

Doc2Hpo: a web application for efficient and accurate HPO concept curation

Cong Liu¹, Fabricio Sampaio Peres Kury¹, Ziran Li¹, Casey Ta¹, Kai Wang^{2,3,*} and Chunhua Weng^{1,*}

¹Department of Biomedical Informatics, Columbia University, New York, NY 10032, USA, ²Center for Cellular and Molecular Therapeutics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA and ³Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

Received February 28, 2019; Revised April 26, 2019; Editorial Decision April 30, 2019; Accepted April 30, 2019

ABSTRACT

We present Doc2Hpo, an interactive web application that enables interactive and efficient phenotype concept curation from clinical text with automated concept normalization using the Human Phenotype Ontology (HPO). Users can edit the HPO concepts automatically extracted by Doc2Hpo in real time, and export the extracted HPO concepts into gene prioritization tools. Our evaluation showed that Doc2Hpo significantly reduced manual effort while achieving high accuracy in HPO concept curation. Doc2Hpo is freely available at <https://impact2.dbmi.columbia.edu/doc2hpo/>. The source code is available at <https://github.com/stormliucong/doc2hpo> for local installation for protected health data.

INTRODUCTION

Variant interpretation remains a major barrier to understanding hereditary disorders, despite the advances in sequencing technology (1). Recent studies have shown that phenotypic information helps prioritize disease-related genes and variants, and multiple computational gene-prioritization tools, such as Phenolyzer (2), Phenomizer (3), PheVor (4) and Exomiser (5), have been developed to aid the interpretation of results from clinical genome or exome sequencing. In particular, we recently demonstrated that extraction of phenotype terms from clinical texts enabled improved interpretation of sequencing data and expedited clinical diagnosis of patients with suspected Mendelian diseases (6). The Human Phenotype Ontology (HPO) (7) is one of the most popular standardized vocabularies of human phenotypic abnormalities. The latest version of HPO contains >13 000 concepts with over 156 000 annotations to human hereditary diseases. The process for clinical researchers to manually curate the standardized HPO concepts from electronic health record (EHR) narratives remains labori-

ous. A newly available tool 'Phenotero' can aid in the encoding of manually curated phenotypes with the HPO standard (8). However, the use of this tool requires extensive knowledge of HPO and depends on manual selection and insertion of HPO concepts. The process can still be error prone given the subjective nature of the process.

Several automated concept extraction tools are now available from the recent community efforts devoted to identifying clinical entities in clinical notes using natural language processing (NLP) techniques, however the accuracy of concept mapping of these tools is still unsatisfactory (9). To address the critical need for efficient and accurate HPO concept curation, we developed Doc2Hpo, a novel web application based on the human-computer collaboration design principle (10) for semi-automatically extracting standardized HPO terms from clinical text with automated concept normalization using HPO. Doc2Hpo automatically highlights clinical entities with their corresponding HPO mappings to facilitate manual review and revision if needed. Users can export the final list of HPO concepts to gene prioritization tools for further analysis.

IMPLEMENTATION

The web service is implemented in JAVA Spring Model-View-Controller (MVC) framework. The frontend annotation interface was built based on a text highlighter, mark.js (<https://markjs.io/>), written in JavaScript. The real-time HPO search function was implemented based on the API provided by <https://hpo.jax.org/app/>. The architecture of Doc2Hpo is illustrated in Figure 1A. The backend parsing engine will automatically recognize the phenotype concepts from the user input, and negation detection is available via the NegEx algorithm (11). Options such as 'allow partial match' are available for users to configure Doc2Hpo's behavior to tailor its performance to their needs. Users can choose one of five NLP engines for HPO concept extraction. The first engine is a string-based method leveraging the Aho–Corasick algorithm (12) for speedy con-

*To whom correspondence should be addressed. Tel: +1 646 734 9159; Fax: +1 212 342 0663; Email: chunhua@columbia.edu
Correspondence may also be addressed to Kai Wang. Tel: +1 267 425 9573; Fax: +1 215 590 3660; Email: wangk@email.chop.edu

