E. Use of computerized surveillance to detect nosocomial pneumonia in neonatal intensive care unit patients. Am J Infect Control 2005;33(8):439–443.
29. Gundlapalli AV, Olson J, Smith SP, et al. Hospital electronic medical record-based public health surveillance system deployed during the 2002 Winter Olympic Games. Am J Infect Control 2007;35(3):163–171.
30. Chapman WW, Fizman M, Chapman BE, Haug PJ. A comparison of classification algorithms to automatically identify chest x-ray reports that support pneumonia. J Biomed Inform 2001;34(1):4–14.
31. Liu V, Clark MP, Mendoza M, et al. Automated identification of pneumonia in chest radiograph reports in critically ill patients. BMC Med Inform Decis Mak 2013;13:90.
32. Elkin PL, Froehling D, Wahner-Roedler D, et al. NLP-based identification of pneumonia cases from free-text radiological reports. AMIA Annu Symp Proc 2008:172–176.
33. Pham AD, Névéol A, Lavergne T, et al. Natural language processing of radiology reports for the detection of thromboembolic diseases and clinically relevant incidental findings. BMC Bioinformatics 2014;15:266.
34. O'Connor SD, Silverman SG, Ip IK, Maehara CK, Khorasani R. Simple cyst-appearing renal masses at unenhanced CT: can

they be presumed to be benign? Radiology 2013;269(3):793–800.

35. Dublin S, Baldwin E, Walker RL, et al. Natural Language Processing to identify pneumonia from radiology reports. Pharmacoepidemiol Drug Saf 2013;22(8):834–841.

36. Hripcsak G, Austin JH, Alderson PO, Friedman C. Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports. Radiology 2002;224(1):157–163.

37. Danforth KN, Early MI, Ngan S, Kosco AE, Zheng C, Gould MK. Automated identification of patients with pulmonary nodules in an integrated health system using administrative health plan data, radiology reports, and natural language processing. J Thorac Oncol 2012;7(8):1257–1262.

38. Percha B, Nassif H, Lipson J, Burnside E, Rubin D. Automatic classification of mammography reports by BI-RADS breast tissue composition class. J Am Med Inform Assoc 2012;19(5):913–916.

39. Esuli A, Marcheggiani D, Sebastiani F. An enhanced CRFs-based system for information extraction from radiology reports. J Biomed Inform 2013;46(3):425–435.

40. Yu S, Kumamaru KK, George E, et al. Classification of CT pulmonary angiography reports by presence, chronicity, and location of pulmonary embolism with natural language processing. J Biomed Inform 2014; 52:386–393.

41. Petkov VI, Penberthy LT, Dahman BA, Poklepovic A, Gillam CW, McDermott JH. Automated determination of metastases in unstructured radiology reports for eligibility screening in oncology clinical trials. Exp Biol Med (Maywood) 2013;238(12):1370–1378.

42. Zopf JJ, Langer JM, Boonn WW, Kim W, Zafar HM. Development of automated detection of radiology reports citing adrenal findings. J Digit Imaging 2012;25(1): 43–49.

43. Sohn S, Ye Z, Liu H, Chute CG, Kullo IJ. Identifying abdominal aortic aneurysm cases and controls using natural language processing of radiology reports. AMIA Jt Summits Transl Sci Proc 2013;2013:249–253.

44. Savova GK, Fan J, Ye Z, et al. Discovering peripheral arterial disease cases from radiology notes using natural language processing. AMIA Annu Symp Proc 2010;2010:722–726.

45. Schuemie MJ, Sen E, 't Jong GW, van Soest EM, Sturkenboom MC, Kors JA. Automating classification of free-text electronic health records for epidemiological studies. Pharmacoepidemiol Drug Saf 2012;21(6):651–658.

46. Garla V, Lo Re V 3rd, Dorey-Stein Z, et al. The Yale cTAKES extensions for document classification: architecture and application. J Am Med Inform Assoc 2011;18(5):614–620.

47. Rubin D, Wang D, Chambers DA, Chambers JG, South BR, Goldstein MK. Natural language processing for lines and devices in portable chest x-rays. AMIA Annu Symp Proc 2010;2010:692–696.

