Functional evaluation of out-of-the-box text-mining tools for data-mining tasks

RECEIVED 23 April 2014
REVISED 17 September 2014
ACCEPTED 5 October 2014
PUBLISHED ONLINE FIRST 21 October 2014

Kenneth Jung[1], Paea LePendu[2], Srinivasan Iyer[3], Anna Bauer-Mehren[4], Bethany Percha[1], Nigam H Shah[2]

AMIA INFORMATICS PROFESSIONALS. LEADING THE WAY.
OXFORD UNIVERSITY PRESS

ABSTRACT

Objective The trade-off between the speed and simplicity of dictionary-based term recognition and the richer linguistic information provided by more advanced natural language processing (NLP) is an area of active discussion in clinical informatics. In this paper, we quantify this trade-off among text processing systems that make different trade-offs between speed and linguistic understanding. We tested both types of systems in three clinical research tasks: phase IV safety profiling of a drug, learning adverse drug–drug interactions, and learning used-to-treat relationships between drugs and indications.

Materials We first benchmarked the accuracy of the NCBO Annotator and REVEAL in a manually annotated, publically available dataset from the 2008 i2b2 Obesity Challenge. We then applied the NCBO Annotator and REVEAL to 9 million clinical notes from the Stanford Translational Research Integrated Database Environment (STRIDE) and used the resulting data for three research tasks.

Results There is no significant difference between using the NCBO Annotator and REVEAL in the results of the three research tasks when using large datasets. In one subtask, REVEAL achieved higher sensitivity with smaller datasets.

Conclusions For a variety of tasks, employing simple term recognition methods instead of advanced NLP methods results in little or no impact on accuracy when using large datasets. Simpler dictionary-based methods have the advantage of scaling well to very large datasets. Promoting the use of simple, dictionary-based methods for population level analyses can advance adoption of NLP in practice.

Key words: electronic health records, natural language processing, text mining

BACKGROUND AND SIGNIFICANCE

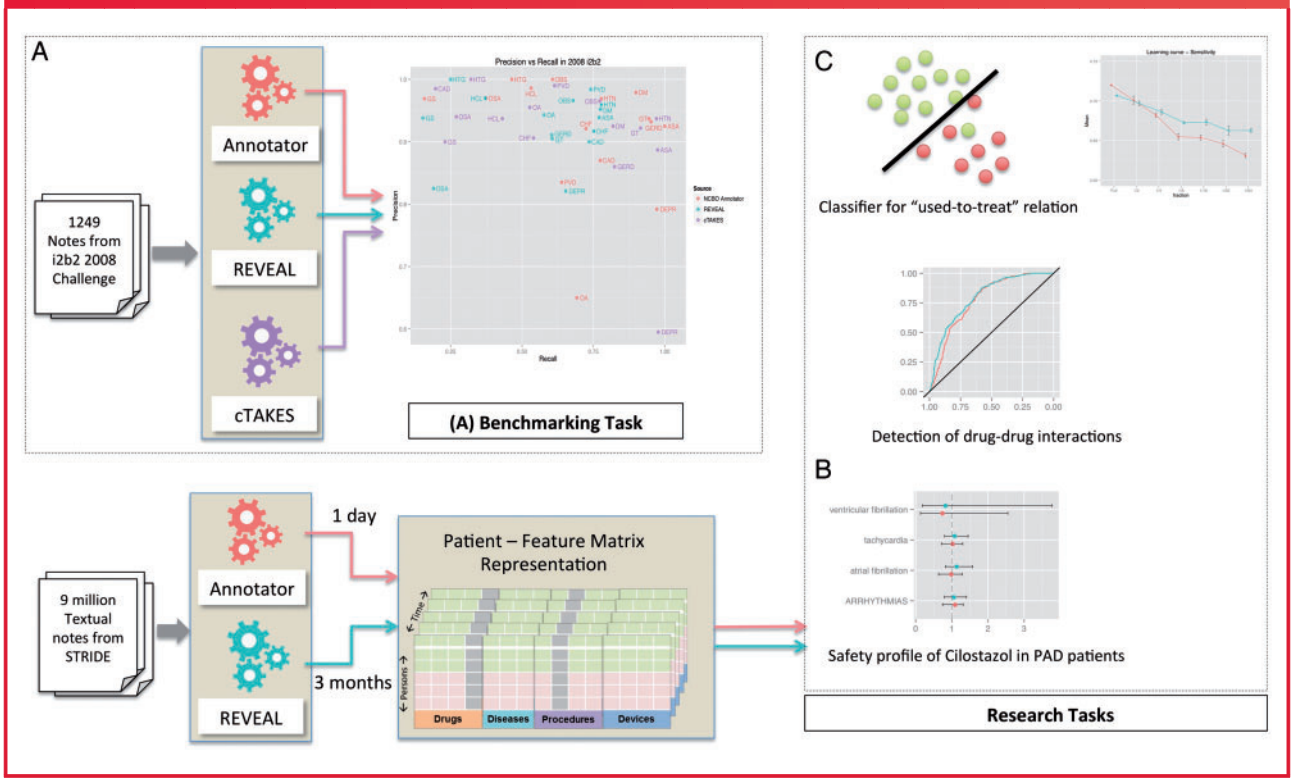
Clinical text from electronic health records (EHRs) has been used for post-marketing surveillance of drug-induced adverse events,[1–6] detection of drug–drug interactions (DDIs),[7,8] discovery and validation of clinical phenotypes,[9–11] and detection of relationships between clinical concepts, such as drug–disease treatment pairs.[12–16] Raw clinical text arguably provides the most complete picture of the state of patients at any point in time since much of the structured data in EHRs, such as administrative codes, are used primarily for purposes other than communication of key clinical information about patients, for example, billing.[17–20] However, clinical text is unstructured data, and basic questions that are easy to state in plain language are often difficult to reduce to practice, for example, find all patients who have peripheral artery disease (PAD) and who are taking cilostazol. A critical first step in the use of clinical text to address such electronic phenotyping problems is finding mentions of entities of interest, such as drugs, diseases, or laboratory values, in the text.[21,22] These may be positive mentions, indicating that the patient has the disease or is taking the drug, or negated mentions, such as when a condition is ruled out. These mentions may then be used to calculate statistics to directly address the question at hand, or as the basis for representing patients for use in data-mining approaches.[21,23,24]

A variety of natural language processing (NLP) systems that address this goal have already been developed and are in widespread use.[25–27] These may be arranged along a spectrum of complexity, from very simple, fast string matching systems to more complex systems incorporating sophisticated statistical learning methods.[28–30] Simpler systems typically provide a minimal set of information about the text, such as mentions of entities of interest and their negation status. More complex systems often provide much richer information about the text, such as part-of-speech tags and parse trees, in addition to mentions of entities of interest. However, this comes at

Correspondence to Dr Kenneth Jung, Program in Biomedical Informatics, 1265 Welch Road, MSOB X-215, MC 5479, Stanford, CA 94305-5479, USA; kjung@stanford.edu

For numbered affiliations see end of article.