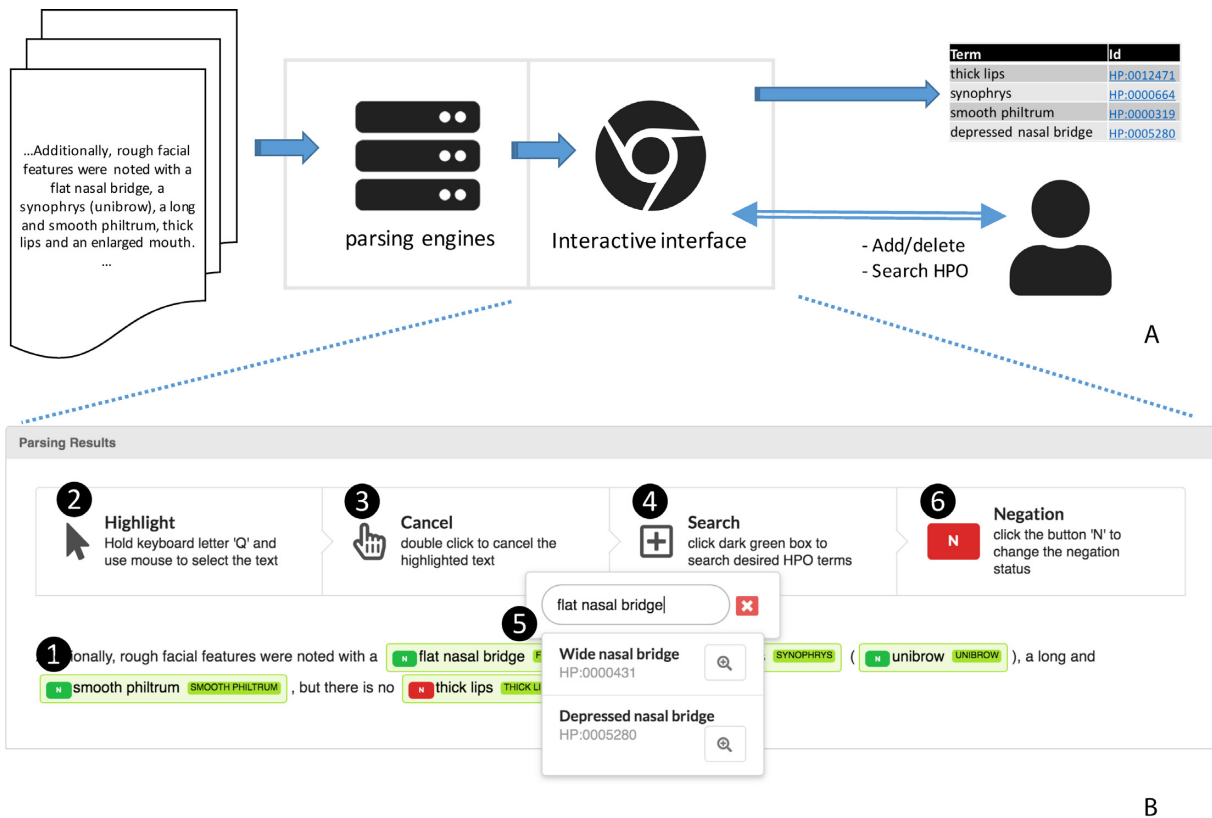


Figure 1. (A) Architecture of Doc2Hpo. (B) Interactive user interface. 1. Automated extraction includes three parts: light green highlights the clinical entity identified within the text, the term in the appended dark green box identifies the standardized HPO concept, and the 'N' button appended indicates the negation modifier (red for negation). 2–3. The user can highlight any text and add an HPO concept. The user can also double click to delete an erroneous extraction; 4–5. The user can change the standardized HPO concept by clicking the dark green box. The user could search for the desired HPO concept and make updates in real-time. 6. The user can click the 'N' button to change the negation modifier.

