Aspiring to Unintended Consequences of Natural Language Processing: A Review of Recent Developments in Clinical and Consumer-Generated Text Processing

D. Demner-Fushman[1]*, N. Elhadad[2]*
[1] National Library of Medicine, National Institutes of Health, Bethesda, MD, USA
[2] Columbia University, New York, NY, USA

Summary

Objectives: This paper reviews work over the past two years in Natural Language Processing (NLP) applied to clinical and consumer-generated texts.

Methods: We included any application or methodological publication that leverages text to facilitate healthcare and address the health-related needs of consumers and populations.

Results: Many important developments in clinical text processing, both foundational and task-oriented, were addressed in community-wide evaluations and discussed in corresponding special issues that are referenced in this review. These focused issues and in-depth reviews of several other active research areas, such as pharmacovigilance and summarization, allowed us to discuss in greater depth disease modeling and predictive analytics using clinical texts, and text analysis in social media for healthcare quality assessment, trends towards online interventions based on rapid analysis of health-related posts, and consumer health question answering, among other issues.

Conclusions: Our analysis shows that although clinical NLP continues to advance towards practical applications and more NLP methods are used in large-scale live health information applications, more needs to be done to make NLP use in clinical applications a routine widespread reality. Progress in clinical NLP is mirrored by developments in social media text analysis: the research is moving from capturing trends to addressing individual health-related posts, thus showing potential to become a tool for precision medicine and a valuable addition to the standard healthcare quality evaluation tools.

Keywords

Natural Language Processing, medical informatics applications, social media, computing methodologies, review

Yearb Med Inform 2016:224-33
http://dx.doi.org/10.15265/IY-2016-017
Published online November 10, 2016

1 Introduction

The promise of a wider use of clinical Natural Language Processing (NLP) in healthcare information technology (IT) has been at our fingertips for decades, with several successful applications integrated in daily care, such as MedLEE [1]. Yet clinical NLP, i.e., natural language processing methods developed and applied to support healthcare by operationalizing clinical information contained in clinical narrative, remains an emerging technology. The gap between the promise and reality is becoming noticeable to the potential beneficiaries of clinical NLP: a recent editorial in Circulation: Cardiovascular Quality and Outcomes notes that most of clinical NLP current successes are restricted to research settings [2]. The authors then state: "NLP tools do not perform well enough for focused clinical tasks like real-time surveillance, quality profiling, and quality improvement initiatives, and the focused NLP tools tend to lose performance in clinical environments outside of their development frame. […] As a result, NLP use in clinical operations has been limited." As a community, NLP researchers need to take urgent steps to amend the situation; the work discussed in this review gives some hope towards this goal.

This review includes any application or methodological publication that leverages texts to facilitate healthcare and addresses the needs of consumers and populations. We focus on the areas where we see much active contribution of NLP techniques in the past two years, with few exceptions for somewhat older papers. Activity in these areas is facilitated by a dramatically growing availability of the texts to researchers and the changing culture that promotes sharing tools and resources.

We omit discussing basic research recently reviewed by Névéol and Zweigenbaum [3] that is however still needed and ongoing. Some examples include exciting new approaches proposed in the context of community challenges: 2012 i2b2 event and time extraction [4], 2014 i2b2/UTHealth modeling of risk factors for heart disease [5], ShARe SemEval 2014 recognition and normalization of disorders [6, 7], and ShARe SemEval 2015 disorder and template filling shared tasks [8]. We refer the reader to the overview papers for each task to learn more about the different proposed approaches. Similarly, we do not discuss NLP methods developed within the Text Retrieval Conference (TREC) Medical Records and Clinical Decision Support tracks that respectively focused on finding patient cohorts using EHR notes and eligibility criteria for the Medical Records track, and finding relevant publications and published case studies given a description of a patient's case for the Clinical Decision Support track. The overviews and the detailed descriptions of individual efforts are available in TREC publications [9, 10]. We also leave the discussion of the international developments to the MEDINFO 2015 panel "Current perspectives on NLP for electronic medical records" [11], recent editions of CLEF eHealth challenges, and the AMIA 2014

* Both authors contributed equally

panel "Clinical Natural Language Processing in Languages Other Than English" [12].

Since Névéol and Zweigenbaum provide information about methods [3], we only mention here that the methods in the included papers range from regular expressions that dominate research in social-media text processing, to event extraction in a supervised setting. More recently, there has been more and more progress in incorporating the principles of distributional semantics into an NLP pipeline, and a shift towards more semantic parsing, however more work in these areas is needed.

In the rest of the paper, we discuss in greater depth disease modeling using clinical texts (section 2), patient cohort selection (section 3), secondary uses of clinical data, i.e., uses outside of direct health care delivery [13] (section 4), support for hospital operations (section 5), and support for individuals and populations that often relies on text analysis in social media (section 6). We conclude the review with a critical reflection on the success and constraints of using NLP methods in clinical settings and propose an outlook for the future.

2 Disease Representation

Representing disease in a computable form has been one of the long-standing research activities in the field of biomedical informatics. Here, we consider computational modeling and phenotyping from the standpoint of representing a disease, either through hard-coded or probabilistic rules over clinical observations in the patient record. In recent years, phenotyping efforts have considered the electronic health record (EHR) as one of the primordial sources [14,15].

Supervised Approaches to Disease Representation

Disease representation approaches that leverage information conveyed in clinical notes and reports have focused mostly on modeling one disease at a time. In this setup, the goal is to classify a given patient record as positive or negative for the presence of a specific disease. The task is often cast as a classification problem, and the set of discriminative features represents a model of the disease, which can be examined and interpreted by experts. Not surprisingly, the challenge lies in selecting the features that yield the most accurate representation of the disease.

Most methods for identifying features from patient notes use the approach we outline next (see [16] for a review and more recently [17–20]). Experts provide a list of phenotype-related terms that will constitute the basis of the disease model. To account for lexical variations in the expert-provided terms, the terms are mapped to standard terminologies (e.g., terms related to signs and symptoms are mapped to SNOMED-CT, and medication names are mapped to RxNorm). The augmented, custom dictionary is then used as part of a standard NLP pipeline that extracts term mentions and modifiers such as negation and uncertainty. In this scenario, the text surrounding the mentions of the controlled vocabulary terms is ignored by the disease-modeling task. The disease model is thus heavily dependent on the expert-provided initial phenotype description.

To ease the reliance on the curated list from domain experts, Yu and colleagues propose to identify features in the following fashion [21]: to identify the terms of interest for a target disease, they look for candidate terms in publicly available documents that are known to discuss the disease (e.g., Wikipedia, Merck manual). These candidate terms are then screened further by examining their distribution over notes in the EHR. When tested on the modeling of two diseases (rheumatoid arthritis and coronary artery disease), the phenotypes with automatically generated features were more accurate at identifying patients with the disease than the ones using expert-curated features.

