Clinical Narrative Analytics Challenges

Ernestina Menasalvas(✉), Alejandro Rodriguez-Gonzalez, Roberto Costumero, Hector Ambit, and Consuelo Gonzalo

Centro de Tecnología Biomédica, Universidad Politécnica de Madrid, Madrid, Spain
{ernestina.menasalvas,alejandro.rodriguezg,roberto.costumero, hector.ambit}@upm.es, chelo@fi.upm.es

Abstract. Precision medicine or evidence based medicine is based on the extraction of knowledge from medical records to provide individuals with the appropriate treatment in the appropriate moment according to the patient features. Despite the efforts of using clinical narratives for clinical decision support, many challenges have to be faced still today such as multilinguarity, diversity of terms and formats in different services, acronyms, negation, to name but a few. The same problems exist when one wants to analyze narratives in literature whose analysis would provide physicians and researchers with highlights. In this talk we will analyze challenges, solutions and open problems and will analyze several frameworks and tools that are able to perform NLP over free text to extract medical entities by means of Named Entity Recognition process. We will also analyze a framework we have developed to extract and validate medical terms. In particular we present two uses cases: (i) medical entities extraction of a set of infectious diseases description texts provided by MedlinePlus and (ii) scales of stroke identification in clinical narratives written in Spanish.

1 Introduction

Electronic Health Records (EHR) and their use in medical institutions are becoming more and more popular and its adoption has been increased during the last years [1].

Physicians complain of having tools to store information but no tools to extract information out of these records. This is a consequence of the unstructured nature of the information contained in the EHRs and consequently still remains a difficult task to perform Query and Answering processes in an accurate way [2].

Clinical narratives lack structure or they have an structure that depends on the hospital or even service of the hospital, contain abbreviations, numbers and they are written in the language of the country. Besides, concepts frequently appear in clinical notes as hypothetical, negated, or expressing temporal relationships and all these issues have to be identify to properly understand and relate concepts in EHRs. Thus traditional NLP process has to be enhaced with

new modules. In particular disambiguation of acronyms is required as the same acronym can have multiple meanings depending on the context.

Multilinguarity affects the development of these systems. Most of the existing solutions are for medical text written in English. In [3] a scheme in which emotion recognition from text through classification with the rough set theory and the support vector machines (SVMs) is proposed for Chinesse language. The experiment results showed that rough set theory and SVMs method are effective in emotion recognition. In [4] a rough set-based semi-naive Bayesian classification method is applied to dependency parsing. The rough set-based classifier is embedded with Nivre deterministic parsing algorithm to conduct dependency parsing task on a Chinese corpus showing the method has a good performance on this task.

One paramount step of the text analysis is the Named Entity Recognition that relies on ontologies. Unified Medical Language System (UMLS) [5] is composed from different ontologies and databases and are organized so every common concept contains a unique identifier. Even that translations of UMLS to different languages are available, these translations only contain a small amount of terms. On the other hand, enrichement of these vocabularies is required in order to add those terms that are specific for a disease, treatment or speciality. H2A [6] is a system composed of several software components to process clinical narratives written in Spanish.

In this paper we present two case studies about the application of Natural Language Processing methods and models. In the first case study we have applied two well-known NLP tools named MetaMap and Apache cTAKES to extract medical diagnosis terms (symptoms, signs, laboratory procedures and tests and diagnostic procedures) as an extension of a previous work [7] to compare the accuracy of both methods. The tools were applied against a set of 30 infectious diseases provided by Medline Plus. In the second case study we have applied our framework for NLP with text written in Spanish to discover scales of stroke in clinical narratives. The rest of the paper is organized as follows: In Sect. 2 the state of the art concerning NLP tools in the health sector is reviewed. The challenges of the NLP process in which we focus are briefly described in Sect. 3 while in Sect. 4 the two case studies are presented: Subsect. 4.1 presents the application of a NER application to extract concepts of infectious diseases and in Subsect. 4.2 the detection of stroke scales on Spanish narratives is presented. To conclude Sect. 5 discusses the achievements so far and presents the outlook of the future developments.

2 Background

Application of Natural Language Processing techniques to extract information from Electronic Health Records has been extensively studied as in the last decade, although most solutions are English-centric.