cept extraction. All HPO terms and their synonyms under 'phenotypic abnormality' (HP:0000118) are used for the string-based search. The second engine uses a locally configured MetaMap (13) server via the Java API to extract UMLS CUI terms. MetaMap first identifies candidate clinical terms through lexical and syntactic analysis and maps them to standard UMLS concepts. The UMLS concepts are then mapped to HPO concepts following the mapping at <http://purl.obolibrary.org/obo/hp.obo>. The third engine employs the online NCBO Annotator (14) API for HPO concept recognition. Different options for NCBO Annotator are exposed to users via the Doc2Hpo interface to customize the parsing. For instance, users would check the 'longest match only' to select only the longest matching HPO term when two different HPO terms would fit. The fourth engine is MetaMap Lite (15), a fast version of MetaMap that provides near-real time named entity recognition. Though not as rigorous as MetaMap, it is much faster. The fifth option, Ensemble by Union, considers the results generated from the previous four engines and takes the union as the final result. It lowers the false negative rate by sacrificing the execution efficiency.

INTERACTIVE USER INTERFACE

Given the complexity of biomedical NLP (16–19), particularly in named entity recognition and concept normaliza-

tion, a human-computer collaborative approach is a pragmatic solution. Doc2Hpo is designed based on this principle. One of the most useful features of Doc2Hpo is the interactive user interface that allows users to easily modify the automated extraction results. Figure 1B shows a typical parsing result and the functions provided by the user interface. Users can highlight the text to add annotations missed by machine, or double click text to delete incorrectly identified phenotypes. The dictionary look-up feature, which allows users to search for relevant concepts, distinguishes Doc2Hpo from other annotation tools. Users can simply click the dark green box to search for or modify the standard HPO concepts. This function makes Doc2Hpo most suitable for the concept recognition tasks, which differs from the named entity recognition task in the way it requires the mapping from text to standardized concepts. Users can also correct the erroneously assigned negation modifiers by clicking the prepended 'N' button. Finally, different outputs are available in a variety of popular formats to provide compatibility with various applications, which we will illustrate in detail in the Use Cases section.

EVALUATION

To evaluate the performance of Doc2Hpo, we compared the performance of Doc2Hpo-aided user curation (referred

to as ‘Doc2Hpo’) against three other approaches: manual HPO curation, fully automated HPO curation without negation detection, and fully automated HPO curation with negation detection. Three different users (two clinicians and one researcher without clinical background) were assigned 18 clinical notes (with more than 100 distinct pairs of note-HPO concepts), obtained from NewYork-Presbyterian/Columbia University Irving Medical Center. To prevent bias, each note was randomly assigned to one or two users for manual concept curation, and then assigned to the other user(s) for Doc2Hpo-aided curation. Fully-automated HPO extraction was performed using the MetaMap parsing engine within Doc2Hpo and evaluating the resulting list of HPO terms directly (i.e. without user intervention). The gold standard was determined by consensus between the two clinicians after reviewing the combined curation results from all four approaches, resulting in 108 distinct pairs of note-HPO concepts. Both accuracy and efficiency were evaluated. Our results showed that Doc2Hpo achieved the highest accuracy and efficiency (Figure 2). Doc2Hpo achieved higher precision (0.916) than manual extraction (0.861), automated extraction (0.386), and automated extraction with negation detection (0.467). Doc2Hpo also achieved higher recall (0.921) than manual extraction (0.658), automated extraction (0.774) and automated extraction with negation detection (0.761). The average F1-score for Doc2Hpo, manual extraction, automated extraction and automated extraction with negation detection were 0.913, 0.72, 0.476 and 0.524, respectively. Compared against manual curation, Doc2Hpo-assistant manual curation made the process more efficient by reducing time consumption from ~7 to ~5 min per 1000 words. In addition, we found that Doc2Hpo might help reduce the knowledge gap between clinicians and non-clinical annotators by reducing the time-consumption differences. More details on evaluation design and results are described in the Supplementary Material.