Figure 1: Our investigation has three parts. (A) First, we benchmark the accuracy of the NCBO Annotator-based workflow, REVEAL, and cTAKES on the task of finding mentions of co-morbidities in the 2008 i2b2 Obesity Challenge dataset (details in figure 2). (B) Second, we evaluate the trade-off of using annotations, and the resulting patient-feature matrix, from 9 million clinical notes from the Stanford Translational Research Integrated Database Environment (STRIDE) generated using the NCBO Annotator-based workflow and the REVEAL natural language processing (NLP) system. The three research tasks are: detection of used-to-treat relationships between drugs and indications, detection of drug–drug interactions, and profiling the safety of cilostazol use in patients with peripheral artery disease (PAD). Each of these evaluations is based on previously published work; the only source of variation is the annotations used as input to the published methods. The patient-feature matrix is described in detail in online supplemental materials S2. We did not run cTAKES on the 9 million clinical notes from STRIDE because it would have required over a year to complete given our computational resources. (C) Finally, we explore the impact of dataset size on the task of detecting the used-to-treat relationship using increasingly smaller subsets of the data (details in figure 6).



a significant cost in the form of increased computational demands and, in the case of supervised learning methods, the need for labeled training text from which to learn.

In this paper, we present a systematic exploration of the trade-offs between simple term recognition and advanced NLP methods when applied to clinical text for a diverse set of use cases. We used a modified, standalone workflow based on the NCBO Annotator[31–33] and REVEAL (Health Fidelity, Palo Alto, California, USA), to find mentions of entities of interest in 9 million clinical notes from the Stanford Translational Research Integrated Database Environment (STRIDE),[34] and investigated trade-offs between these two sets of term-mentions when they were used in several clinical research tasks (figure 1).

The NCBO Annotator-based workflow is a minimalist system that relies on a large dictionary of terms, their mappings to UMLS concept IDs (concept unique identifiers, CUIs),[35] and the NegEx negation detection system,[36] to find mentions of biomedical concepts in clinical text and establish their negation status. There is little 'understanding' of the text aside from recognizing a set of words and phrases, and their negations. The workflow is deployed quite directly, using computationally efficient string matching and regular expression engines[37,38] that operate directly on the text without any pre-processing. In contrast, REVEAL, a commercial NLP system based on the popular MedLEE system,[39] performs various pre-processing steps such as parsing and word sense disambiguation en route to encoding words and phrases into UMLS codes. These steps represent a deeper understanding of the structure of the text than that of the NCBO Annotator-based workflow, and are expected to improve the quality of the annotations.

However, it took roughly 3 months to process our dataset with REVEAL, while the NCBO Annotator processed the same dataset in a few hours. It is thus worth exploring what is gained from this extra computational time when we employ the

resulting term-mentions (which we refer to as annotations) in subsequent tasks.

MATERIALS AND METHODS

Overview of our approach

Our investigation has three parts (figure 1). First, we compared the accuracy of the NCBO Annotator-based workflow and REVEAL on the task of finding mentions of entities of interest in the 2008 i2b2 Obesity Challenge dataset,[40] which is a set of discharge notes manually annotated with the presence/absence of 16 indications related to obesity. This test provides a baseline measurement of the accuracy of the systems on a task that does not, by itself, count as direct clinical research.

Second, we compared the NCBO Annotator and REVEAL in three research tasks that use the resulting annotations in different ways to address questions of greater clinical significance. Each of these evaluations is based on previously published work. The only source of variation is the annotations used as input to the published methods; in all other respects, the analyses are identical. The first task is to profile adverse events in patients with PAD who are taking cilostazol versus other PAD patients.[41] The second task is to detect adverse DDIs.[7] The third task uses mentions of drugs and diseases to detect used-to-treat relationships between drugs and diseases.[12] Finally, we explored the impact of dataset size on the last of these tasks by repeating the used-to-treat detection analysis using increasingly smaller random subsets of patients.

Data sources

We used two sources of clinical text in our evaluations. First, we used the manually annotated dataset from the 2008 i2b2 Obesity Challenge, which consists of 1292 discharge notes that have been manually annotated by domain experts with the presence/absence of 16 indications related to obesity and its comorbidities. Second, we used 9 million unstructured clinical notes from STRIDE. These notes covered approximately 1.2 million patients and 18 years of data from the Stanford hospital system and Lucile Packard Children's Hospital.

Processing clinical text

The NCBO Annotator finds mentions of biomedical concepts in unstructured text in two steps. First, it finds mentions of terms from a dictionary compiled from 22 clinically relevant ontologies, such as SNOMED-CT and MedDRA. We applied a series of syntactic and semantic suppression rules to the terms from the 22 ontologies to create a *clean lexicon*. We keep terms that are predominantly noun phrases[42] based on an analysis of over 20 million MEDLINE abstracts; we remove uninformative phrases based on term frequency analysis of over 50 million clinical documents from the Mayo Clinic[43]; and we suppress terms having fewer than four characters by default because the majority of these tend to be ambiguous abbreviations.

We then map the terms to UMLS CUIs. This mapping was tuned on the STRIDE corpus by identifying ambiguous terms that belong to more than one semantic group (drug, disease, device, procedure)[43,44] and suppressing their least likely interpretation. For example 'clip' is more likely to be a device than a drug in clinical text, so we suppress the interpretation as 'Corticotropin-Like Intermediate Lobe Peptide.' Mentions of drugs are expanded to their ingredients using RxNorm.[45] Finally, NegEx regular expressions are used to flag negative mentions (eg, 'myocardial infarction was ruled out') and to determine if a term is mentioned in the history or family history section of the note.[36] The result is a list of present, positive mentions of biomedical concepts, which are about the patient, in the input text (see online supplementary materials figure S1 for more information).

In contrast, REVEAL is a commercial text processing system based on Columbia University's MedLEE system and developed by Health Fidelity under an exclusive license.[46] We obtained a virtual machine for the REVEAL system under an academic license from Health Fidelity and used it to process the i2b2 and STRIDE datasets. Like the NCBO Annotator, REVEAL identifies negated mentions, which are ignored. Only positive mentions are used in further analysis. We used the same term to concept mapping used by the NCBO Annotator with REVEAL to assign CUIs. REVEAL-derived annotations thus also benefited from the tuning of this mapping to the STRIDE dataset.