Electronic Health Record (EHR) contain valuable clinical information expressed in narrative form. This information is nowadays stored in digital form, but the content still lacks from structure, typos are common, etc.

The analysis of different uses of information extraction from textual documents in EHR has been analyzed in [8]. According to that publication, this extraction poses new challenges due to the problems mentioned before. The growth in the use of EHRs has generated a significant development in Medical Language Processing systems (MLP), information extraction techniques and applications [8–23].

A medical text processor is described in Friedman et al. which processes radiology reports [16]. Clinical documents are analyzed in order to transform them into terms pertaining to a controlled vocabulary. The MedLEE system is presented in [13]. MedLEE was developed to extract structures and to encode clinical information from textual patient reports. The first version of MedLEE was evaluated in chest radiology reports. MedLEE was extended to work on mammography reports and discharge summaries [14], electrocardiography, echocardiography and pathology reports [15]. The performance of MedLEE using differents lexicons (LUMLS, M-CUR, M+UMLS) was evaluated in [19].

Patient discharge summaries (PDS) were processed using MENELAS [23] to extract information from them. MENELAS can analyze reports in French, English and Dutch. cTAKES, a clinical Text Analysis and Knowledge Extraction System is introduced in [22]. cTAKES is an open-source NLP system that uses rule-based and machine learning techniques to process and extract information to support clinical research. The cTAKES components are sentence boundary detectors, tokenizers, normalizers, Part-of-Speech (PoS) taggers, shallow parsers and Named Entity Recognition (NER) annotators. HITEx (Health Information Text Extraction) [24], an open-source application based on the GATE framework, were developed to solve common problems in medical domains such as diagnosis extraction, discharge medications extraction and smoking status extraction. HITEx has been also used in [25] to extract the main diagnosis from a set of 150 discharge summaries. Co-morbidity and smoking status showed a positive performance.

MedTAS/P [10] is a system based on the Unstructured Information Management Architecture (UIMA) [26] open source framework that uses NLP techniques, machine learning and rules to map free-text pathology reports automatically into concepts represented by CDKRM (Cancer Disease Knowledge Representation Model) for storing cancer characteristics and their relationships. Fiszman et al. [12] introduced Sym Text, a NLP tool to extract relevant clinical information from radiology (Ventilation/Perfusion lung scan) reports. To evaluate the use of current NLP techniques in an automatic knowledge acquisition domain, a system is introduced in Taboada et al. [27]. The system reuses OpenNLP, Stanford parsers, SemRep and UMLS NormalizeString service as building blocks. Using an ontology, clinical practice guidelines documents are enriched. In Thomas et al. [28], an NLP program to identify patients with prostate cancer and to retrieve pathological information from their EMR is evaluated. The results show that NLP can do it accurately.

Some systems have been developed [29, 30] to process clinical text in German. An approach called Left-Associative Grammar (LAG) was used in MediTas [30],

to parse summary sections of cytopathological findings reports for a Medical Text Analysis System for German. For the German SNOMED II version another Natural Language Processing (NLP) parser is presented in [29]. The parser divides a medical term into fragments which might contain other SNOMED terms.

There are nearly 500 million Spanish speakers worldwide, however, tools to extract medical information from Spanish EHR are practically non existent. Savana Médica [31] is one of the solutions that are starting to be present in the Spanish medical environment.

The introduction of the TIDA architecture (currently renamed to H2A: Human Health Analytics) proposed for a medical decision support system was done in [32,33]. This architecture constitutes a software that analyzes text, images and the structure data from the patient in order to give the doctors answers to complex questions. Previous works on the analysis of negation detection in Spanish for medical documents has been introduced in [34] and analysis on the creation of models for performing NLP in the medical domain has been explained in [35].