USE CASES

Genetic counselors or physicians: automated EHR phenotype-driven gene analysis

Phenotype-driven gene analysis tools, like Phenolyzer, require physicians to input a list of HPO terms to find causal genes and variants, which requires extensive knowledge of HPO. Doc2Hpo complements these gene analysis tools by minimizing manual effort in phenotype concept curation. For instance, physicians could type or paste narrative patient descriptions, and use Doc2Hpo to identify standardized HPO concepts. They could examine and manually correct the results (if necessary) to guarantee parsing quality. By clicking the ‘Phenolyzer’ link in the ‘Physicians’ section, a Phenolyzer-compatible phenotype list will be generated in the clipboard and the users will be directed to Phenolyzer, where the standardized phenotypes can be pasted into the input box for further gene prioritization analysis.

Phenotype data curator: standardized phenotypes collection

Clinical data curators need to identify and coordinate the flow of phenotypic data from EHR into clinical study

databases. Phenotype concept extraction based on manual chart review can be laborious, inaccurate, incomplete, and inconsistent from person to person. Data curators can employ Doc2Hpo in this use case to facilitate phenotype extraction. After the Doc2Hpo parsing engines extract the initial set of recommended phenotypes, data curators can customize the results using the interactive user interface to easily add, remove, or replace current concepts. Finally, they could download the resulting collection of standardized phenotypes in Excel or PDF format.

Annotators: NLP-assisted textual annotation

Doc2Hpo could serve as an annotation platform to enhance the efficiency of creating training corpora for standardized clinical concept recognition. A human annotator could attach a free-text training sample and annotate it using the interactive user interface. The automated annotated results could minimize annotation workload and provide guidance, while the dictionary lookup utility could improve efficiency and reduce errors. The final annotated results could be downloaded as a JSON file, which is a stand-off annotation file that identifies the start and end positions in the text it applies to with concrete standardized HPO concepts recognized.

DISCUSSION

We present a semi-automatic tool to enable interactive, efficient and accurate HPO term curation from clinical narratives with minimal human effort. To the best of our knowledge, this is the first web application to enable text annotation with HPO using a real-time vocabulary look-up function. Additionally, we made the source code of the web application available, so that users can run a local installation of Doc2Hpo behind institutional firewalls to annotate clinical texts with protected health information. Our main contribution is this novel platform where machine and human could work collaboratively to improve the efficiency and accuracy for standardizing HPO concept recognition. Doc2Hpo outperforms fully automated extraction approaches in precision by allowing manual detection and removal of hypothetical phenotypes, terms not related to patients, or educational text. It is also better than the manual approach because the real-time dictionary look-up utility facilitates finding more standardized terms. Furthermore, by using Doc2Hpo, non-clinical annotators could achieve a similar performance as clinicians, which indicated the potential of using Doc2Hpo to reduce the cost of data curation for clinical studies. However, due to the time and effort required to produce the gold standard, the evaluation of this study is relatively small, and large-scale evaluations are wanted to confirm the findings.

Doc2Hpo provides multiple options for parsing engines, and different parsing engines tend to have different trade-offs between accuracy and speed. In general, the ensemble parsing engine can detect more HPO concepts than the individual ones, but takes the most time to complete. Among the individual parsing engines, MetaMap can recognize more concepts but costs more processing time. By turning on negation detection, we can reduce the false positives if the

