Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 7.056

IJCSMC, Vol. 9, Issue. 12, December 2020, pg.30 – 40

Data Science Applications inside Healthcare

Dr. K V N R Sai Krishna
Krishna Swaraj Educational Society, Repalle, drkvnrsaikrishna@gmail.com

Dr. A. Srinivas Rao
ALIET, VIJAYAWADA, akella.srinivas08@gmail.com

DOI: 10.47760/ijcsmc.2020.v09i12.005

Abstract: ***The massive amounts of data that are generated in the healthcare process and stored in electronic health record (EHR) systems are an underutilized resource that, with the use of data science applications, can be exploited to improve healthcare. To foster the development and use of data science applications in healthcare, there is a fundamental need for access to EHR data, which is typically not keenly available to researchers and developers. A relatively rare exception is the large EHR database, comprising information commencing further than two million patients that has been prepared available to an inadequate group of researchers. We will explain a number of data science applications that have been developed using this database, signifying the prospective recycle of EHR data to bear healthcare and public health actions, as well as make easy medical research. However, in categorize to understand the full feasible of this source; it requirements to be completed available to a larger population of researchers, as well as to industry actors. Aggregate data will be feed keen on a pipeline for open e-access, while non-aggregated data will be provided to researchers within an ethical permission framework. We believe that prospective to encourage on the rise diligence in the region of the development of data science applications that will in due course increase the effectiveness and helpfulness of healthcare Sector.***

Keywords: electronic health record, communications, data mining, text mining, Healthcare

Introduction

Data formed in the healthcare position is very precious for further analysis and development of enhanced healthcare processes, such as real-time monitoring, verdict support, and analytical analytics. Electronic health record (EHR) systems are used in almost all healthcare institutions.

An invaluable opportunity for secondary data use and growth of systems to give support to clinicians in their daily work, hospital managements in their occupation on process and healthcare deliverance improvements, and researchers in their work. Resources for health and medical research are at present available through bio banks and countrywide registers such as cancer registers and reason of death registers for researchers with appropriate moral authorization.

Extreme advance of tools and techniques in the preceding few years to routinely process an assortment of statistics sources because of the digitization of the world, to enable further analysis and tool development. Case in point, as is extensively known, the Internet contains information in various formats, and a number of systems have been industrial to make this information enthusiastically on hand for easy access, such as looking into and in order pulling out tools.

The move to digitized solutions has also in use position in healthcare. The machinery have, however, not been developed at the same pace.

On essential reason is that the health data has not been clearly available for the research society and attentiveness in order to create such tools. Health record data contains aware of in sequence about individuals an aspect that is extremely important and that requires particular considerations. To deal with these issues, we recommend developing a communications that enables access to recognized EHR data for extra investigation and arrangement growth.

Communications will consist of a variety of preprocessing tools, and will consist of two pipelines: one providing access to ordered aggregated and completely de-identified data, and one requiring ethical authorization before access to creative data is provided. This communications will be based on a large clinical database, the EPR (Electronic Patient Record) quantity, which has been collected. The EPR quantity contains over two million patients from all medical and surgical departments from the entire hospital both inpatient and outpatient records written by several different professionals at Hospital.

The quantity is identified with regard to names of patients and private identity numbers. The private identity number has been replaced by a serial number to ensure that the patient can be followed through the care process. The database contains both structured data. such as age, gender, ICD-10 diagnosis codes, ATC-drug codes, blood and lab-oratory values, admission and

discharge dates, timestamps – and unstructured data (free text), e.g. daily notes by clinicians and discharges summaries.

E H R Resources for System Development

Internationally, some research groups have been able to obtain access to health record data from one or two clinics, but almost never from a whole hospital or city council. Moreover, access is usually limited only to the research group, which limits reproducibility and generalizability of research findings. Access to this type of data is limited mostly due to legal reasons, but also because such large repositories are often compound and not easy to dig out data from.

In particular, the parts of the EHRs that are written in free text, such as expulsion summaries and daily notes, are often most difficult to obtain access to given their susceptible nature, but compose a large part of the healthcare documentation.