Luo and colleagues [22] propose a different approach to curated features. In their phenotyping scenario, input data come from clinical reports. Sentences are translated into a graph representation, using Unified Medical Language System (UMLS) for concept/node mapping. The sentence subgraphs are then mined to identify frequent and duplicate graphs. Luo et al. tested their approach on modeling three types of lymphomas from pathology reports. They showed that using sentence subgraphs as features yields better disease classification than using words alone (n-grams) or UMLS terms. More importantly, for the sake of disease modeling, they confirmed that the semantic modeling of the report content yields better phenotyping performance.

Semi-supervised Approaches to Phenotyping

Although semi- and distantly supervised methods still require an annotated gold standard, they are less reliant on manual annotations. In order to reduce this complex and time-consuming step, Halpern and colleagues introduced the concept of anchor variables, which they defined as an intermediary key observation about the patient, that might in turn be relevant for phenotyping (e.g., an anchor might be defined as "has cardiac etiology") [23]. They learned the anchors in an automated fashion through unlabeled data and ontological knowledge. They also provided an interface to elicit from domain experts the relevance of the anchors for a specific phenotyping task.

Beyond Single-disease Modeling

Recently, methods to identify sub-groups of a given disease have been investigated. We note, however, that most approaches to date do not rely yet on the notes and reports (e.g., [24]). Approaches to model a large number of diseases at once have also been proposed in the literature. There, the point is not precise disease modeling, but rather, to model a map of patient characteristics across diseases. With such multiple disease models, patients can be characterized according to all their conditions, rather than according to a single condition, and diseases can be studied in the context of each other. Like in the subgroup identification work, most approaches currently examine structured data and ignore patient notes [25, 26]. One exception is the UPhenome model [27], which leverages notes. It uses a probabilistic graphical model to jointly learn a representation over the words in the notes, the structured part of

the patient record, and a very large set of latent variables representing phenotypes to be learned. When evaluated on 50 random phenotypes from the 750 learned, clinical experts found that the learned disease models were coherent and representative of the corresponding diseases.

3 Patient Cohort Selection

Both the prospective clinical studies that need to find and enroll eligible patients and the retrospective studies that are increasingly relying on secondary use of EHR data turn to clinical notes to extract some of the patients' characteristics. Systems that identify cohorts eligible for a clinical trial are described in the 2010 review of Weng and colleagues [28]. Several studies have since included NLP in addressing the important problem of identifying eligible patients across an institution's EHR. This section does not discuss a large body of work in two related areas: the analysis of eligibility criteria in the description of clinical studies, e.g., in ClinicalTrials.gov (see [29] for an example), and the approaches to standardize formal representations of the inclusion and exclusion criteria [28].

Extracting Cohort Characteristics from Clinical Texts

Extraction of cohort characteristics from clinical texts ranges from the relatively simple task of identifying a single eligibility criterion [30] to the identification of several criteria within or across trials [31, 32]. Methods to augment the extracted information and represent the patient in a more sophisticated fashion have also been proposed [33].

When comparing eight automated approaches to extracting patients' metastatic status from unstructured radiology reports, Petkov and colleagues [30] found that the best performing algorithm consisted of a sequence of rules encoding positive and negative relations among metastatic terms and a set of "ignore phrases" (e.g., "evaluate for metastasis"). This approach resulted in sensitivity 94%, specificity 100%, positive predictive value 90%, negative predictive value 100%, and accuracy of 99% on a set of 5,523 patients with 10,492 radiology reports.

Kreuzthaler and colleagues assessed the necessity and accuracy of information extraction (IE) from clinical notes for a cohort study on metabolic syndrome [31]. For this study, about 50% of the needed attributes were in semi-structured document templates. Using the Apache UIMA framework and regular expressions for information extraction, the authors achieved a 0.90 F-score, identifying the main challenges for information extraction to be typing errors, inconsistency, and redundancy and spelling variants in the notes. In the outlook for NLP in cohort identification, the authors were fairly optimistic with respect to the efforts needed for adapting IE frameworks to specific information needs, provided the spelling variants and errors could be normalized to a standard vocabulary.

Ni and colleagues assessed the effectiveness of information extraction methods in automated eligibility screening for clinical trials in a urban tertiary care pediatric emergency department (ED) [32]. To that end, the authors collected eligibility criteria for 13 clinical trials as well as demographics, laboratory data, and clinical notes for 202,795 patients visiting their ED during the trial recruitment period. The eligibility criteria were based on 15 EHR fields, seven of which were derived using text in the clinical notes. The structured fields were used to develop logical constraint filters. These filters were combined with descriptive criteria derived from the notes using information extraction, concept identification, negation detection, and elements of discourse, as well as supervised term expansion based on UMLS hyponymy of the terms identified in the notes of eligible patients. This allows to reduce the workload from 98 encounters that clinicians would have on average per trial to identify all eligible patients in the gold standard set to eight screened encounters per trial.

Because much of the eligibility criteria have a temporal aspect, Raghavan and colleagues argued that the matching of a patient's report to the criteria needs to integrate the timing of events documented in the record [33]. Thus, they proposed an approach to creating a timeline of a patient in a supervised fashion. The primary contribution of the work is that it orders the different events of the patient's record spread across the different parts of the report. They represent the events and their ordering through a finite state transducer, which enables search for the best ordering. Of note, while the goal of this work is patient matching with eligibility criteria, the creation of such a timeline is a promising venue for NLP techniques.

Beyond Matching Rules

Methods for matching patients with eligibility criteria that go beyond matching rules have been proposed as well in the recent years. Li et al. [34] explored methods for linking medications and their attributes in two corpora, 3,000 clinical trial announcements and 1,655 clinical notes, which represent the types of texts that will need to be linked through the common criteria representation. Li et al. compared binary classification of links to CRF-based multi-layered sequence labeling, in which each layer deals with one type of label. Both methods had comparable performance and achieved F-measures in the 80s on two different collections.

Shivade and colleagues annotated sentences related to one of the four potential eligibility criteria related to cardiac problems in 80 of the records provided in the 2014 i2b2 challenge on identifying risk of heart disease [35]. They piloted an approach inspired by the Recognizing Textual Entailment (RTE) task [36] to decide whether the criterion can be inferred from patient's record sentences. Of the four relatively simple RTE methods, semantic methods outperformed lexical methods. However, the results were low for all tested methods.

Miotto and Weng derived a target patient representation for 13 diversified clinical trials, one for each trial [37]. The target representation consisted of four vectors, one of which was based on clinical notes, represented through their distribution over a learned topic model. For a given trial, EHR data of a new unseen patient was matched to the "target patient" using pairwise cosine similarity. Ranked patients with a similarity score above an empirically set threshold were considered eligible. When evaluating 262 participants of the 13 trials, half of

which were used for training and the other half were combined with 30,000 randomly selected patients for testing, binary classification of patients as eligible or not achieved 0.95 AUC. This approach indicates that the efforts in structuring patients' criteria and matching them to EHR data (or the literature for decision support, as is the case in the CDS track of TREC [10]) are a promising research direction.