Accuracy on the 2008 i2b2 dataset

We used the 1249 discharge notes from the 2008 i2b2 Obesity Challenge to evaluate the accuracy of the systems. Each note contains ground truth labels for whether or not each of 16 indications is explicitly mentioned in the text. The ground truth labels are at the level of entire notes instead of specific locations in each note, so our evaluation counted any positive mention of the target indications as indicating the presence of the indication in a given note. The UMLS CUIs for each indication are listed in online supplementary materials table S2. Note that for this evaluation, we count positive mentions of the descendants of these CUIs as positive mentions of the listed CUIs, and this step was performed for all of the evaluated systems. For this task, we also evaluated the accuracy of cTAKES V.3.0.0,[27] an NLP system specialized for clinical text that was originally developed at the Mayo Clinic. We included cTAKES in this evaluation because it is another widely used clinical text-processing system and is able to perform approximate matches, for example, 'joint with pain' is recognized as 'joint pain.' It was not used further in the functional evaluation because it was computationally prohibitive—our calculations based on processing the i2b2 dataset indicated that it would take well over a year to process all 9 million STRIDE notes, given our computational resources.

Safety profiling of cilostazol

Leeper *et al*[41] analyzed the electronic medical records from the Stanford clinical data warehouse using text-mining to identify 232 PAD patients taking cilostazol and a control group of 1160 patients with PAD but not taking this drug. Over a mean follow-up of 4.2 years, they observed no association between cilostazol use and any major adverse cardiovascular event including stroke, myocardial infarction, or death. We used the methods

described in detail in Leeper *et al*[41] and reviewed in online supplementary materials S1 to calculate ORs for a set of adverse events in patients with PAD who are taking cilostazol versus those who are not taking cilostazol. Mentions of clinical concepts in clinical notes from STRIDE are used to build the case (PAD and cilostazol) and control (PAD only) cohorts as described in Patrick and Li,[16] and to match them for potential confounders. The ORs are based on the positive mentions of the adverse events in each group. In this analysis, we compare the ORs and CIs obtained from annotations output from the NCBO Annotator-based workflow and REVEAL. For this evaluation, we used the 2-hop ontological expansion, described in Lependu *et al*[47] and in online supplementary materials S1, to generate sets of recognized CUIs for each adverse event, and these were used for all systems.

Adverse DDIs

Iyer *et al*[7] used mentions of drug and event concepts from clinical notes to identify DDIs leading to adverse events among 1165 drugs and 14 adverse events. Positive mentions of drugs and adverse events were used to create timelines of mentions for each patient, and these were used to calculate adjusted ORs for the drug–drug–event associations. They validated the results on a gold standard of 1698 DDIs curated from existing knowledge bases.

In this study, we detected adverse DDIs using mentions of drugs and adverse events in clinical notes from STRIDE, using methods described in detail in Iyer *et al*[7] and reviewed in online supplementary materials S1. Accuracy was tested on a gold standard set of known DDIs that was assembled in Iyer *et al* from Drugbank[48] and the Medi-Span Drug Therapy Monitoring System (Wolters Kluwer Health, Indianapolis, Indiana, USA).

The output of the NCBO Annotator-based workflow and REVEAL on the STRIDE dataset was used to calculate ORs and CIs for the drug–drug–adverse event triplets in the gold standard, and receiver operator characteristic (ROC) curves were calculated using thresholds on the ORs. As in the cilostazol study described above, we used the 2-hop ontological expansion to generate sets of recognized CUIs for each adverse event. We compare the ROC curves derived from the NCBO Annotator and the ROC curves derived from REVEAL using the method of DeLong *et al*.[49]

Learning used-to-treat relationships from clinical text

Jung *et al*[12] described a data-mining approach for identifying off-label usages using features derived from free text clinical notes and features extracted from two databases on known usage (Medi-Span and DrugBank). In that effort, we trained a highly accurate predictive model to detect novel used-to-treat relationships among 1602 unique drugs and 1472 unique indications. We validated 403 predicted uses across independent data sources and prioritized them based on drug safety and cost.

We evaluated the utility of mentions of biomedical concepts found by the NCBO Annotator-based workflow and REVEAL, respectively, in detecting used-to-treat relationships between drugs and indications, using methods described in detail in Jung *et al*[12] and reviewed in online supplementary materials S1. We followed these methods exactly, except that we used only input features derived from clinical text. This is because we are principally interested in the difference in the predictive value of annotations from the NCBO Annotator-based workflow and REVEAL.

As described in Platt *et al*,[6] a gold standard of positive and negative examples of used-to-treat relationships compiled from Medi-Span (Wolters Kluwer Health) was split randomly 4:1 into training and test sets. Features calculated from mentions of drugs and indications in the data were used as inputs to SVM classifiers. The resulting classifiers were tested on the hold out test sets. We used the e1071 library in R to fit the models, setting the misclassification cost hyperparameter for the SVMs using 10-fold cross validation in the training set.

Learning used-to-treat relationships with limited data

Intuitively, a smaller dataset will have less information about rare associations. In those circumstances, analysis may benefit from more advanced NLP than using simpler methods. We assessed the impact of dataset size on the learning task described above. To create the reduced datasets, we sampled subsets of patients without replacement from the whole population of patients in our STRIDE dataset. The reduced datasets were 1/2, 1/4, 1/8, 1/16, 1/32, and 1/64 the size of the original dataset. We repeated this sampling process 10 times for each sample size. For each sample, we used the mentions of drugs and indications found by either the NCBO Annotator-based workflow or REVEAL to construct features for SVM classifiers as before. A classifier was trained and evaluated on a held out test set for each sample as before.