Some huge patient record databases or text collections are accessible for research, including the i2b2 quantity contains of several clinical in English that has been used in a number of shared challenges. CMC quantity, containing 2,216 patient records in English, MIMIC II database, which consists of 30,000 intensive care patient records written in English.

A Finnish clinical corpus, containing 2,800 sentences from nursing notes and finally. Database, containing 11 million English patient records from general practices. Both academia and industry have developed methods within computer science, statistics, computational linguistics and machine learning. This is an evolving research area also called data science - to process abundant data and produce meaningful information

Data Science Applications for Healthcare

Healthcare Associated Infections fall injuries and bedsores – in total three million patients yearly. Such prolong the treatment of the patient, because suffering for the patient, and is costly for society, with its ten million inhabitants, it is estimated that are responsible for 750,000 extra healthcare days at the hospital, costing an additional of 700 million yearly, without taking into account the suffering of the patients EPR quantity has been used for several research projects that are of practical importance for healthcare.

These projects have included work on healthcare associated infections detection, detection of a post-marketing setting, text simplification of the EHRs for public, automatic ICD-10 diagnosis code assignment, mining of cancer records and pathology reports for future improvement of cancer screening, and co-morbidity studies. For the successful development of such

applications, basic text processing tools are needed. Clinical notes in EHRs are to process for several reasons: they contain a large amount of misspellings, non-standard words and abbreviations, incomplete sentences, and medical jargon.

Therefore, we have developed a set of basic tools to process clinical text written in Local languages. These include factuality level classification, negation detection, spelling error detection, abbreviation normalization, named entity recognition as well as tools for expanding medical vocabularies.

We have also initiated studies on characterizing the domain-specific language in this type of text and performed studies on how well general language tools and techniques work on clinical notes, such as syntactic parsers and distributional semantic models studies that are important for the future development of tools adapted for this domain.

The development of these tools have also involved the creation of seven reference standards, manually annotated for protected health information , factuality levels of diagnostic expressions, clinical named entities, indications and cervical cancer symptoms, classifications of healthcare associated infections and clinical abbreviations. Many of the above mentioned tools are trained on the annotated corpora. We would like to share these valuable resources with other researchers.

Automatic Surveillance of Healthcare-Associated Infections

A healthcare-associated infection is an infection obtained by a patient during healthcare treatment. There is a requirement to report annually the number of healthcare-associated infection in each hospital, which is currently carried out in one of two ways:

by compulsory reporting of healthcare-associated infection cases, but also through so called Point Prevalence Measurements , which are carried out twice a year at all hospitals. Point Prevalence Measurements are conducted manually by assessing all the patients admitted on one particular day and deciding whether those patients have surveyed from a healthcare-associated infection.

The estimates obtained through Point Prevalence Measurements are not very reliable due to the limited sample size: only 1-2% of all patients admitted during a year are analyzed. Measurements made more frequently would give healthcare institutions. We have developed several prototype tools for detecting HAIs in EHRs. The selected patients can thereafter be assessed by a clinician. The tool is trained on health records that have been manually annotated, or classified, by a physician. The system has access to the clinical text, body temperature, drug

lists and microbiology reports; it obtains 87% recall and 83% precision using the random forest algorithm.

In Figure 1 a tentative system for HAI surveillance is depicted. The sys-tem follows the patient between caregivers, utilizing the fact that the Swedish health-care system is connected throughout the country, which means that the measurements can be carried out centrally by pulling information out of several EHR systems and pushing back risk assessments.