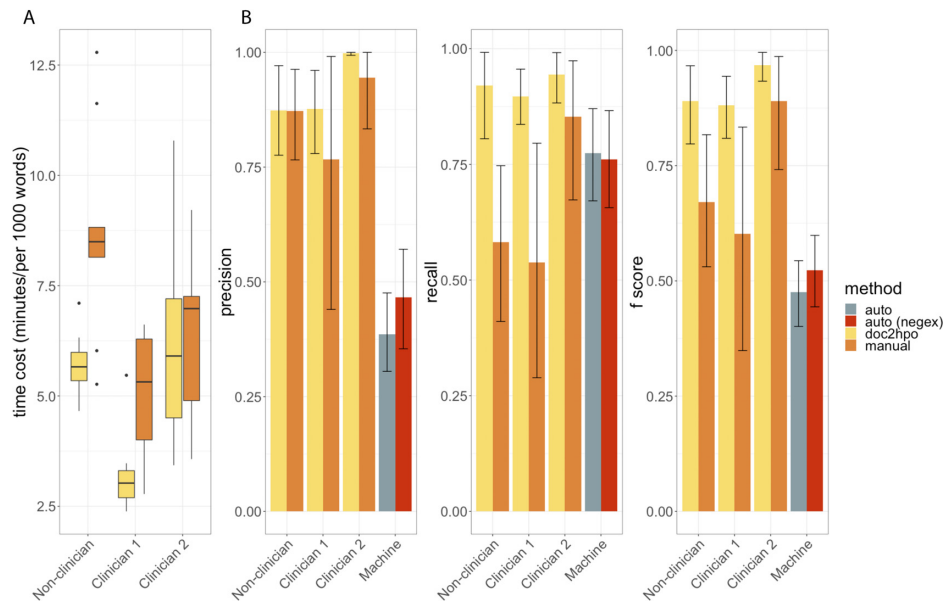


Figure 2. (A) Box and whisker plots of time consumption by using Doc2Hpo and manual curation across three different annotators. (B) The accuracy (precision, recall, and F1 score) for different approaches across different annotators. The error bars indicate the 95% confidence intervals. ‘Auto’ refers to fully automated HPO curation based on MetaMap without negation detection. ‘Auto (negex)’ refers to fully automated HPO curation based on MetaMap with negation detection. ‘Doc2hpo’ refers to Doc2Hpo-aided user curation. ‘Manual’ refers to user curation without using Doc2Hpo.

user only wants the non-negated phenotypes. Based on our experience, MetaMapLite and NCBO Annotator achieve the best balance between efficiency and accuracy.

As we showed in our evaluation, it is still challenging for current NLP systems to automatically extract a comprehensive set of phenotype terms from clinical narratives. Particular difficulties for automated parsers include recognizing spelling variations (‘renal hypertension’ – ‘renovascular hypertension’), synonymy (‘double kidney’ – ‘extra kidney’), and hyponymy/hypernymy (‘skin depigmentation’ – ‘depigmentation/hyperpigmentation of skin’). By allowing users to insert new annotations for these difficult cases that existing automated parsers cannot handle well, Doc2Hpo provides means to improve the recall rate.

Other work closely related to Doc2Hpo lie in two categories, automated concept recognition and manual text annotation for NLP tasks. The NCBO Annotator (<https://bioportal.bioontology.org/annotator>) is a popular tool to get annotations for biomedical text with classes from the ontologies (14). Users could input free-text and select the desired ontologies, and the system will return the corresponding recognized concepts. Our work differs from NCBO Annotator in three aspects. First, as reported in recent works (9,20,21), the accuracy of machine-only based concept recognition is still low. To ensure high quality results, our interactive user interface provides several easy-to-use functions enabling users to modify the extraction results if necessary. This interactive user interface is the most important feature that distinguishes Doc2Hpo from other online concept recognition tools, including Text Annotator developed by the Monarch Initiative (<https://monarchinitiative.org/annotate/text>) (22). Second, multiple parsing engines were integrated in the Doc2Hpo web application, allowing users to select the appropriate parsing

engine that fits their needs. Third, by implementing a negation detector, Doc2Hpo excludes negated mentions of other phenotypes and retains only patient-specific phenotypes. Finally, as a bridge to connect clinical notes and gene prioritization tools, Doc2Hpo provides multiple output formatting options to facilitate entry for different gene prioritization tools.

Different from other popular text annotation platforms, such as BRAT (<http://brat.nlplab.org/>) (23), ANAFORA (<https://github.com/weitechen/anafora>) (24), which mainly focus on named entity recognition and relation extraction, Doc2Hpo is designed specifically to annotate text for phenotype concept recognition. By using the dictionary lookup function, it could increase human productivity. In addition, although human judgement could not be replaced completely by automated annotation, our evaluation does show the recall of the annotation could be improved by using a first-round automated parsing result.