RESULTS

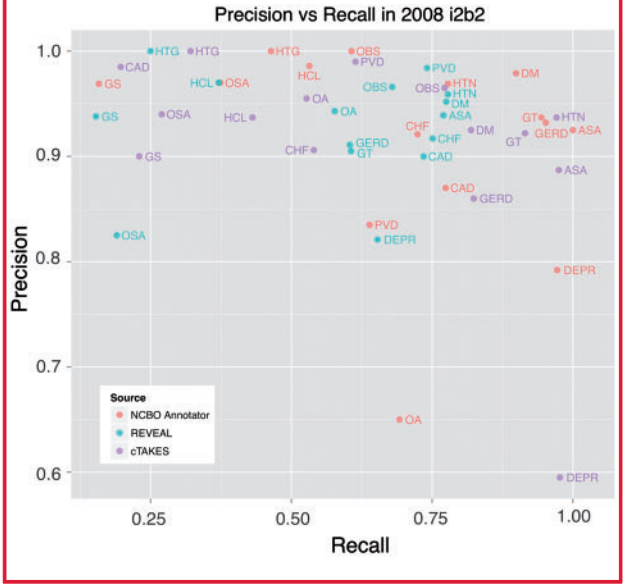
Accuracy of annotations using the 2008 i2b2 Obesity Challenge

Figure 2 summarizes the precision and recall of the NCBO Annotator, REVEAL, and cTAKES on the 2008 i2b2 Obesity Challenge dataset. Precision and recall is shown for each indication with the exception of 'venous insufficiency,' for which REVEAL was not able to detect any mentions. The full set of results is presented in online supplementary materials table S3. All systems achieve high precision, but there is considerable variation in recall, and no system is best across all indications with respect to either precision or recall. Overall, indications that are difficult to detect for one system (eg, gallstones) are difficult for all systems.

Safety profiling of cilostazol

The goal of this task is to profile adverse events associated with the use of cilostazol in patients with PAD. The output is ORs and CIs for each adverse event. Figure 3 shows results from Leeper *et al*,[41] which were obtained using the NCBO Annotator-based workflow, along with results from the same analysis performed using annotations from REVEAL. There is no significant difference in either the ORs or CIs for any of the

Figure 2: The precision and recall of the NCBO Annotator, REVEAL, and cTAKES in the 2008 i2b2 dataset is plotted here for each indication. There is considerable variation in recall across both systems and indications, but generally indications that are hard to detect are hard to detect for all systems (eg, Gallstones, labeled here as GS). There is no universally best system across all indications with respect to either precision or recall. ASA, asthma; CAD, coronary artery disease; CHF, congestive heart failure; DM, diabetes; DEPR, depression; GS, gallstones; GERD, gastro-esophageal reflux disease; GT, gout; HCL, hypercholesterolemia; HTN, hypertension; HTG, hypertriglyceridemia; OA, osteoarthritis; OBS, obesity; OSA, obstructive sleep apnea; PVD, peripheral vascular disease.



adverse events except for the event 'sudden cardiac death,' for which REVEAL found no instances in the data.

Learning adverse DDIs

The goal of this task is to use mentions of drugs and indications in clinical notes from STRIDE to detect adverse DDIs following the method of Iyer *et al*.[7] The output is a set of adjusted ORs for a gold standard set of known and negative DDIs and associated adverse events. Figure 4A shows ROC curves for the gold standard as we vary the OR threshold for signaling an adverse DDI. There is no significant difference in the ROC curves ($p = 0.275$). Figure 4B shows the area under the area under the ROC curve (AUC) for each of nine adverse events separately. For all adverse events, there is no significant difference in the AUC between the two systems. Note that this analysis excludes the adverse event 'serotonin syndrome,' for which the NCBO Annotator is significantly better than REVEAL ($p<1e-6$; see online supplementary materials figure S4).

Learning used-to-treat relationships

The goal of this task is to use mentions of drugs and indications in clinical notes from STRIDE to construct features that are useful for identifying which drugs are being used to treat which indications, according to the methods in Jung *et al*.[12] The output is a set of performance metrics for an SVM classifier, including positive predictive value, specificity, sensitivity, and F1 based on a hold out test set. Figure 5 shows the performance of classifiers trained and tested using features derived from the NCBO Annotator-based workflow and REVEAL, and shows that there is no significant difference in performance ($p = 0.29$ by McNemar's test).

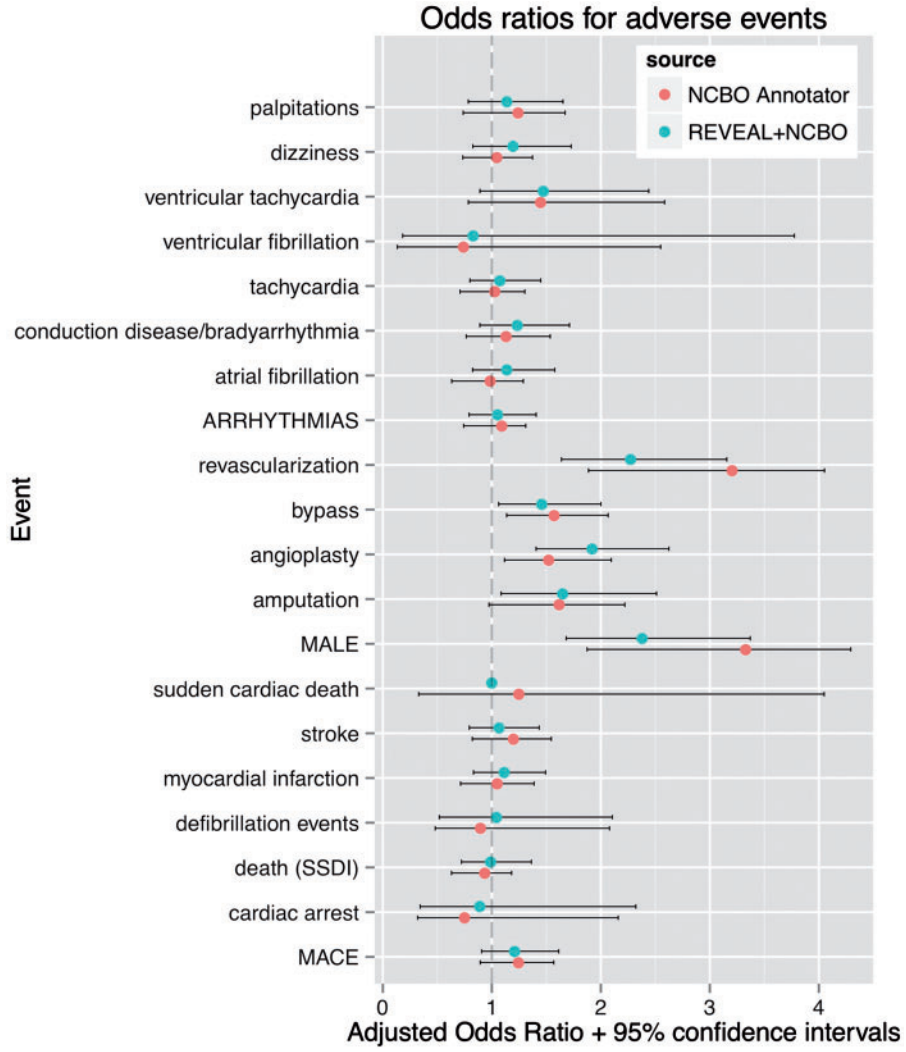
Learning used-to-treat relationships with limited data

We explored whether or not more advanced NLP methods, as embodied in REVEAL, are advantageous when data are limited so that rare associations are less well represented in the data. Figure 6 shows the relationship between dataset size and accuracy of the classifiers in the used-to-treat task. Each plot shows the mean performance and SE of the mean over 10 random samples of patients. The gap in accuracy as the dataset size decreases remains quite modest, even at only 1/64th of the dataset size. However, the classifiers trained on output from REVEAL consistently show higher sensitivity with smaller datasets starting at datasets 1/4 the size of the full dataset. This 1/4 size corresponds to roughly 2.25 million notes, which would correspond to approximately 250 000 patients if each patient had the median number of notes. These results agree with the notion that more advanced NLP will be advantageous when detecting rare associations or when data are limited.