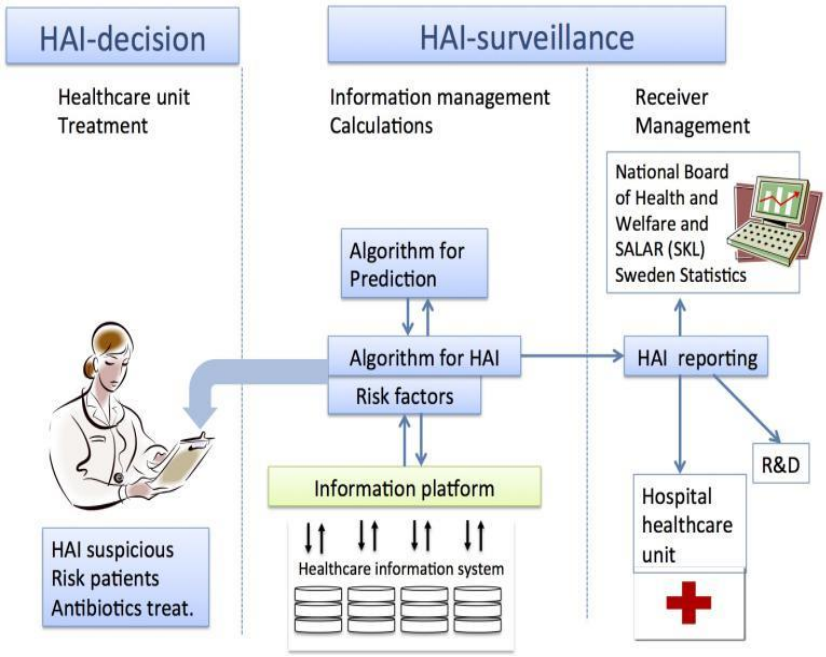

Figure 1

Fig. 1. A provisional system for monitoring and calculating Healthcare-Associated Infections (HAI) - HAI-Surveillance, but also for predicting patients with possible HAIs - HAI-Decision. Information is collected to produce statistics, but also to produce warnings and alerts to clinicians treating patients at healthcare units. The system could be used centrally in Sweden using the county councils' joint service platform and intranet.

Detection and Exploration of Adverse Drug Events

Adverse drug events constitute the most common form of iatrogenic injury, causing approximately 3.7% of hospital admissions worldwide, and one of the most common causes of death: in Sweden, they have been identified as the seventh most common cause of death.

The safety of drug is thus a major public health issue, necessitating their continuous monitoring, including post marketing due to the unavoidable limitations of clinical trials in terms of duration and sample size (number of patients). This activity, known as drug safety surveillance or

pharmacy co vigilance, primarily relies on collecting information voluntarily reported by clinicians or users of the target drugs. Such individual case reports, however, come with severe limitations, such underreporting and low reliability

Fig. 2. - An exploratory data analysis tool for investigating adverse drug events

Diagnosis Code Assignment

Assigning diagnosis codes that correspond to a given disease or health condition is necessary in order to estimate the prevalence and incidence of diseases and health conditions, as well as monitor differences therein over space and time. For such statistics to be, to some degree, comparable, a standard known as the International Statistical Classification of Diseases and Related Health Problems (ICD), created by the World Health Organization, is in use. The process of assigning diagnosis codes is generally carried out by either expert coders or physicians. In both cases, diagnosis code assignment is expensive and time-consuming, yet essential. According to one estimate, the cost of diagnosis coding and associated errors is approximately $25 billion per annum in the US.

The National Board of Health and Welfare also estimates that 20 percent of the assigned ICD-10 diagnosis codes are erroneous that efforts have long been made to provide computer-aided diagnostic coding. Using the EPR quantity, we have explored the repurposing of distributional semantics – i.e., models of word meaning that exploit word co-occurrence patterns in large corpora to obtain estimates of semantic similarity between words for the task of recommending diagnosis codes to assign to a care episode.

This approach leverages historical encoding of diagnoses and the words used in the clinical notes of the corresponding care episodes to create a predictive model that recommends possible diagnosis codes to assign to a new care episode on the basis of the data – primarily in the form of free-text – that is available for that care episode.