4 Other Secondary Uses of Clinical Data

Besides disease modeling and cohort selection, there are several applications in the realm of secondary use of clinical data that leverage clinical notes. They roughly fall into three types of applications: predictive analytics, pharmacovigilance and drug repurposing, and characterization of population and care patterns.

Predictive Analytics

The task at hand in predictive analytics is to predict an outcome or an event of interest in the future (e.g., a patient is readmitted to the hospital). In the recent approaches to predictive analytics tasks that consider clinical notes content, text-related features consist of bag-of-words. For instance, Poulain and colleagues described a retrospective study for risk of suicide in U.S veterans, and used n-grams extracted from a small cohort of clinical records to identify potential predictors of suicide [38].

Goodwin and Harabagiu developed a probabilistic graph-based method to predict progression of clinical findings for individual patients [39]. Re-using clinical notes annotated for the 2014 i2b2 challenge, they infered chronological ordering of the findings (obesity, hypertension, diabetes, hyperlipidemia, and coronary artery disease) and used probabilistic inference on the graphical model to make predictions. Although the computational approach is interesting, it needs to be further explored using currently available NLP methods, rather than gold standard annotations to build predictive models.

When predicting 30-day readmission, Caruana and colleagues took all mentions of UMLS terms in the notes with a mapping to the Core Problem List Subset of SNOMED-CT, and created predictive models that can scale to the number of patients and features [40]. For instance, their models, trained on 195,000 patients and tested on 100,000 patients, can incorporate about 4,000 features per patient, most of them being terms extracted from the notes.While research in outcome prediction in the intensive care unit (ICU) has prevalently focused on physiological signals, there is emerging work on incorporating clinical documentation for outcome prediction. Ghassemi and colleagues explored modeling of mortality at different time ranges (in the hospital, 30-day post discharge, and 1-year post discharge) [41]. Text-derived features consisted of the distribution over a topic model learned across a large corpus of ICU notes. When text-features were added to baseline clinical features, such as severity scores and demographics, mortality prediction yielded better performance for all time ranges, and the discriminative topics correlated with known causes of death.

To study progression of disease, Perotte and colleagues cast their work in a survival analysis framework [42]. They proposed a method to incorporate longitudinal data and documentation prior to onset of disease into the survival model for progression. They used topic models, as learned from a large corpus of notes on patients with chronic kidney disease. In their experiments on chronic kidney disease progression, they showed that a model, which incorporates longitudinal topic models and laboratory test data, performs the best at predicting which patients are more likely to progress faster. Furthermore, like in the mortality study, they found that the significant topic models for progression correlated with known risks of progression.

Pharmacovigilance and Drug Repurposing

Leveraging content of clinical notes to identify potential adverse drug events (ADEs) in a systematic fashion is another active area of research. Wang and colleagues extracted drug and disorder mentions from the clinical notes of 1.6 million patients to create a pool of drug-disorder pairs [43]. The pairs were then used as instances for learning potential ADEs. For the task of drug-drug interactions, Iyer and colleagues started from a similar approach of extracting all drugs and disorders, but they included a temporal aspect in their statistical analysis [44]. Rather than relying on global occurrence counts derived from notes mentions, Henriksson and colleagues focused on identifying explicit relations between drug and disorder mentions in the clinical notes, including within sentences and across sentences [45]. They experimented with distributional semantics, specifically word2vec, and showed a positive impact on the learning of explicit ADE relations in notes. Of note, social media is emerging as an additional, complementary source for ADE detection to clinical data (for detailed reviews, see [46–48]).

Characterizing Populations and Patterns of Care

Even for clinical events for which there exist well-defined concepts in standard ontologies, there is value in using simple keyword searches for identifying patient cohorts relevant to this event. For instance, when identifying dialysis, Abhyankar and colleagues showed that combining search over structured codes with simple keyword search of notes identified populations with a better overall performance [49]. Researchers have proposed similar methods with good success for a range of tasks, including identifying documentation patterns of Framingham criteria in patients with and without heart failure [50], determining the prevalence of different indications for colonoscopies [51], measuring physician adherence with guidelines for medication use and behavioral modification in gout patients [52], identifying patterns of opioids over-prescription [53], and tracking the population of congestive heart failure patients in a state-wide health information exchange [54].

The use of distributional semantics was also found to be helpful in characterizing

patterns of clinical documentations. Sullivan and colleagues used topic-model representations of clinical notes to help in detecting potential misdiagnosis in the case of epilepsy syndrome in a pediatric population [55]. McCoy and colleagues, in order to study the relevance of research domain criteria in psychiatric care, mapped clinical notes to prevalence of documentation according to five domain criteria (negative valence, positive valence, cognitive functioning, arousal, social processes) [56]. There, a direct keyword search approach makes less sense. For each domain, they identified a set of domain-relevant corpora of Web pages. They then transformed domain-specific web pages and clinical notes in a vector space model, using latent semantic indexing. Clinical notes were then scored according to their similarity to each domain-specific vector. McCoy and colleagues showed that their approach not only helps them score documentation with respect to domain criteria as described above, but also characterizes populations and outcomes, such as length of stay, according to these inferred scores.

5 Supporting Hospital Operations

Clinical NLP can have a practical impact on administrative as well as point-of-care aspects of hospital operations. Some practical impact can already be seen in such established areas as medical coding and billing. The work in this area continues to grow and is paralleled by research and some advances into practice in quality improvement and clinical decision support.

Supporting Hospitals with Billing and Reporting Activities

Efforts in supporting the needs of hospitals with billing and syndromic surveillance have been reported since the last two years. Perotte and colleagues proposed to leverage the content of discharge summaries to identify billing codes without restriction to a clinical domain on a corpus of 26,000 discharge summaries [57]. Their feature set is a simple bag-of-words from clinical notes, but the classification itself leverages the hierarchical nature of the ICD-9 tree. With the adoption of ICD-10 coding, Subotin and Davis proposed a diagnosis code assignment method, which also considers a bag-of-words approach, but combines a series of assignments based in part on the structure of the ICD-10 classification [58]. Their experiments on a corpus of 28,000 patient records show promising results for this new and complex terminology.

Towards the goal of syndromic surveillance, Haas and colleagues proposed to classify ED (emergency department) triage notes into one of three high-level categories: gastro-intestinal, respiratory, and fever-rash [59]. Starting from a master list of terms pertinent to each category, they iteratively added terms to the list by searching the triage notes for terms similar according to lexical, context-based metric. More recently, Lopez Pineda and colleagues focused on predicting influenza in the ED. They experimented with data from four different hospitals to predict influenza from the clinician-authored reports, rather than triage notes and chief complaints [60].