The Doc2Hpo server undergoes continuous improvement. In the future, we will incorporate more parsing engines, such as cTakes (25) and ClinPhen (26), to provide a more comprehensive ‘ensemble’ parsing results for further annotation. Doc2Hpo can also be easily extended for other standard phenotype or disease concept recognition ontologies such as OMIM (27) and Orphanet (28) to fulfil the fast expansion of phenotype related ontologies provided by the Monarch Initiative (22).

DATA AVAILABILITY

The original clinical notes used for the current study are available from the corresponding authors on reasonable requests and institutional approvals.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Authors contributions: C.L. designed and implemented the application. C.L. and C.W. designed the evaluation. C.L., F.K. and Z.L. performed manual and Doc2Hpo based HPO term extraction. C.L. performed the HPO term extraction using fully automatic approach and summarized the experiment results. C.L. drafted the manuscript. C.T. proof read the manuscript and made critical suggestions. C.W. and K.W. conceived and co-directed the study.

FUNDING

National Institutes of Health (NIH)/ National Library of Medicine (NLM) / National Human Genome Research Institute (NHGRI) [R01LM012895-01 to Kai Wang (K.W.), Chunhua Weng (C.W.)]. Funding for open access charge: NIH/NLM/NHGRI [R01LM012895-01].

Conflict of interest statement. None declared.