Text Mining in the Cancer Domain

Cancer is a disease that is treatable with a high success rate in its early stages, but with few early symptoms. In later stages, it is a serious illness. An infection with a human papiloma virus (HPV) is necessary for the development of cervical cancer, and as vaccines against HPV types 6, 11, 16 and 18 provides a high degree of protection against infection, vaccination programs are believed to reduce the cases of cervical cancer. Since screenings with pap smears, where women are investigated for pre-cancerous changes, have been implemented, the number of cervical cancer cases has nearly halved.

However, not all women take part in screening and other methods of finding early symptoms would therefore be valuable. Health records contain a patient's medical history; the free text part of the records can reveal what previous diseases and symptoms a patient has experienced. By applying text mining methods on records of cervical cancer patients early, possibly unknown symptoms can be found.

These symptoms could be of great value for detection of the disease. We have investigated symptoms de-scribed in the health records of patients with a cervical cancer diagnosis from the EPR quantity, by performing named entity recognition and negation detection.

Temporal Modeling of Clinical Events

Temporal information is a crucial aspect for developing accurate models of e.g. disease progression and treatment effects. For instance, knowing that a particular symptom occurred before or after a patient was treated with a specific medication alters the conclusions that can be drawn from how well a medication worked for a particular problem. Time information can be extracted from EHR data through document timestamps and other structured information, but is often also documented in free text.

An Envisioned Infrastructure for EHR Data Access

We have hitherto been successful in organizing and utilizing our EHR database for research, as described above; however, the database is currently far from being utilized to its full potential. To fulfill our vision of facilitating the development of useful data science applications in the healthcare domain, our goal is to provide access to this data, in a refined form, to both researchers and suppliers of healthcare-related IT tools.

To provide the data on a large scale in a sustainable manner, there is a need for an infrastructure, the details of which are described below.

The intension is that this infrastructure will provide a workbench for data science application development in the healthcare domain.

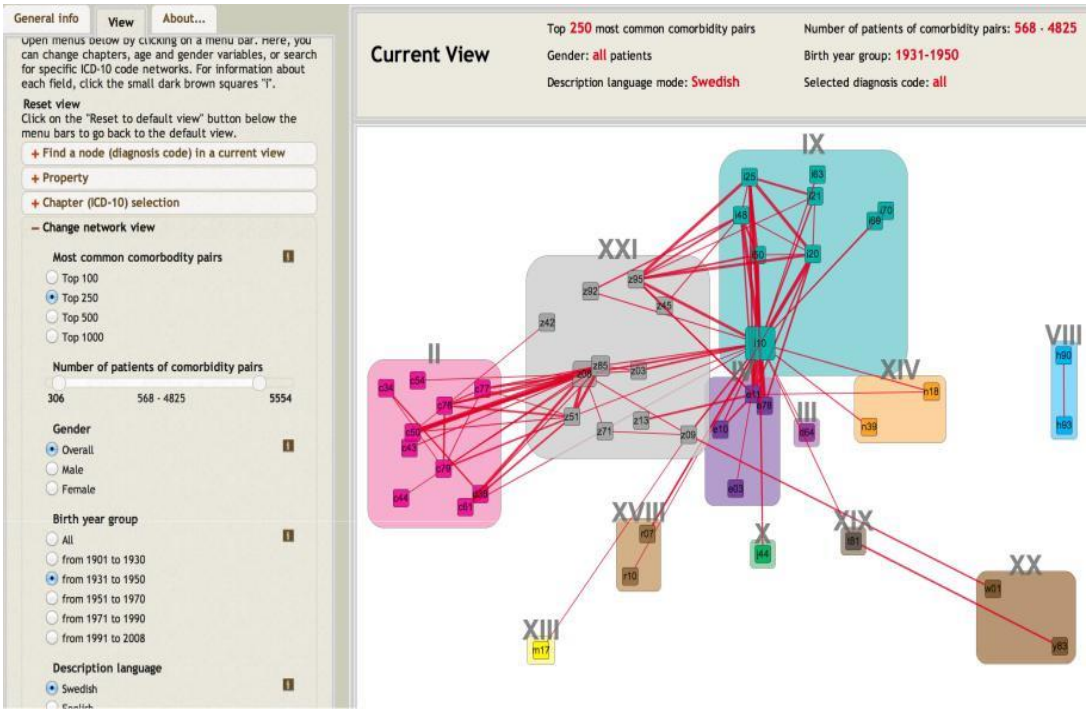

Fig. 3. Screenshot of the Co morbidity View demonstrator applied on 605,587 patient records