3 Challenges of Extracting Valuable Information from EHR and Medical Texts

3.1 Traditional NLP Pipeline

The NLP process is a pipeline (see Fig. 1) that detects sentences, tokens, Part-of-Speech (PoS), phrases and parse tree and is able to find entities such as locations, people, or in the case of healthcare, drugs, diseases or parts of the body. The main modules of the process are the following:

- Part-of-Speech (PoS) tagging and Parsing for Spanish EHRs. NLP techniques applied to PoS and Shallow Parsing to train models are mainly based on supervised learning and that, at least for English, is a solved problem. Semi-supervised learning is also used to bootstrap the creation of corpora that is used to train the models, especially in the cases of specialized corpora as the process of annotation is very time-consuming. Unsupervised learning is currently trending as a problem to solve so annotation can be skipped. These models trained have been developed and used in different frameworks on the medical domain as seen in Sect. 2. The main challenge has to deal with the interoperability using different frameworks and most of them are English centric and some of them are proprietary. Generally, the improvement in those models depends on the corpus used to train them; and the lack of these annotated corpora in the clinical domain which are accessible, specially in languages such as Spanish, is a challenge yet to be solved.
- Named Entity Recognition (NER) is responsible of extracting the entities that are relevant in a domain and getting the relationships among them. NERs typically rely on the use of ontologies and dictionaries to detect, structure and analyse the data contained in a particular domain. UMLS is the most frequently used thesaurus in the health sector however the translation of UMLS

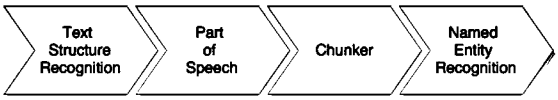


Fig. 1. Named Entity Recognition (NER) pipeline for clinical domain

to other languages rather than English, is not complete what clearly decreases the power of tools using them.

3.2 Discovering the Meaning of Numbers and Metrics

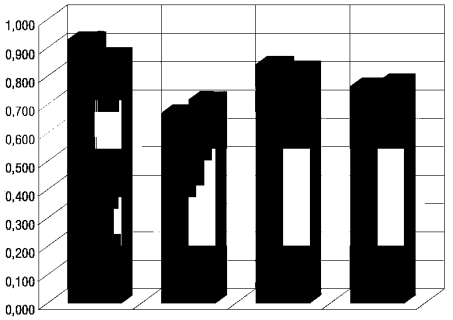
If one observes a clinical narrative, it is easy to spot many features that are particular for this kind of texts. In particular, an interesting problem is that of numbers that appear in the text which typically are followed by some metrics. This can be the case for a treatment (eg. ibuprofen 2 cp/d), laboratory tests (eg. glucose 140 mg), or blood pressure measure (eg. 140/92 mmHg). Dates is another typical feature in clinical narratives which can be absolute (eg. MRI on 14/02/2016) or relative (eg. patient suffered from headache two days ago). But numbers can also make reference to the status of the patient regarding a disease (eg. the cancer has spread and the patient is on stadium IV). In Sect. 4.2 we present an application of H2A in which the NER module developed is able to find numbers and metrics in particular to detect the scales that are reported for patients suffering a stroke.

3.3 Identifying Diagnosis Terms and Elements

The identification of diagnosis elements in medical texts is a crucial task. It is mainly used for the development of medical diagnosis systems, since nowadays it is very difficult to find open databases with information regarding the symptoms, signs or procedures to diagnose a disease (also known as diagnosis criterion; see DCM model [36]). Other relevant uses can be found in the construction of human symptoms disease networks [37], a challenging linea of research where this information is very important.

4 Materials and Methods

In what follows we present results of the experiments conducted with our framework applied for two different problems: (i) find names and symptoms of infectious diseases in English text and (ii) find the scale that is reported for a patient suffering a stroke.



	MetaMap	cTAKES
Precision	0,929	0,874
Recall	0,672	0,718
Specifty	0,843	0,826
F1	0,765	0,780

Fig. 2. Comparison of results of MetaMap and cTakes

4.1 Extracting Clinical Terms from Medical Texts

As an extension of a previous work [7] we present the resutls of using a Apache cTAKES to retrieve clinical terms from MedlinePlus texts. We have used the same set of 30 infectious diseases. The idea of this use case is to show a comparative in accuracy regarding the extraction of generalist medical terms that only affect to terms used in the diagnosis context. As has been outlined in [7] only those semantic types that belong to the classes related with diagnostic elements were used to filter. The experiment was performed using our framework in which Apache cTAKES is used as NER. We have manually analyzed the results and made a comparison between MetaMap ([7]) and Apache cTAKES.