REFERENCES

- Brookes,A.J. and Robinson,P.N. (2015) Human genotype-phenotype databases: aims, challenges and opportunities. *Nat. Rev. Genet.*, **16**, 702–715.
- Yang,H., Robinson,P.N. and Wang,K. (2015) Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. *Nat. Methods*, **12**, 841–843.
- Kohler,S., Schulz,M.H., Krawitz,P., Bauer,S., Dolken,S., Ott,C.E., Mundlos,C., Horn,D., Mundlos,S. and Robinson,P.N. (2009) Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am. J. Hum. Genet.*, **85**, 457–464.
- Singleton,M.V., Guthery,S.L., Voelkerding,K.V., Chen,K., Kennedy,B., Margraf,R.L., Durtschi,J., Eilbeck,K., Reese,M.G., Jorde,L.B. *et al.* (2014) Phevor combines multiple biomedical ontologies for accurate identification of disease-causing alleles in single individuals and small nuclear families. *Am. J. Hum. Genet.*, **94**, 599–610.
- Robinson,P.N., Kohler,S., Oelrich,A., Sanger Mouse Genetics,P., Wang,K., Mungall,C.J., Lewis,S.E., Washington,N., Bauer,S., Seelow,D. *et al.* (2014) Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Res.*, **24**, 340–348.
- Son,J.H., Xie,G., Yuan,C., Ena,L., Li,Z., Goldstein,A., Huang,L., Wang,L., Shen,F., Liu,H. *et al.* (2018) Deep phenotyping on electronic health records facilitates genetic diagnosis by clinical exomes. *Am. J. Hum. Genet.*, **103**, 58–73.
- Kohler,S., Carmody,L., Vasilevsky,N., Jacobsen,J.O.B., Danis,D., Gouridine,J.P., Gargano,M., Harris,N.L., Matentzoglou,N., McMurry,J.A. *et al.* (2019) Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Res.*, **47**, D1018–D1027.
- Hombach,D., Schwarz,J.M., Knierim,E., Schuelke,M., Seelow,D. and Kohler,S. (2019) Phenotero: Annotate as you write. *Clin. Genet.*, **95**, 287–292.
- Groza,T., Kohler,S., Doelken,S., Collier,N., Oelrich,A., Smedley,D., Couto,F.M., Baynam,G., Zankl,A. and Robinson,P.N. (2015) Automatic concept recognition using the human phenotype ontology reference and test suite corpora. *Database*, **2015**, bav005.
- Friedman,C.P. (2009) A “fundamental theorem” of biomedical informatics. *J. Am. Med. Inform. Assoc.*, **16**, 169–170.
- Chapman,W.W., Bridewell,W., Hanbury,P., Cooper,G.F. and Buchanan,B.G. (2001) A simple algorithm for identifying negated findings and diseases in discharge summaries. *J. Biomed. Inform.*, **34**, 301–310.
- Aho,A.V. and Corasick,M.J. (1975) Efficient string matching: an aid to bibliographic search. *Commun. ACM*, **18**, 333–340.
- Aronson,A.R. (2001) Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc. AMIA Symp.*, **2001**, 17–21.
- Tchechmedjiev,A., Abdaoui,A., Emonet,V., Melzi,S., Jonnagaddala,J. and Jonquet,C. (2018) Enhanced functionalities for annotating and indexing clinical text with the NCBO Annotator. *Bioinformatics*, **34**, 1962–1965.
- Demner-Fushman,D., Rogers,W.J. and Aronson,A.R. (2017) MetaMap Lite: an evaluation of a new Java implementation of MetaMap. *J. Am. Med. Inform. Assoc.*, **24**, 841–844.
- Verspoor,K., Cohen,K.B., Lanfranchi,A., Warner,C., Johnson,H.L., Roeder,C., Choi,J.D., Funk,C., Malenkiy,Y., Eckert,M. *et al.* (2012) A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools. *BMC Bioinformatics*, **13**, 207.
- Hunter,L. and Cohen,K.B. (2006) Biomedical language processing: what’s beyond PubMed? *Mol. Cell*, **21**, 589–594.
- Huang,C.C. and Lu,Z. (2016) Community challenges in biomedical text mining over 10 years: success, failure and the future. *Brief. Bioinform.*, **17**, 132–144.
- Liao,K.P., Cai,T., Savova,G.K., Murphy,S.N., Karlson,E.W., Ananthakrishnan,A.N., Gainer,V.S., Shaw,S.Y., Xia,Z., Szolovits,P. *et al.* (2015) Development of phenotype algorithms using electronic medical records and incorporating natural language processing. *BMJ*, **350**, h1885.
- Funk,C., Baumgartner,W. Jr., Garcia,B., Roeder,C., Bada,M., Cohen,K.B., Hunter,L.E. and Verspoor,K. (2014) Large-scale biomedical concept recognition: an evaluation of current automatic annotators and their parameters. *BMC Bioinformatics*, **15**, 59.
- Oelrich,A., Collier,N., Smedley,D. and Groza,T. (2015) Generation of silver standard concept annotations from biomedical texts with special relevance to phenotypes. *PLoS One*, **10**, e0116040.
- Mungall,C.J., McMurry,J.A., Kohler,S., Balhoff,J.P., Borromeo,C., Brush,M., Carbon,S., Conlin,T., Dunn,N., Engelstad,M. *et al.* (2017) The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res.*, **45**, D712–D722.
- Stenotorp,P., Pyysalo,S., Topić,G., Ohta,T., Ananiadou,S. and Tsujii,J.i. (2012) *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 102–107.
- Chen,W.-T. and Styler,W. (2013) *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*. NIH Public Access, Vol. **2013**, p. 14.
- Savova,G.K., Masanz,J.J., Ogren,P.V., Zheng,J., Sohn,S., Kipper-Schuler,K.C. and Chute,C.G. (2010) Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J. Am. Med. Inform. Assoc.*, **17**, 507–513.
- Deisseroth,C.A., Birgmeier,J., Bodle,E.E., Kohler,J.N., Matalon,D.R., Nazarenko,Y., Genetti,C.A., Brownstein,C.A., Schmitz-Abe,K., Schoch,K. *et al.* (2018) ClinPhen extracts and prioritizes patient phenotypes directly from medical records to expedite genetic disease diagnosis. *Genet. Med.*, doi:10.1038/s41436-018-0381-1.
- Hamosh,A., Scott,A.F., Amberger,J.S., Bocchini,C.A. and McKusick,V.A. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33**, D514–D517.
- Rath,A., Oly,A., Dhombres,F., Brandt,M.M., Urbero,B. and Ayme,S. (2012) Representation of rare diseases in health information systems: the Orphanet approach to serve a wide range of end users. *Hum. Mutat.*, **33**, 803–808.