Quality Improvement

In addition to billing and reporting activities, exploratory work in assisting healthcare organizations in improving the quality of data is ongoing. Yetisgen and colleagues developed statistical and knowledge based methods that combined publicly available tools in pipelines to support the Surgical Care and Outcomes Assessment Program (SCOAP) [61]. The program aims to improve quality and compare effectiveness of surgical procedures across multiple Washington state hospitals. The F-scores for the 25 extracted elements in this study performed on the notes for 618 patients from one institution varied for both the statistical and rule-based methods, but are encouraging enough to warrant further research.

Raju and colleagues used keyword extraction to identify and compute adenoma detection rate using colonoscopy reports [62]. This method outperformed manual screening by correctly identifying 91.3% of screening examinations as compared to 87.8% identified manually. Similarly, Gawron and colleagues [63] primarily used regular expressions for adenoma detection, achieving 0.98 and 0.99 accuracy in identifying screening indication and complete procedure, respectively. The correct location and histology of the polyp was identified with 0.94 positive predictive value, 0.94 sensitivity, and 0.94 F1 score. The numbers of polyps and adenomas were underestimated by the method.

Although not yet directly applicable to clinical narrative, an approach to formalizing quality measures developed by Dentler and colleagues allows to structure the requirements and issue database SQL queries to compute quality measures [64].

Clinical Decision Support

Despite the promise of opportunities for NLP techniques to contribute to clinical decision support (CDS) tools [65], there are few instances of applications that operate at the point of care and that make use of NLP technology. Dean and colleagues reported on a real-time pneumonia screening tool in the ED that provides care recommendations when pneumonia is inferred from the patient's radiology report [66,67]. While they did not observe a significant difference in mortality across the EDs where the tool was deployed and the control EDs, they found that EDs with access and use of the tool had increased adherence with recommendations for pneumonia care. Demner-Fushman and colleagues reported on an in-depth evaluation of their evidence-based decision support tool [68]. The evaluation showed stable use of an application that extracts concepts from the patients' progress notes to automatically generate searches over several resources identified as useful by the NIH Clinical Center interdisciplinary teams. More recently, there has been renewed work in problem-list generation based on patient notes [69, 70] and through patient record summarization [71] (for a review of patient record summarization techniques that use NLP, see [72]).

6 Supporting the Needs of Individuals and Populations

Along with the general explosion of social media use, more and more health consumers discuss their health and their care ecosystem online, whether on general social media sites or in dedicated discussion boards. For instance, in a study of 294,000 Yelp New York City restaurant reviews, Harrison and colleagues found 468 reports consistent with foodborne illness, of which only 3% had been reported through the official channels [73]. In a 2011 literature review, Smith concluded that consumer language is an under-researched area inside and outside of healthcare [74]. We review here recent advances in NLP for health consumer language, as well as exciting avenues for learning from patient-generated texts.

NLP for Health Consumer Language

In the past couple years, there has been some evidence suggesting that the language used in health texts should be adapted to the level of health literacy of health consumers, for them to comprehend the text. A study of FDA Drug Safety Communications revealed that changing the existing communications to plain language significantly increased consumers' level of comprehension of the communications [75]. Ramesh and colleagues, linking 20 de-identified progress note reports and 20 de-identified discharge summary reports to MedlinePlus, UMLS, and Wikipedia, showed Wikipedia to significantly improve self-reported EHR note comprehension by AMT workers [76].

There has also been much research recently in building basic NLP tools to support the automated analysis of the language authored by health consumers and patients. Basic approaches developed recently for consumer language include: spelling correction, for which Zhou and colleagues rely on Google Spell Checker [77], whereas Kilicoglu and colleagues are developing a stand-alone publicly available tool and corpora [78]; automated evaluation of errors in consumer language processing made by NLP tools [79]; extraction of patient demographics and personal medical information [80, 81]; Keyword in Context (KWIC) analysis to evaluate patients' experience with primary care reported in a survey [82]; and a framework for finding health mentions online [83].

Recent developments in enriching the existing consumer vocabularies include a system developed to assist with collaborative updates an existing consumer health vocabulary [84]; a crowdsourcing approach to identifying medical terms in patient-authored texts [85]; unsupervised lexicon generation representative of the sub-language used in an online consumer community [86]; mining pairs of professional terms and their equivalent consumer terms from Wikipedia [87]; and an approach to expanding a seed vocabulary of consumer-friendly terms [88].

Towards Interventions?

The above resources and methods serve as foundation for the more complex methods that are needed to accomplish some higher-level NLP tasks listed next. When analyzing online communications, attribution to the author of a post might be very important, if, for example, clinicians would like to intervene online. Lee and colleagues discussed prevention of back pain through the detection of risk factors, as the individuals tweeting about certain activities and health problems are likely to tweet about acute back pain shortly after [89]. In this study, the attribution of back pain to the authors was defined by the use of personal pronouns. A more sophisticated approach to attribution of disorders to patients in health forums was proposed by Driscoll and colleagues, who casted disorder attribution as a classification task and used Brown clusters assignments and syntactic features [90]. Other potential interventions could be based on a potential relationship between posting to an online weigh loss forum and weight changes. Hekler and colleagues suggested that the increased use of past-tense and motion words, such as go, car, and arrive, were associated with lower weekly weights of an online forum users [91]. On the other hand, increased use of conjunctions and exclusion words (e.g., but, without, exclude) were associated with higher weights.

Language Use and Social Support

Understanding the interactions between patient discourse and social support has been investigated in the past several years in the context of online health discussion boards and online support groups. Using a range of linguistic features, Wang and colleagues [92] developed a supervised machine learning approach for predicting if a post falls into a predefined message type (e.g., positive emotional disclosure, question asking, etc.). The developed method was consistent with human judgments in establishing that when people convey their negative experiences, thoughts, and feelings, others provide them with emotional support. In another study, Vlahovic and colleagues [93] confirmed that there is a strong link between a discussion participant's satisfaction and the type of support they receive and provide.

While Wang and colleagues found that requesting support and talking about exclusively negative events triggered support from others, Lewallen and colleagues [94] did not find that greater use of negative emotions predicted peer responsiveness; however, three other factors did: greater message length, lower use of second-person pronouns, and lower use of positive emotion words.

NLP techniques have also been used to study the associations between participation and various outcomes. For instance, Zhang and colleagues [95] investigated the impact of different factors on post sentiment, as assessed automatically when using a learned, forum-specific, sentiment analysis tool. They found that there is a significant increase in sentiment of posts as patients keep on posting in time, with different patterns of sentiment trends for initial posts in threads and reply posts. Zhao and colleagues [96] leveraged sentiment analysis as well, but for the task of identifying influential users in the community. Their working hypothesis is that influence can be approximated through a user's ability to affect the sentiment of others. They proposed a novel metric that incorporates this hypothesis.

Social Media as a Source for Healthcare Quality Assessment

Social media is also an excellent venue to measure patients' perception of healthcare quality. Not surprisingly, active research in this area is ongoing. Wallace and colleagues proposed a factorial latent Dirichlet allocation (f-LDA) model to uncover patients' sentiments about important aspects of healthcare, such as interpersonal manner, technical competence, and systems issues, expressed in RateMDs reviews [97]. They showed that f-LDA predictions of positive and negative sentiment correlate well with state-level measures of quality healthcare.