DISCUSSION

Recognizing mentions of drugs and diseases in clinical text is a key step in using the unstructured text from EHRs to address many questions of clinical interest. In this paper, we have performed a systematic comparison of the trade-off between simple term recognition and the deeper linguistic understanding of clinical text provided by advanced NLP, as embodied by the NCBO Annotator and REVEAL, respectively. The NCBO Annotator uses mgrep and an extensive dictionary mapping strings to biomedical concepts, along with the NegEx negation detection module, to efficiently find mentions of the concepts in text. In contrast, REVEAL performs extensive preprocessing, including parsing, word sense disambiguation, and other core NLP tasks, en route to identifying mentions of drugs and diseases. It is significantly more computationally expensive than the NCBO Annotator-based workflow. For example, the average time to process one clinical note using REVEAL is 10 s, whereas with the NCBO Annotator-based workflow it is 0.01 s. Furthermore, the NCBO Annotator is a freely available tool while REVEAL is a commercial product. These tools were evaluated on a set of clinical research tasks: safety profiling of cilostazol, learning adverse DDIs, and learning used-to-treat relationships between drugs and diseases. We also evaluated the accuracy of the systems in finding positive mentions of 16 diseases in a manually annotated set of clinical notes from i2b2. We found

Figure 3: Profile of adverse events in peripheral artery disease patients with and without exposure to cilostazol. The plot shows ORs and 95% CIs calculated using annotations from the Stanford Translational Research Integrated Database Environment (STRIDE) using the NCBO Annotator-based workflow and REVEAL. There is no change in the conclusions of this analysis depending on the text processing system being used. Note that REVEAL did not find any instances of 'sudden cardiac death' in the data; for this event, we set the OR to 1. MACE, major adverse cardiac event.



little difference in accuracy between the methods in any of the three clinical research tasks.

The clinical research tasks we undertook used aggregate statistics over an entire corpus of clinical text. In such tasks, all that matters is that we accurately count mentions of drugs and indications in the text. We note that the best performing systems in the textual portion of the 2008 i2b2 Obesity Challenge were similar in spirit to the NCBO Annotator.[40] In the summary paper on that challenge,[50] Uzuner writes, 'Most of the factual and objective pieces of information were identified by simple rule-based systems armed with dictionaries of terms and negation extraction modules.' Our findings mirror that viewpoint and it seems that for the set problems we examined, having a good negation detection module, a comprehensive dictionary, and a best-effort mapping of strings to concepts are the key ingredients necessary for excellent accuracy. More advanced NLP techniques do not appear to add much value to such tasks, and take a much longer time to run. We note that using REVEAL to find strings of interest, and then using our mapping of strings to concepts consistently performed better at the i2b2 annotation task than the default mapping provided by REVEAL itself (see online supplementary materials table S3). This suggests that the quality of the mapping of strings to concepts is one of the key differences between the systems. In work evaluating extensions of cTAKES for document classification, Garla *et al*[51] found that much of their tuning consisted of adding terms and lexical variants of terms to their dictionary.

Figure 4: Detection of adverse drug–drug interactions. The analysis of Iyer *et al*[7] was carried out using either the NCBO Annotator-based workflow or REVEAL to process clinical notes from the Stanford Translational Research Integrated Database Environment (STRIDE). (A) There is no significant difference in the receiver-operator characteristic (ROC) curves ($p = 0.275$ by DeLong's test) for the two systems. (B) Area under the ROC curves (AUC) for each of nine adverse events separately. There is no significant difference in performance for any adverse event ($p > 0.05$).