48. Trick WE, Chapman WW, Wisniewski MF, Peterson BJ, Solomon SL, Weinstein RA. Electronic interpretation of chest radiograph reports to detect central venous catheters. Infect Control Hosp Epidemiol 2003;24(12):950–954.

49. Flynn RW, Macdonald TM, Schembri N, Murray GD, Doney AS. Automated data capture from free-text radiology reports to enhance accuracy of hospital inpatient stroke codes. Pharmacoepidemiol Drug Saf 2010;19(8):843–847.

50. Friedlin J, McDonald CJ. A natural language processing system to extract and code concepts relating to congestive heart failure from chest radiology reports. AMIA Annu Symp Proc 2006:269–273.

51. Zhou Y, Amundson PK, Yu F, Kessler MM, Benzinger TL, Wippold FJ. Automated classification of radiology reports to facilitate retrospective study in radiology. J Digit Imaging 2014;27(6):730–736.

52. Gerstmair A, Daumke P, Simon K, Langer M, Kotter E. Intelligent image retrieval based on radiology reports. Eur Radiol 2012;22(12):2750–2758.

53. Mamlin BW, Heinze DT, McDonald CJ. Automated extraction and normalization of findings from cancer-related free-text radiology reports. AMIA Annu Symp Proc 2003:420–424.

54. Lacson R, Andriole KP, Prevedello LM, Khorasani R. Information from Searching Content with an Ontology-Utilizing Toolkit (iSCOUT). J Digit Imaging 2012;25(4):512–519.

55. Do BH, Wu A, Biswal S, Kamaya A, Rubin DL. Informatics in radiology: RADTF—a semantic search-enabled, natural language processor-generated radiology teaching file. RadioGraphics 2010;30(7):2039–2048.

56. Dang PA, Kalra MK, Schultz TJ, Graham SA, Dreyer KJ. Informatics in radiology: Render—an online searchable radiology study repository. RadioGraphics 2009;29(5):1233–1246.

57. Hersh W, Mailhot M, Arnott-Smith C, Lowe H. Selective automated indexing of findings and diagnoses in radiology reports. J Biomed Inform 2001;34(4):262–273.

58. Warden GI, Lacson R, Khorasani R. Leveraging terminologies for retrieval of radiology reports with critical imaging findings. AMIA Annu Symp Proc 2011;2011:1481–1488.

59. Dreyer KJ, Kalra MK, Maher MM, et al. Application of recently developed computer algorithm for automatic classification of unstructured radiology reports: validation study. Radiology 2005;234(2):323–329.

60. Dang PA, Kalra MK, Blake MA, et al. Natural language processing using online analytic processing for assessing recommendations in radiology reports. J Am Coll Radiol 2008;5(3):197–204.

61. Dang PA, Kalra MK, Blake MA, Schultz TJ, Halpern EF, Dreyer KJ. Extraction of recommendation features in radiology with natural language processing: exploratory study. AJR Am J Roentgenol 2008;191(2): 313–320.

62. Sistrom CL, Dreyer KJ, Dang PP, et al. Recommendations for additional imaging in radiology reports: multifactorial analysis of 5.9 million examinations. Radiology 2009; 253(2):453–461.

63. Ip IK, Mortele KJ, Prevedello LM, Khorasani R. Focal cystic pancreatic lesions: assessing variation in radiologists' management recommendations. Radiology 2011;259(1):136–141.

64. Lakhani P, Menschik ED, Goldszal AF, Murray JP, Weiner MG, Langlotz CP. Development and validation of queries using structured query language (SQL) to determine the utilization of comparison imaging in radiology reports stored on PACS. J Digit Imaging 2006;19(1):52–68.

65. Gershanik EF, Lacson R, Khorasani R. Critical finding capture in the impression section of radiology reports. AMIA Annu Symp Proc 2011;2011:465–469.

66. Duszak R Jr, Nossal M, Schofield L, Picus D. Physician documentation deficiencies in abdominal ultrasound reports: frequency, characteristics, and financial impact. J Am Coll Radiol 2012;9(6):403–408.