Researchers examined Twitter to measure patient-perceived quality of care in UK and US hospitals, respectively. Greaves and colleagues used commercial software that relies on POS tagging, syntactic parsing, compositional sentiment lexica, and a sentiment grammar to classify tweets about hospitals as positive or negative [98]. The average sentiment about a hospital was computed as a proportion of positive tweets to the total number of tweets. The correlation between the overall patient experience score from the NHS inpatient survey and the automated Twitter sentiment analysis score was low, which might be explained by a relatively low agreement between manually rated sentiment and automated sentiment analysis. Hawkins and colleagues used publicly-available software to derive sentiment scores, and then calculated a mean sentiment score for each of 297 US hospitals with at least 50 patient experience tweets [99]. The Twitter sentiment scores did not correlate with a formal US nationwide patient experience survey and weakly correlated with the Hospital Compare 30-day hospital readmission rate. Despite weak or absent correlation with the official hospital satisfaction metrics in both studies, the authors recommended to continue monitoring these feeds to better understand the experiences of healthcare consumers.

Drawing on data from two South Korean online communities predominantly used by parents to discuss pediatric services, Jung and colleagues defined six quality factors for social media–based hospital service quality analysis and used keywords corresponding to the factors for recommending hospitals [100].

Understanding Consumer Health Questions Online

Although consumers' health information needs are well studied (primarily using search engine logs analysis), consumer-health question answering (QA) is a relatively new area, with most of the work focusing on question analysis. Roberts and Demner-Fushman analyzed several consumer and professional question answering venues and found that the form of consumer questions is highly dependent upon the individual online resource, especially in the amount of background information provided [101]. Professionals, on the other hand, provide very little background information, and often ask much shorter questions. The content of consumer questions is also highly dependent upon the resource. While professional questions commonly discuss treatments and tests, consumer questions focus disproportionately on symptoms and diseases. Further, consumers place far more emphasis on certain types of health problems (e.g., sexual health). Cohen and colleagues have shown that interactive question answering sites could efficiently address consumer health question answering through either short answers by a small number of dedicated physicians, enabling high throughput, or physician experts operating as moderators in patient forums [102]. Luo and colleagues used syntactic and semantic analysis to align a new question with the questions previously submitted to NetWellness, a website through which highly qualified volunteers provide answers to consumers' health questions [103].

7 Discussion

To counterbalance the somewhat pessimistic outlook expressed in the Circulation: Cardiovascular Quality and Outcomes editorial, which rightfully indicated there were very few actual NLP systems in daily healthcare use, we note the success stories and recent improvements in the approaches to established tasks, such as NLP support for coding for billing purposes and quality improvements, patient record summarization, as well as a growing contribution to retrospective studies and phenotyping algorithms. We also note the successful integration of an NLP-based algorithm for finding congestive heart failure in a live health information exchange [54].

To get more of the success stories for clinical NLP in practice, the NLP research community needs for EHR vendors to buy into the technology and collaborate with NLP researchers. Another important aspect of success is educating clinicians about the systems that target their activity; many clinicians indicated lack of information about CDSSs [62] in their orientation as a factor contributing to their delayed use of these systems. One of the missing elements in measuring success is the lack of appropriate evaluation metrics. Surveys are widely used to study clinicians' satisfaction with systems, but practical measures of the impact on healthcare outcomes still need to be developed.

Although we see an increasing use of publicly available tools, the pipelines that use the tools for identical purposes at different institutions, e.g., ejection fraction detection or adenoma detection, are still sometimes programmed at the institutions. Some progress has been made in porting clinical models and NLP methods [104, 105], however more work on porting pipelines with easy domain adaption needs to be done.

Seeing that community-wide challenges and the datasets they make publicly available, such as i2b2, ShARe, THYME, and TREC, do facilitate fundamental research, we need more and larger publicly available clinical text collections. We appreciate the individual researcher's efforts to make datasets and code more available as well, and hope to see even more sharing in the future.

Overall, we see three directions in clinical NLP development: patterns to share for simple tasks, more sophisticated methods yet to be developed for more complex tasks, and tasks that have yet to be addressed, and therefore are of unknown complexity. We also see that the latest NLP methods are

not used in applications: they are explored, published, and shelved. We hope the worthy new methods will get more attention in being seen through to practice. More NLP research is needed to support meeting quality measures and health information exchange and interoperability.

In the realm of identifying and processing health-related texts in social media, we see that some researchers stay within the analysis realm, but it is interesting to see a growing number of publications aspiring to interventions based on the real-time processing of online consumer-generated texts. Most of text processing in this area is using very simple, yet effective techniques from an NLP standpoint.

Acknowledgments

This work was in part supported by the Intramural Research Program of the NIH, National Library of Medicine (DDF) and award R01 GM114355 from the National Institute of General Medical Studies (NE).

We thank the anonymous reviewers and the section editors for the encouragement, thorough reviews, and helpful suggestions.