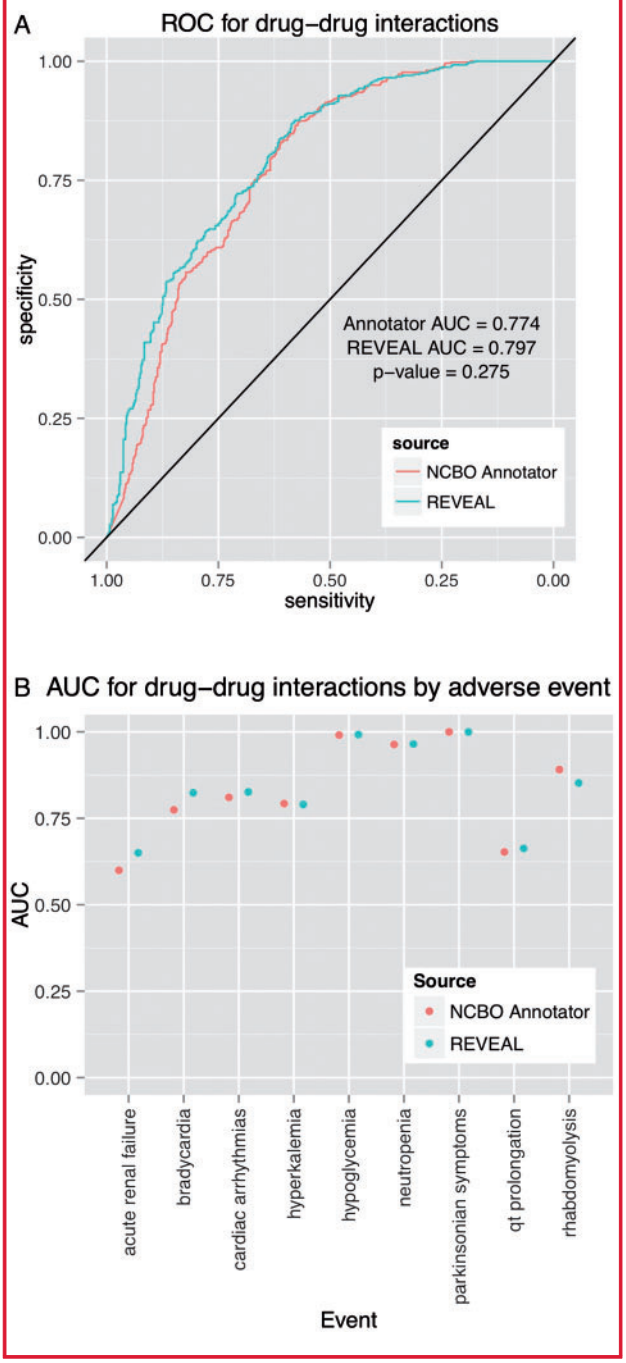
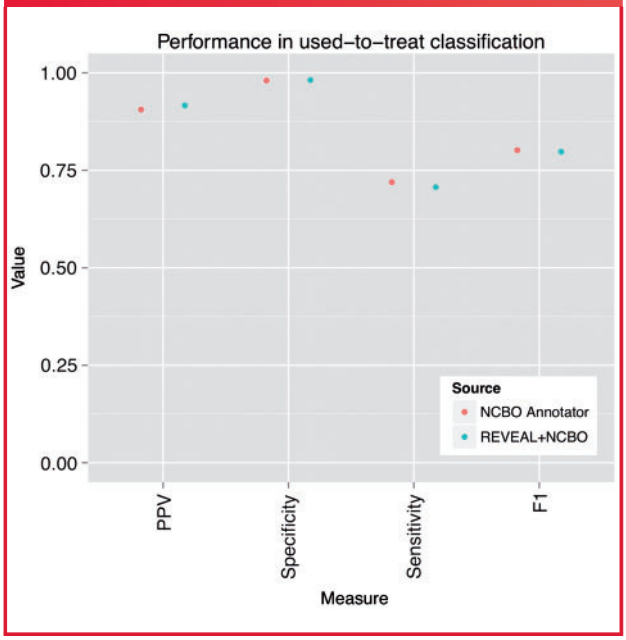
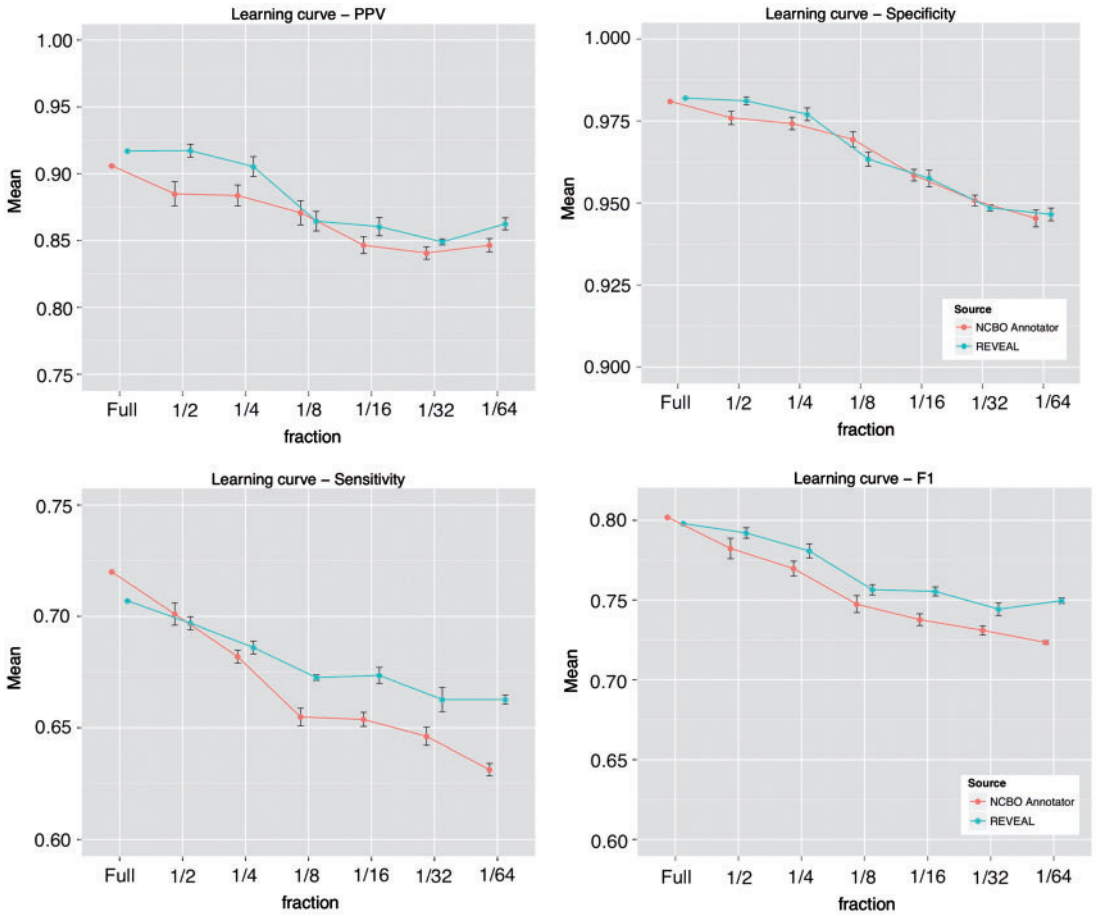


Figure 5: Detecting used-to-treat relationships. We carried out the analysis described in Jung *et al*[12] using either the NCBO Annotator or REVEAL to annotate clinical notes from the Stanford Translational Research Integrated Database Environment (STRIDE). There is no significant difference in performance between classifiers trained and tested using features derived from either ($p = 0.29$ by McNemar's test). PPV, positive predictive value.



These results do not mean that the simple approach embodied by dictionary-based approaches, such as the NCBO Annotator, is necessarily best for all problems. For instance, the used-to-treat task is formulated as a population level problem instead of asking whether a drug is being used to treat a disease as asserted in a particular note. For the latter type of question, in which we want to infer complex relationships between entities within a given text, the richer linguistic information output by a full NLP system, such as part-of-speech tags, a dependency parse, etc, can be very useful.[52] Furthermore, we found that features derived from REVEAL were more predictive of the used-to-treat relationship than features derived from the NCBO Annotator-based workflow as we decreased the size of the dataset. This is consistent with the intuition that as the dataset size decreases, rare associations may be more difficult to detect using simple text processing methods. Thus, it may be worthwhile to use a full-featured system when the dataset is relatively small. It should also be noted that while the upfront computational cost of running REVEAL on a large corpus may appear big, it is a one-time cost. And finally, we note that the systems were used 'out of the box' (ie, without any special tuning for the evaluation tasks). Given that the analysis methods were originally developed for dictionary-based approaches, they could be more effective in using that output. It is certainly possible that different methods that take explicit advantage of

Figure 6: Learning curves for the used-to-treat task. We sampled random subsets of patients and used the associated notes to generate features based on the annotations of those notes by either the NCBO Annotator or REVEAL. This was repeated 10 times for each fraction of the full Stanford Translational Research Integrated Database Environment (STRIDE) dataset. The mean performance metric across the 10 runs is plotted, along with the SEM. REVEAL has higher sensitivity in smaller datasets, and generally has higher precision/positive predictive value (PPV).



the richer information provided by more advanced NLP methods could outperform the original methods. However the gain would come at a significant computational cost, and require expertise that currently only exists in specialized NLP research teams.