As it can be seen in Fig. 2, the mean results are quite similar between the executions using MetaMap or cTAKES. Precision is higher on MetaMap, while recall is higher in cTAKES. Specificity and F1 score are roughly the same, the former being higher in MetaMap, while the latter is higher in cTAKES. The main differences are found in the analysis of individual diseases. cTAKES typically performs better on laboratory or test results or locating rare symptoms, but increases in most cases the number of false positives, incorrectly annotating several elements as findings. In this case, it could also be relevant that the number of true negatives is higher because the NLP process annotates more elements, but the validation usually classifies them correctly. The tools analyzed have a good performance in general in terms of NER process. However, the general behaviour could be improved by adding specific vocabularies of acronyms or complex terms to the validation terms. Future work will be focused on the analysis of further NLP tools as well the creation of hybrid approaches where more than one NLP tool could be applied to capture the knowledge within the texts.

4.2 Stroke Scales Detection in Clinical Narratives in Spanish

Neurologists have defined different protocols to be able to determine the severity of their patients when they are affected by a stroke. This is reflected on the report as values of scale of the stroke. There are different scales: Barthel, Modified Rankin Scale, National Institute of Health Stroke Scale (NIHSS), Canadian

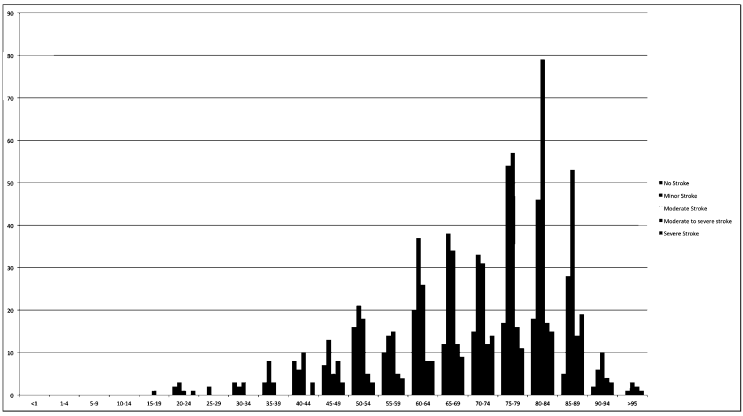


Fig. 3. Severity of the patients extracted from clinical notes

Neurological Scale, ... In order to detect these scales in the narrative we have improved the NLP process in particular enhancing the NER for Spanish narratives. In particular the following functionalities have been added:

- Detection of different possible metrics, such as the detection of the doses of a particular drug, the doses that the patient must take of it, or the different values that are indicative of different values in a laboratory test.
- Detection of contextual temporal information, retrieving absolute and relative dates that allow the automatic correlation of events in the order that happened and causation could be analyzed after when associating these dates to particular entities.
- Detection of particular scales or indexes that indicate the severity of the patient or a grade in a particular condition, like the values of the different stroke scales for patients, or the stadium of the cancer patients.

As an example of use case, EHR can be automatically analyzed to detect stroke scales such as: NIHSS, Modified Rankin Scale, Canadian Neurological Scale or the Barthel Index. These scales are typically used to determine the severity of a patient related to a stroke or the functional and neurological status of the patient after suffering from that condition. These scales are automatically extracted from the free-text written by doctors so their value and an interpretation can be given.

Figure 3 shows a possible application generating graphical distributions by age and severity of patients records according to their NIHSS values to analyze a given population.

5 Conclusions and Future Work

Clinical narratives analysis lays on the root of personalized medicine. In this paper some of the challenges that narrative analytics has to face have been

reviewed. In the coming years we will witness the arousal of new systems that would integrate text analysis as part of the discovery processes. We have analyzed the possibility of extending the NERs with information that can be required in a particular medical service. We haven also analyse the importance of NER in english texts.

The future work will go into extension of NERs both for English and Spanish language that are enriched with context aware semantics. The application to other languages is no doubt equally important.