References

1. Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. J Am Med Inform Assoc 1994;1:161–74.
2. Maddox TM, Matheny MA. Natural Language Processing and the Promise of Big Data Small Step Forward, but Many Miles to Go. Circ Cardiovasc Qual Outcomes 2015 8:463–5.
3. Névéol A, Zweigenbaum P. Clinical Natural Language Processing in 2014: Foundational Methods Supporting Efficient Healthcare. Yearb Med Inform 2015;10:194–8.
4. Uzuner O, Bodnari A, Shen S, Forbush T, Pestian J, South BR. Evaluating the state of the art in coreference resolution for electronic medical records. J Am Med Inf Assoc 2012;19:786–91.
5. Uzuner Ö, Stubbs A. Practical applications for natural language processing in clinical research: The 2014 i2b2/UTHealth shared tasks. J Biomed Inform 2015 Dec;58 Suppl:S1-5.
6. Pradhan S, Elhadad N, South BR, Martinez D, Christensen L, Vogel A, et al. Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. J Am Med Inform Assoc 2015;22:143–54.
7. Pradhan S, Elhadad N, Chapman W, et al. SemEval-2014 Task 7: Analysis of clinical text. Proceedings of the International Workshop on Semantic Evaluation (SemEval); 2014. p. 54–62.
8. Elhadad N, Pradhan S, Lipsky Gorman S, et al. SemEval-2015 Task 14: Analysis of clinical text. Proceedings of the International Workshop on Semantic Evaluation (SemEval); 2015. p. 303–10.
9. Text REtrieval Conference (TREC) Proceedings. http://trec.nist.gov/proceedings/proceedings.html
10. Roberts K, Simpson M, Demner-Fushman D, Voorhees E, Hersh W. State-of-the-art in biomedical literature retrieval for clinical cases: a survey of the TREC 2014 CDS track. Inf Retr J 2015;19(1-2):113–48.
11. Tognola G, Fiszman M, Del Fiol G, Roberts A, Taboada M, Zweigenbaum P. Current perspectives on NLP for electronic medical records. In: Medinfo 2015, São Paulo, Brasil; 2015.
12. Neveol A, Dalianis H, Savova G, Zweigenbaum P. Clinical Natural Language Processing in Languages Other Than English. AMIA Annu Symp Proc 2014. p. 183-5.
13. Safran C, Bloomrosen M, Hammond WE, Labkoff S, Markel-Fox S, Tang PC, et al. Toward a National Framework for the Secondary Use of Health Data: An American Medical Informatics Association White Paper. J Am Med Inform Assoc 2007;14:1–9.
14. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet 2012;13:395–405.
15. Chen ES, Sarkar IN. Mining the electronic health record for disease knowledge. Methods Mol Biol Clifton NJ 2014;1159:269–86.
16. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. J Am Med Inform Assoc 2014;21:221–30.
17. Liao KP, Ananthakrishnan AN, Kumar V, Xia Z, Cagan A, Gainer VS , et al. Methods to Develop an Electronic Medical Record Phenotype Algorithm to Compare the Risk of Coronary Artery Disease across 3 Chronic Disease Cohorts. PLoS ONE 2015 Aug 24;10(8):e0136651
18. Castro V, Shen Y, Yu S, Finan S, Pau CT, Gainer V, et al. Identification of subjects with polycystic ovary syndrome using electronic health records. Reprod Biol Endocrinol 2015 Oct 29;13:116.
19. Liao KP, Cai T, Savova GK, Murphy SN, Karlson EW, Ananthakrishnan AN, et al. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. BMJ 2015 Apr 24;350:h1885.
20. Wei W-Q, Teixeira PL, Mo H, Cronin RM, Warner JL, Denny JC. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. J Am Med Inform Assoc 2016 Apr;23(e1):e20-7.
21. Yu S, Liao KP, Shaw SY, Gainer VS, Churchill SE, Szolovits P, et al. Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources. J Am Med Inform Assoc 2015 Sep;22(5):993-1000.
22. Luo Y, Sohani AR, Hochberg EP, Szolovits P. Automatic lymphoma classification with sentence subgraph mining from pathology reports. J Am Med Inform Assoc 2014 Sep-Oct;21(5):824-32.
23. Halpern Y, Choi Y, Horng S, Sontag D. Using Anchors to Estimate Clinical State without Labeled Data. AMIA Annu Symp Proc 2014 Nov 14;2014:606-15.
24. Doshi-Velez F, Ge Y, Kohane I. Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis. Pediatrics 2014;133:e54–63.
25. Hidalgo CA, Blumm N, Barabási A-L, Christakis NA. A Dynamic Network Approach for the Study of Human Phenotypes. PLoS Comput Biol 2009 Apr;5(4):e1000353.
26. Ho JC, Ghosh J, Sun J. Marble: High-throughput Phenotyping from Electronic Health Records via Sparse Nonnegative Tensor Factorization. Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2014. p.115–24.
27. Pivovarov R, Perotte AJ, Grave E, Angiolillo J, Wiggins CH, Elhadad N. Learning probabilistic phenotypes from heterogeneous EHR data. J Biomed Inform 2015 Dec;58:156-65.
28. Weng C, Tu SW, Sim I, Richesson R. Formal representation of eligibility criteria: a literature review. J Biomed Inform 2010 Jun;43(3):451-67.
29. Pfiffner PB, Oh J, Miller TA, Mandl KD. ClinicalTrials.gov as a data source for semi-automated point-of-care trial eligibility screening. PloS One 2014 Oct 21;9(10):e111055.
30. 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 NJ 2013 Dec;238(12):1370-8.
31. Kreuzthaler M, Schulz S, Berghold A. Secondary use of electronic health records for building cohort studies through top-down information extraction. J Biomed Inform 2015 53:188–95.
32. Ni Y, Kennebeck S, Dexheimer JW, McAneney CM, Tang H, Lingren T, et al. Automated clinical trial eligibility prescreening: increasing the efficiency of patient identification for clinical trials in the emergency department. J Am Med Inform Assoc 2015;22(1):166–78.
33. Raghavan P, Fosler-Lussier E, Elhadad N, Lai AM. Cross-narrative Temporal Ordering of Medical Events. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics; 2014. p. 998–1008.
34. Li Q, Zhai H, Deleger L, Lingren T, Kaiser M, Stoutenborough L, et al. A sequence labeling approach to link medications and their attributes in clinical notes and clinical trial announcements for information extraction. J Am Med Inform Assoc 2013;20(5):915–21.
35. Shivade C, Hebert C, Lopetegui M, de Marneffe MC, Fosler-Lussier E, Lai AM. Textual inference for eligibility criteria resolution in clinical trials. J Biomed Inform 2015 Dec;58 Suppl:S211-8.
36. Dagan I, Glickman O, Magnini B. The PASCAL Recognising Textual Entailment Challenge. In: Quiñonero-Candela J, Dagan I, Magnini B, et al., editors. Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment. Berlin Heidelberg: Springer; 2006. p. 177–90.
37. Miotto R, Weng C. Case-based reasoning using electronic health records efficiently identifies eli-