These caveats notwithstanding, our results suggest that for a variety of questions of clinical interest, it is feasible to use very simple and fast approaches in lieu of more complex approaches in deriving useful information from the unstructured data in EHRs.

CONCLUSION

Widespread adoption of EHRs is creating a new source of data as a by-product of routine clinical care. These data are increasingly recognized as an asset that can be used to address problems in public health, healthcare economics, quality of care, drug safety surveillance, and even personalized medicine.[10,22,53–57] However, extracting actionable information from EHRs is a challenging problem because much of its value resides in unstructured text. NLP has been applied to this problem to good effect. In this paper, we have explored the trade-off between using a free, simple but fast term recognition system and a more advanced commercial NLP system. We evaluated the systems in a variety of tasks that address questions of clinical interest. These tasks ranged from canonical studies that use the mentions of drugs and diseases to calculate ORs, such as assessing the safety profile of a particular drug in a well-defined patient population, to a machine learning approach to finding used-to-treat relationships at the population level. We achieve the same accuracy in all three clinical research tasks using the NCBO Annotator-based workflow and REVEAL. Thus, although tasks that use detailed linguistic information about clinical text can benefit from the richer information provided by tools such as REVEAL, there are important research problems that can be tackled with much simpler and

faster dictionary-based methods. Given the increasing availability of data from EHR systems, both the variety of problems that can be addressed via text-mining as well as the amount of textual data that needs to be processed has increased significantly. We believe that it is possible to use simple faster, dictionary-based methods that scale well to very large datasets, trading off deep linguistic understanding for computational efficiency. When using very large datasets, advances in algorithms may be less important than using larger, comprehensive datasets, an observation that has been called the 'unreasonable effectiveness of data' in other endeavors.[58] We note that the continued rapid growth of clinical datasets is almost guaranteed, while significant advances in clinical NLP are more difficult. Finally, we acknowledge that for discovering novel pieces of knowledge, advanced NLP methods have a higher chance of success.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge Mark Musen for helpful discussions on the experiments and manuscript.

CONTRIBUTORS

KJ devised and ran the experiments. PL, SI, and ABM contributed code. PL, BP, and NHS contributed text to the manuscript.

FUNDING

This work was supported by NLM grant RO1 LM011369, NIGMS grant RO1 GM101430, NHGRI U54 HG004028, and the Smith Stanford Graduate Fellowship.

COMPETING INTERESTS

None.

PROVENANCE AND PEER REVIEW

Not commissioned; externally peer reviewed.

SUPPLEMENTARY MATERIAL

Supplementary material is available online at http://jamia.oxfordjournals.org/.

REFERENCES

1. Harpaz R, Haerian K, Chase HS, *et al*. Mining electronic health records for adverse drug effects using regression based methods. Proceedings of the 1st ACM International Health Informatics Symposium. Arlington, Virginia, USA. 1883008: ACM; 2010:100–7. http://dl.acm.org/citation.cfm?id=1883008
2. Haerian K, Varn D, Vaidya S, *et al*. Detection of pharmacovigilance-related adverse events using electronic health records and automated methods. *Clin Pharmacol Ther* 2012;92:228–34.
3. Wang X, Hripcsak G, Markatou M, *et al*. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. *J Am Med Inform Assoc* 2009;16:328–37.
4. Friedman C. Discovering novel adverse drug events using natural language processing and mining of the electronic health record. AMIE 2009: Proceedings of the 12th Conference on Artificial Intelligence in Medicine 2009:1–5.
5. Liu M, McPeek Hinz ER, Matheny ME, *et al*. Comparative analysis of pharmacovigilance methods in the detection of adverse drug reactions using electronic medical records. *J Am Med Inform Assoc* 2013;20:420–6.
6. Platt R, Carnahan RM, Brown JS, *et al*. The U.S. Food and Drug Administration's Mini-Sentinel program: status and direction. *Pharmacoepidemiol Drug Saf* 2012;21(Suppl 1):1–8.
7. Iyer SV, Harpaz R, Lependu P, *et al*. Mining clinical text for signals of adverse drug-drug interactions. *J Am Med Inform Assoc* 2014;21:353–62.
8. Duke JD, Han X, Wang Z, *et al*. Literature based drug interaction prediction with clinical assessment using electronic medical records: novel myopathy associated drug interactions. *PLoS Comput Biol* 2012;8:e1002614.
9. Lyalina S, Percha B, LePendu P, *et al*. Identifying phenotypic signatures of neuropsychiatric disorders from electronic medical records. *J Am Med Inform Assoc* 2013;20(e2):e297–305.
10. Pathak J, Bailey KR, Beebe CE, *et al*. Normalization and standardization of electronic health records for high-throughput phenotyping: the SHARPn consortium. *J Am Med Inform Assoc* 2013;20(e2):e341–8.
11. Davis MF, Sriram S, Bush WS, *et al*. Automated extraction of clinical traits of multiple sclerosis in electronic medical records. *J Am Med Inform Assoc* 2013;20(e2):e334–40.
12. Jung K, LePendu P, Chen WS, *et al*. Automated detection of off-label drug use. *PloS ONE* 2014;9:e89324.
13. Chen ES, Hripcsak G, Xu H, *et al*. Automated acquisition of disease drug knowledge from biomedical and clinical documents: an initial study. *J Am Med Inform Assoc* 2008;15:87–98.
14. Zhu X, Cherry C, Kiritchenko S, *et al*. Detecting concept relations in clinical text: insights from a state-of-the-art model. *J Biomed Inform* 2013;46:275–85.
15. de Bruijn B, Cherry C, Kiritchenko S, *et al*. Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. *J Am Med Inform Assoc* 2011;18:557–62.
16. Patrick J, Li M. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. *J Am Med Inform Assoc* 2010;17:524–7.
17. Poissant L, Taylor L, Huang A, *et al*. Assessing the accuracy of an inter-institutional automated patient-specific health problem list. *BMC Med Inform Decis Mak* 2010;10:10.
18. Birman-Deych E, Waterman AD, Yan Y, *et al*. Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Med Care* 2005;43:480–5.
19. Carroll RJ, Thompson WK, Eyler AE, *et al*. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. *J Am Med Inform Assoc* 2012;19(e1):e162–9.