References

1. Ben-Assuli, O.: Electronic health records, adoption, quality of care, legal and privacy issues and their implementation in emergency departments. Health Policy **119**(3), 287–297 (2015)
2. Hanauer, D.A., Mei, Q., Law, J., Khanna, R., Zheng, K.: Supporting information retrieval from electronic health records: a report of university of michigans nine-year experience in developing and using the electronic medical record search engine (EMERSE). J. Biomed. Inf. **55**, 290–300 (2015)
3. Teng, Z., Ren, F., Kuroiwa, S.: Emotion recognition from text based on the rough set theory and the support vector machines. In: 2007 International Conference on Natural Language Processing and Knowledge Engineering, pp. 36–41. IEEE (2007)
4. Ji, Y., Shang, L., Dai, X., Ma, R.: Apply a rough set-based classifier to dependency parsing. In: Wang, G., Li, T., Grzymala-Busse, J.W., Miao, D., Skowron, A., Yao, Y. (eds.) RSKT 2008. LNCS (LNAI), vol. 5009, pp. 97–105. Springer, Heidelberg (2008). doi:10.1007/978-3-540-79721-0_18
5. Humphreys, B.L., Lindberg, D.A.: The UMLS project: making the conceptual connection between users and the information they need. Bull. Med. Libr. Assoc. **81**(2), 170 (1993)
6. Rodriguez, A., Gonzalo, C., Menasalvas, E., Costumero, R., Ambit, H.: H2a - human health analytics: a natural language processing system for electronic health records. In: Proceedings of the AMIA Symposium. IJCRS-Chile (2016, to appear)
7. Rodríguez-González, A., Martínez-Romero, M., Costumero, R., Wilkinson, M.D., Menasalvas-Ruiz, E.: Diagnostic knowledge extraction from medlineplus: an application for infectious diseases. In: Overbeek, R., Rocha, M.P., Fdez-Riverola, F., Paz, J.F. (eds.) 9th International Conference on Practical Applications of Computational Biology and Bioinformatics. AISC, vol. 375, pp. 79–87. Springer, Heidelberg (2015). doi:10.1007/978-3-319-19776-0_9
8. Meystre, S.M., Savova, G.K., Kipper-Schuler, K.C., Hurdle, J.F., et al.: Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med. Inform. **35**, 128–144 (2008)
9. Christensen, L.M., Haug, P.J., Fiszman, M.: Mplus: a probabilistic medical language understanding system. In: Proceedings of the ACL 2002 Workshop on Natural Language Processing in the Biomedical Domain, vol. 3, pp. 29–36. Association for Computational Linguistics (2002)
10. Coden, A., Savova, G.K., Sominsky, I.L., Tanenblatt, M.A., Masanz, J.J., Schuler, K., Cooper, J.W., Guan, W., de Groen, P.C.: Automatically extracting cancer disease characteristics from pathology reports into a disease knowledge representation model. J. Biomed. Inf. **42**(5), 937–949 (2009)

11. Doan, S., Mike Conway, T., Phuong, M., Ohno-Machado, L.: Natural language processing in biomedicine: a unified system architecture overview. arXiv preprint arXiv:1401.0569 (2014)
12. Fiszman, M., Haug, P.J., Frederick, P.R.: Automatic extraction of pioped interpretations from ventilation/perfusion lung scan reports. In: Proceedings of the AMIA Symposium, pp. 860–864 (1998)
13. Friedman, C., Hripcsak, G., DuMouchel, W., Johnson, S.B., Clayton, P.D.: Natural language processing in an operational clinical information system. Nat. Lang. Eng. **1**(01), 83–108 (1995)
14. Friedman, C.: Towards a comprehensive medical language processing system: methods and issues. In: Proceedings of the AMIA Annual Fall Symposium, p. 595. American Medical Informatics Association (1997)
15. Friedman, C.: A broad-coverage natural language processing system. In: Proceedings of the AMIA Symposium, p. 270. American Medical Informatics Association (2000)
16. Friedman, C., Alderson, P.O., Austin, J.H., Cimino, J.J., Johnson, S.B.: A general natural-language text processor for clinical radiology. J. Am. Med. Inf. Assoc. **1**(2), 161–174 (1994)
17. Friedman, C., Hripcsak, G.: Natural language processing and its future in medicine. Acad. Med. **74**(8), 890–895 (1999)
18. Friedman, C., Knirsch, C., Shagina, L., Hripcsak, G.: Automating a severity score guideline for community-acquired pneumonia employing medical language processing of discharge summaries. In: Proceedings of the AMIA Symposium, p. 256. American Medical Informatics Association (1999)
19. Friedman, C., Liu, H., Shagina, L., Johnson, S., Hripcsak, G.: Evaluating the UMLS as a source of lexical knowledge for medical language processing. In: Proceedings of the AMIA Symposium, p. 189. American Medical Informatics Association (2001)
20. Goryachev, S., Sordo, M., Zeng, Q.T.: A suite of natural language processing tools developed for the I2B2 project. In: AMIA Annual Symposium Proceedings, vol. 2006, p. 931. American Medical Informatics Association (2006)
21. Hripcsak, G., Austin, J.H.M., Alderson, P.O., Friedman, C.: Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports 1. Radiology **224**(1), 157–163 (2002)
22. Savova, G.K., Masanz, J.J., Ogren, P.V., Zheng, J., Sohn, S., Kipper-Schuler, K.C., Chute, C.G.: Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. J. Am. Med. Inf. Assoc. **17**(5), 507–513 (2010)
23. Zweigenbaum, P.: Menelas: an access system for medical records using natural language. Comput. Method Prog. Biomed. **45**(1), 117–120 (1994)
24. Goryachev, S.: Hitex manual. https://www.i2b2.org/software/projects/hitex/hitex_manual.html
25. Zeng, Q.T., Goryachev, S., Weiss, S., Sordo, M., Murphy, S.N., Lazarus, R.: Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC Med. Inf. Decis. Making **6**(1), 30 (2006)
26. Ferrucci, D., Lally, A.: UIMA: an architectural approach to unstructured information processing in the corporate research environment. Nat. Lang. Eng. **10**(3–4), 327–348 (2004)
27. Taboada, M., Meizoso, M., Martínez, D., Riaño, D., Alonso, A.: Combining open-source natural language processing tools to parse clinical practice guidelines. Expert Syst. **30**(1), 3–11 (2013)