gible patients for clinical trials. J Am Med Inform Assoc 2015;22:e141–50.
38. Poulin C, Shiner B, Thompson P, Vepstas L, Young-Xu Y, Goertzel B, et al. Predicting the Risk of Suicide by Analyzing the Text of Clinical Notes. PLoS ONE 2014; 9(1):e85733.
39. Goodwin T, Harabagiu SM. A Probabilistic Reasoning Method for Predicting the Progression of Clinical Findings from Electronic Medical Records. AMIA Summits Transl Sci Proc 2015:61–5.
40. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2015. p. 1721-30.
41. Ghassemi M, Naumann T, Doshi-Velez F, Brimmer N, Joshi R, Rumshisky A, et al. Unfolding physiological state. Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2014. p. 75–84.
42. Perotte A, Ranganath R, Hirsch JS, Blei D, Elhadad N. Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis. J Am Med Inform Assoc 2015;22(4):872-80.
43. Wang G, Jung K, Winnenburg R, Shah NH. A method for systematic discovery of adverse drug events from clinical notes. J Am Med Inform Assoc 2015;22(6):1196-204.
44. Iyer SV, Harpaz R, LePendu P, Bauer-Mehren A, Shah NH. Mining clinical text for signals of adverse drug-drug interactions. J Am Med Inform Assoc 2014;21(2):353-62.
45. Henriksson A, Kvist M, Dalianis H, Duneld M. Identifying adverse drug event information in clinical notes with distributional semantic representations of context. J Biomed Inform 2015;57:333–49.
46. Liu X, Chen H. A research framework for pharmacovigilance in health social media: Identification and evaluation of patient adverse drug event reports. J Biomed Inform 2015 Dec;58:268-79.
47. Lardon J, Abdellaoui R, Bellet F, Asfari H, Souvignet J, Texier N ,et al. Adverse Drug Reaction Identification and Extraction in Social Media: A Scoping Review. J Med Internet Res 2015 Jul 10;17(7):e171
48. Sarker A, Ginn R, Nikfarjam A, O'Connor K, Smith K, Jayaraman S , et al. Utilizing social media data for pharmacovigilance: A review. J Biomed Inform 2015 Apr;54:202-12.
49. Abhyankar S, Demner-Fushman D, Callaghan FM, McDonald CJ. Combining structured and unstructured data to identify a cohort of ICU patients who received dialysis. J Am Med Inform Assoc 2014;21 Sep-Oct;21(5):801-7
50. Vijayakrishnan R, Steinhubl SR, Ng K, Sun J, Byrd RJ, Daar Z , et al. Prevalence of Heart Failure Signs and Symptoms in a Large Primary Care Population Identified Through the Use of Text and Data Mining of the Electronic Health Record. J Card Fail Jul;20(7):459-64.
51. Patterson OV, Forbush TB, Saini SD, Moser SE, DuVall SL. Classifying the Indication for Colonoscopy Procedures: A Comparison of NLP Approaches in a Diverse National Healthcare System. Stud Health Technol Inform 2015;216:614–8.
52. Kerr GS, Richards JS, Nunziato CA, Patterson OV, DuVall SL, Aujero M et al. Measuring Physician Adherence With Gout Quality Indicators: A Role for Natural Language Processing. Arthritis Care Res 2015;67(2):273-9.
53. Carrell DS, Cronkite D, Palmer RE, Saunders K, Gross DE, Masters ET, et al. Using natural language processing to identify problem usage of prescription opioids. Int J Med Inf 2015;84(12):1057-64.
54. Wang Y, Luo J, Hao S, Xu H, Shin AY, Jin B, et al. NLP based congestive heart failure case finding: A prospective analysis on statewide electronic medical records. Int J Med Inf 2015;84(12):1039-47.
55. Sullivan R, Yao R, Jarrar R, Buchhalter J, Gonzalez G. Text Classification towards Detecting Misdiagnosis of an Epilepsy Syndrome in a Pediatric Population. AMIA Annu Symp Proc 2014;2014:1082-7.
56. McCoy TH, Castro VM, Rosenfield HR, Cagan A, Kohane IS, Perlis RH. A Clinical Perspective on the Relevance of Research Domain Criteria in Electronic Health Records. Am J Psychiatry 2015;172(4):316-20
57. Perotte A, Pivovarov R, Natarajan K, Weiskopf N, Wood F, Elhadad N. Diagnosis code assignment: models and evaluation metrics. J Am Med Inform Assoc 2014;21(2):231-7.
58. Subotin M, Davis A. A System for Predicting ICD-10-PCS Codes from Electronic Health Records. Proceedings of the 2014 Workshop on Biomedical Natural Language Processing (BioNLP); 2015. p. 59–67.
59. Haas SW, Travers D, Waller A, Mahalingam D, Crouch J, Schwartz TA , et al. Emergency Medical Text Classifier: New system improves processing and classification of triage notes. Online J Public Health Inform 2014 Oct 16;6(2):e178.
60. López Pineda A, Ye Y, Visweswaran S, Cooper GF, Wagner MM, Tsui FR. Comparison of machine learning classifiers for influenza detection from emergency department free-text reports. J Biomed Inform 2015;58:60-9.
61. Yetisgen M, Klassen P, Tarczy-Hornoch P. Automating Data Abstraction in a Quality Improvement Platform for Surgical and Interventional Procedures. EGEMS (Wash DC) 2014;2(1):1114.
62. Raju GS, Lum PJ, Slack RS, Thirumurthi S, Lynch PM, Miller E, et al. Natural language processing as an alternative to manual reporting of colonoscopy quality metrics. Gastrointest Endosc 2015;82(3):512-9
63. Gawron AJ, Thompson WK, Keswani RN, Rasmussen LV, Kho AN. Anatomic and advanced adenoma detection rates as quality metrics determined via natural language processing. Am J Gastroenterol 2014;109(12):1844-9.
64. Dentler K, Numans ME, ten Teije A, Cornet R, de Keizer NF. Formalization and computation of quality measures based on electronic medical records. J Am Med Inform Assoc 2014;21(2):285-91.
65. Pai VM, Rodgers M, Conroy R, Luo J, Zhou R, Seto B. Workshop on using natural language processing applications for enhancing clinical decision making: an executive summary. J Am Med Inform Assoc 2014 Feb;21(e1):e2-5.
66. Dean NC, Jones BE, Ferraro JP, Vines CG, Haug PJ. Performance and utilization of an emergency department electronic screening tool for pneumonia. JAMA Intern Med 2013;173(8):699-701.
67. Dean NC, Jones BE, Jones JP, Ferraro JP, Post HB, Aronsky D, et al. Impact of an Electronic Clinical Decision Support Tool for Emergency Department Patients With Pneumonia. Ann Emerg Med 2015;66(5):511-20.
68. Demner-Fushman D, Seckman C, Fisher C, Thoma GR. Continual development of a personalized decision support system. Stud Health Technol Inform 2013;192:175-9.
69. Devarakonda M, Tsou C-H. Automated Problem List Generation from Electronic Medical Records in IBM Watson. Twenty-Seventh IAAI Conference; 2015. http://www.aaai.org/ocs/index.php/IAAI/IAAI15/paper/view/9516
70. Tsou C-H, Devarakonda M, Liang JJ. Toward Generating Domain-Specific / Personalized Problem Lists from Electronic Medical Records. 2015 AAAI Fall Symposium Series. http://www.aaai.org/ocs/index.php/FSS/FSS15/paper/view/11733
71. Hirsch JS, Tanenbaum JS, Lipsky Gorman S, Liu C, Schmitz E, Hashorva D et al. HARVEST, a longitudinal patient record summarizer. J Am Med Inform Assoc 2015;22(2):263-74.
72. Pivovarov R, Elhadad N. Automated methods for the summarization of electronic health records. J Am Med Inform Assoc 2015;22(5):938-47.
73. Harrison C, Jorder M, Stern H, Stavinsky F, Reddy V, Hanson H, et al. Using Online Reviews by Restaurant Patrons to Identify Unreported Cases of Foodborne Illness — New York City, 2012–2013. Morb Mortal Wkly Rep MMWR 2014;63(20):441-5.
74. Smith CA. Consumer language, patient language, and thesauri: a review of the literature. J Med Libr Assoc JMLA 2011;99(2):135-44.
75. McCormack L, Craig Lefebvre R, Bann C, Taylor O, Rausch P. Consumer Understanding, Preferences, and Responses to Different Versions of Drug Safety Messages in the United States: A Randomized Controlled Trial. Drug Saf 2016;39(2):171-84.
76. Polepalli Ramesh B, Houston T, Brandt C, Fang H, Yu H. Improving patients' electronic health record comprehension with NoteAid. Stud Health Technol Inform 2013;192:714-8.
77. Zhou X, Zheng A, Yin J, Chen R, Zhao X, Xu W, et al. Context-Sensitive Spelling Correction of Consumer-Generated Content on Health Care. JMIR Med Inform 2015;3(3):e27.
78. Kilicoglu H, Fiszman M, Roberts K, Demner-Fushman D. An Ensemble Method for Spelling Correction in Consumer Health Questions. AMIA Annu Symp Proc 2015:727-36.
79. Park A, Hartzler AL, Huh J, McDonald DW, Pratt W. Automatically Detecting Failures in Natural Language Processing Tools for Online Community Text. J Med Internet Res 2015;17(8):e212.
80. Liu Y, Xu S, Yoon H-J, Tourassi G. Extracting patient demographics and personal medical information from online health forums. AMIA Annu Symp Proc 2014 Nov 14;2014:1825-34.
81. Gupta S, MacLean DL, Heer J, Manning CD. Induced lexico-syntactic patterns improve information extraction from online medical forums. J Am Med Inform Assoc 2014 Sep-Oct;21(5):902-9.
82. Maramba ID, Davey A, Elliott MN, Roberts M, Roland M, Brown F, et al. (2015) Web-Based Textual Analysis of Free-Text Patient Experience Comments From a Survey in Primary Care. JMIR