20. Xu H, Fu Z, Shah A, *et al*. Extracting and integrating data from entire electronic health records for detecting colorectal cancer cases. *AMIA Annu Symp Proc* 2011;2011: 1564–72.
21. Boland MR, Hripcsak G, Shen Y, *et al*. Defining a comprehensive verotype using electronic health records for personalized medicine. *J Am Med Inform Assoc* 2013; 20(e2):e232–8.
22. Pathak J, Kho AN, Denny JC. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. *J Am Med Inform Assoc* 2013;20(e2):e206–11.
23. Bauer-Mehren A, Lependu P, Iyer SV, *et al*. Network analysis of unstructured EHR data for clinical research. *AMIA Jt Summits Transl Sci Proc* 2013;2013:14–18.
24. Cole TS, Frankovich J, Iyer S, *et al*. Profiling risk factors for chronic uveitis in juvenile idiopathic arthritis: a new model for EHR-based research. *Pediatr Rheumatol Online J* 2013; 11:45.
25. Chen L, Friedman C. Extracting phenotypic information from the literature via natural language processing. *Stud Health Technol Inform* 2004;107(Pt 2):758–62.
26. D'Avolio LW, Nguyen TM, Farwell WR, *et al*. Evaluation of a generalizable approach to clinical information retrieval using the automated retrieval console (ARC). *J Am Med Inform Assoc* 2010;17:375–82.
27. Savova GK, Masanz JJ, Ogren PV, *et al*. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010;17:507–13.
28. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. *J Am Med Inform Assoc* 2011;18:544–51.
29. Deshmukh VG, Meystre SM, Mitchell JA. Evaluating the informatics for integrating biology and the bedside system for clinical research. *BMC Med Res Methodol* 2009;9:70.
30. Meystre SM, Savova GK, Kipper-Schuler KC, *et al*. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform* 2008:128–44. http://www.schattauer.de/en/magazine/subject-areas/journals-a-z/imia-yearbook/imia-yearbook-2008/issue/840/manuscript/9830.html
31. Noy NF, Shah NH, Whetzel PL, *et al*. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res* 2009;37(Web Server issue):W170–3.
32. Lependu P, Iyer SV, Fairon C, *et al*. Annotation analysis for testing drug safety signals using unstructured clinical notes. *J Biomed Semantics* 2012;3(Suppl 1):S5.
33. Jonquet C, Shah NH, Musen MA. The Open Biomedical Annotator. *Summit on Translat Bioinforma* 2009;2009: 56–60.
34. Lowe HJ, Ferris TA, Hernandez PM, *et al*. STRIDE–An integrated standards-based translational research informatics platform. *AMIA Annu Symp Proc* 2009;2009:391–5.
35. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004;32(Database issue):D267–70.
36. Chapman WW, Bridewell W, Hanbury P, *et al*. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform* 2001;34: 301–10.
37. Shah NH, Bhatia N, Jonquet C, *et al*. Comparison of concept recognizers for building the Open Biomedical Annotator. *BMC Bioinformatics* 2009;10(Suppl 9):S14.
38. Unitex: http://www-igm.univ-mlv.fr/~unitex/
39. Friedman C. A broad-coverage natural language processing system. *Proc AMIA Symp* 2000:270–4. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2243979/pdf/procamiasymp00003-0305.pdf
40. Uzuner O. Recognizing obesity and comorbidities in sparse data. *J Am Med Inform Assoc* 2009;16:561–70.
41. Leeper NJ, Bauer-Mehren A, Iyer SV, *et al*. Practice-based evidence: profiling the safety of cilostazol by text-mining of clinical notes. *PloS ONE* 2013;8:e63499.
42. Xu R, Musen MA, Shah NH. A comprehensive analysis of five million UMLS metathesaurus terms using eighteen million MEDLINE citations. *AMIA Annu Symp Proc* 2010;2010: 907–11.
43. Wu S, Liu H, Li D, *et al*. UMLS term occurrences in clinical notes: a large scale corpus analysis. *J Am Med Inform Assoc* 2012;19(e1):e149–56.
44. Bodenreider O, McCray AT. Exploring semantic groups through visual approaches. *J Biomed Inform* 2003;36: 414–32.
45. Nelson SJ, Zeng K, Kilbourne J, *et al*. Normalized names for clinical drugs: RxNorm at 6 years. *J Am Med Inform Assoc Journal of the American* 2011;18:441–8.
46. Health Fidelity: http://healthfidelity.com/health-fidelity-announces-research-program-support-advanced-use-unstructured-clinical-data
47. Lependu P, Iyer SV, Bauer-Mehren A, *et al*. Pharmacovigilance using clinical notes. *Clin Pharmacol Ther* 2013;93:547–55.
48. Knox C, Law V, Jewison T, *et al*. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Res* 2011;39(Database issue):D1035–41.
49. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988;44:837–45.
50. Uzuner O, South BR, Shen S, *et al*. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc* 2011;18:552–6.
51. Garla V, Lo Re VIII, Dorey-Stein Z, *et al*. The Yale cTAKES extensions for document classification: architecture and application. *J Am Med Inform Assoc* 2011;18:614–20.
52. Goldstein I, Uzuner O. Specializing for predicting obesity and its co-morbidities. *J Biomed Inform* 2009;42: 873–86.
53. Denny JC, Bastarache L, Ritchie MD, *et al*. Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol* 2013;31:1102–10.

54. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet* 2012;13:395–405.
55. Murdoch TB, Detsky AS. The inevitable application of big data to health care. *JAMA* 2013;309:1351–2.
56. Murff HJ, FitzHenry F, Matheny ME, *et al*. Automated identification of postoperative complications within an electronic medical record using natural language processing. *JAMA* 2011;306:848–55.
57. Shah NH. Mining the ultimate phenome repository. *Nat Biotechnol* 2013;31:1095–7.
58. Halevy A, Norvig P, Pereira F. The Unreasonable Effectiveness of Data. *IEEE Intell Syst* 2009;24:8–12.

AUTHOR AFFILIATIONS

[1]Program in Biomedical Informatics, Stanford University, Stanford, California, USA
[2]Center for Biomedical Informatics Research, Stanford University, Stanford, California, USA