28. Thomas, A.A., Zheng, C., Jung, H., Chang, A., Kim, B., Gelfond, J., Slezak, J., Porter, K., Jacobsen, S.J., Chien, G.W.: Extracting data from electronic medical records: validation of a natural language processing program to assess prostate biopsy results. World J. Urology **32**(1), 99–103 (2014)
29. Hohnloser, J.H., Holzer, M., Fischer, M.R., Ingenerf, J., Günther-Sutherland, A.: Natural language processing, automatic snomed-encoding of free text: An analysis of free text data from a routine electronic patient record application with a parsing tool using the german snomed ii. In: Proceedings of the AMIA Annual Fall Symposium, p. 856. American Medical Informatics Association (1996)
30. Pietrzyk, P.M.: A medical text analysis system for german-syntax analysis. Method Inf. Med. **30**(4), 275–283 (1991)
31. Savana Médica: Savana médica (2015)
32. Costumero, R., Gonzalo, C., Menasalvas, E.: TIDA: a spanish EHR semantic search engine. In: Saez-Rodriguez, J., Rocha, M.P., Fdez-Riverola, F., De Paz, J.F., Santana, L.F. (eds.) PACBB 2014. AISP, vol. 294, pp. 235–242. Springer, Heildelberg (2014)
33. Costumero, R., Garcia-Pedrero, A., Sánchez, I., Gonzalo, C., Menasalvas, E.: 1 electronic health records analytics: natural language processing and image annotation. In: Big Data and Applications, p. 1 (2014)
34. Costumero, R., Lopez, F., Gonzalo-Martín, C., Millan, M., Menasalvas, E.: An approach to detect negation on medical documents in Spanish. In: Ślęzak, D., Tan, A.H., Peters, J.F., Schwabe, L. (eds.) BIH 2014. LNCS (LNAI), vol. 8609, pp. 366–375. Springer, Heidelberg (2014). doi:10.1007/978-3-319-09891-3_34
35. Costumero, R., García-Pedrero, Á., Gonzalo-Martín, C., Menasalvas, E., Millan, S.: Text analysis and information extraction from Spanish written documents. In: Ślęzak, D., Tan, A.-H., Peters, J.F., Schwabe, L. (eds.) BIH 2014. LNCS (LNAI), vol. 8609, pp. 188–197. Springer, Heidelberg (2014). doi:10.1007/978-3-319-09891-3_18
36. Rodríguez-González, A., Alor-Hernández, G.: An approach for solving multi-level diagnosis in high sensitivity medical diagnosis systems through the application of semantic technologies. Comput. Biol. Med. **43**(1), 51–62 (2013)
37. Zhou, X., Menche, J., Barabási, A.-L., Sharma, A.: Human symptoms-disease network. Nat. Commun. **5** (2014)