Med Inform 2015 May 6;3(2):e20.
83. Yin Z, Fabbri D, Rosenbloom ST, Malin B. A Scalable Framework to Detect Personal Health Mentions on Twitter. J Med Internet Res 2015 Jun 5;17(6):e138.
84. Doing-Harris KM, Zeng-Treitler Q. Computer-Assisted Update of a Consumer Health Vocabulary Through Mining of Social Network Data. J Med Internet Res 2011 May 17;13(2):e37.
85. MacLean DL, Heer J. Identifying medical terms in patient-authored text: a crowdsourcing-based approach. J Am Med Inform Assoc 2013 Nov-Dec;20(6):1120-7.
86. Elhadad N, Zhang S, Driscoll P, Brody S. Characterizing the Sublanguage of Online Breast Cancer Forums for Medications, Symptoms, and Emotions. AMIA Annu Symp Proc 2014 Nov 14;2014:516-25.
87. Vydiswaran VGV, Mei Q, Hanauer DA, Zheng K. Mining Consumer Health Vocabulary from Community-Generated Text. AMIA Annu Symp Proc 2014 Nov 14;2014:1150-9.
88. Seedor M, Peterson KJ, Nelsen LA, Cocos C, McCormick JB, Chute CG, et al. Incorporating Expert Terminology and Disease Risk Factors into Consumer Health Vocabularies. Pac Symp Biocomput 2013:421-32.
89. Lee H, McAuley JH, Hübscher M, Allen HG, Kamper SJ, Moseley GL. Tweeting back: predicting new cases of back pain with mass social media data. J Am Med Inform Assoc 2016 May;23(3):644-8.
90. Driscoll P, Lipsky Gorman S, Elhadad N. Learning Attribution Labels for Disorder Mentions in Online Health Forums. Proceedings of the SIGIR Workshop on Health Search and Discovery; 2013. p. 3–6.
91. Hekler EB, Dubey G, McDonald DW, Poole ES, Li V, Eikey E. Exploring the Relationship Between Changes in Weight and Utterances in an Online Weight Loss Forum: A Content and Correlational Analysis Study. J Med Internet Res 2014 Dec 8;16(12):e254.
92. Wang Y-C, Kraut RE, Levine JM. Eliciting and Receiving Online Support: Using Computer-Aided Content Analysis to Examine the Dynamics of Online Social Support. J Med Internet Res 2015 Apr 20;17(4):e99.
93. Vlahovic TA, Wang Y, Kraut RE, Levine ML. Support Matching and Satisfaction in an Online Breast Cancer Support Community. Proc 32nd Annu ACM Conf Hum Factors Comput Syst (CHI'2014). New York: ACM Press. p. 1625–34.
94. Lewallen AC, Owen JE, Bantum EO, Stanton AL. How language affects peer responsiveness in an online cancer support group: implications for treatment design and facilitation. Psychooncology 2014 Jul;23(7):766-72.
95. Zhang S, Bantum E, Owen J, Elhadad N. Does sustained participation in an online health community affect sentiment? AMIA Annu Symp Proc 2014 Nov 14;2014:1970-9.
96. Zhao K, Yen J, Greer G, Qiu B, Mitra P, Portier K. Finding influential users of online health communities: a new metric based on sentiment influence. J Am Med Inform Assoc 2014 Oct;21(e2):e212-8.
97. Wallace BC, Paul MJ, Sarkar U, Trikalinos TA, Dredze M. A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews. J Am Med Inform Assoc 2014 Nov-Dec;21(6):1098-103.
98. Greaves F, Laverty AA, Cano DR, Moilanen K, Pulman S, Darzi A, et al. Tweets about hospital quality: a mixed methods study. BMJ Qual Saf 2014 Oct;23(10):838-46.
99. Hawkins JB, Brownstein JS, Tuli G, Runels T, Broecker K, Nsoesie EO, et al. (2015) Measuring patient-perceived quality of care in US hospitals using Twitter. BMJ Qual Saf 2016 Jun;25(6):404-13.
100 Jung Y, Hur C, Jung D, Kim M. Identifying Key Hospital Service Quality Factors in Online Health Communities. J Med Internet Res 2015 Apr 7;17(4):e90.
101 Roberts K, Demner-Fushman D. Interactive use of online health resources: A comparison of consumer and professional questions. J Am Med Inform Assoc 2016 May 4.
102 Cohen R, Elhadad M, Birk O. Analysis of free online physician advice services. PLoS One 2013;8(3):e59963.
103 Luo J, Zhang G-Q, Wentz S, Cui L, Xu R. SimQ: Real-Time Retrieval of Similar Consumer Health Questions. J Med Internet Res 2015 Feb 17;17(2):e43.
104 Carroll RJ, Thompson WK, Eyler AE, Mandelin AM, Cai T, Zink RM, et al. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. J Am Med Inform Assoc 2012 Jun;19(e1):e162-9.
105 Khalifa A, Meystre S. Adapting existing natural language processing resources for cardiovascular risk factors identification in clinical notes. J Biomed Inform2015;58 Suppl:S128–32.

Correspondence to:
Dina Demner-Fushman
National Library of Medicine, National Institutes of Health
Bldg. 38A, Room 10S-1022
8600 Rockville Pike MSC-3824
Bethesda, MD 20894, USA
Tel: +1 301 435 5320
Fax: +1 301 402 0341
E-mail: ddemner@mail.nih.gov