67. Lakhani P, Langlotz CP. Automated detection of radiology reports that document non-routine communication of critical or significant results. J Digit Imaging 2010; 23(6):647–657.

68. Lakhani P, Kim W, Langlotz CP. Automated extraction of critical test values and communications from unstructured radiology reports: an analysis of 9.3 million reports from 1990 to 2011. Radiology 2012;265(3):809–818.

69. Ip IK, Mortele KJ, Prevedello LM, Khorasani R. Repeat abdominal imaging exam-

inations in a tertiary care hospital. Am J Med 2012;125(2):155–161.

70. Lacson R, Prevedello LM, Andriole KP, et al. Factors associated with radiologists' adherence to Fleischner Society guidelines for management of pulmonary nodules. J Am Coll Radiol 2012;9(7):468–473.

71. Raja AS, Ip IK, Prevedello LM, et al. Effect of computerized clinical decision support on the use and yield of CT pulmonary angiography in the emergency department. Radiology 2012;262(2):468–474.

72. Sippo DA, Warden GI, Andriole KP, et al. Automated extraction of BI-RADS final assessment categories from radiology reports with natural language processing. J Digit Imaging 2013;26(5):989–994.

73. Fiszman M, Haug PJ, Frederick PR. Automatic extraction of PIOPED interpretations from ventilation/perfusion lung scan reports. Proc AMIA Symp 1998:860–864.

74. Do BH, Wu AS, Maley J, Biswal S. Automatic retrieval of bone fracture knowledge using natural language processing. J Digit Imaging 2013;26(4):709–713.

75. Sevenster M, Bozeman J, Cowhy A, Trost W. Automatically pairing measured findings across narrative abdomen CT reports. AMIA Annu Symp Proc 2013;2013:1262–1271.

76. Sevenster M, van Ommering R, Qian Y. Automatically correlating clinical findings and body locations in radiology reports using MedLEE. J Digit Imaging 2012;25(2):240–249.

77. Fiszman M, Chapman WW, Aronsky D, Evans RS, Haug PJ. Automatic detection of acute bacterial pneumonia from chest x-ray reports. J Am Med Inform Assoc 2000;7(6):593–604.

78. Fiszman M, Chapman WW, Evans SR, Haug PJ. Automatic identification of pneumonia related concepts on chest x-ray reports. Proc AMIA Symp 1999:67–71.

79. Voll K, Atkins S, Forster B. Improving the utility of speech recognition through error detection. J Digit Imaging 2008;21(4):371–377.

80. Sinha U, Dai B, Johnson DB, et al. Interactive software for generation and visualization of structured findings in radiology reports. AJR Am J Roentgenol 2000;175(3):609–612.

81. Farkas R, Szarvas G. Automatic construction of rule-based ICD-9-CM coding systems. BMC Bioinformatics 2008;9(Suppl 3):S10.

82. Roberts K, Rink B, Harabagiu SM, et al. A machine learning approach for identifying anatomical locations of actionable findings in radiology reports. AMIA Annu Symp Proc 2012;2012:779–788.

83. Bozkurt S, Rubin D. Automated detection of ambiguity in BI-RADS assessment categories in mammography reports. Stud Health Technol Inform 2014;197:35–39.

84. Pestian JP, Brew C, Matykiewicz P, et al. A shared task involving multi-label classification of clinical free text. Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing. Prague, Czech Republic: Association for Computational Linguistics, 2007; 97–104.

85. Integrating the Healthcare Enterprise. IHE Radiology. http://www.ihe.net/Radiology/. Accessed November 26, 2014.

86. Radiology Reporting Initiative. Radiological Society of North America. http://www.rsna.org/Reporting_Initiative.aspx. Accessed November 26, 2014.

87. Radreport.org. Radiological Society of North America. http://www.radreport.org/. Accessed March 11, 2015.

88. Cramer JA, Eisenmenger LB, Pierson NS, Dhatt HS, Heilbrun ME. Structured and templated reporting: An overview. Appl Radiol 2014.