Technical Solutions

The communications requires a technical solution that conveniently provides access to the EHR data to the various intended users, while doing so in a secure fashion, which is critical given the inherently sensitive nature of the data. The infrastructure will be designed as a pipeline, allowing the user to select the data it wants and to obtain the data via e-access in a form that fits the user's needs (Figure 4).

An important prerequisite is thus that the entire database is appropriately preprocessed and indexed to ensure that the required information can be readily extracted.

There will essentially be two ways of accessing EHR data analyze the data from different views and/or download aggregated data at levels contiguous at least one hundred patients. This will allow us to make available users with secure access to de-identified data.

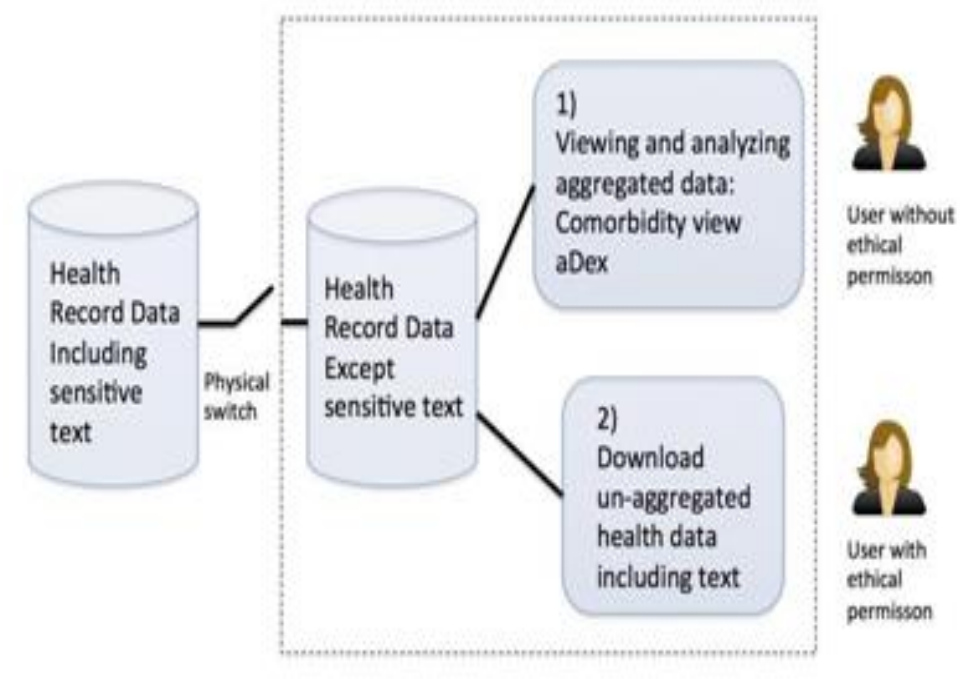

Fig. 4. The proposed technical solution that provides access to Database in a pipeline fashion.

Conclusions

We have here provided an overview of research conducted using a database of electronic health records – the EPR quantity – representative the potential of exploiting and reusing such data to create data science applications that are intended to support and, ultimately, improve healthcare. The ability to develop such applications, which are often data-intensive, hinges to a great extent on having access to data, which is currently challenging to obtain.

To recognize the full possible of data science applications in the healthcare domain, health record data needs to be made available to both researchers and industry actors, such as system developers. We have outlined a vision to create a communications, around the EPR quantity, efficiently providing access to EHR data in aggregated as well as non-aggregated form. However, making sensitive data available to the large number of potential users requires paying careful attention to various ethical issues and complying with in sequence security standards and

regulations: Database makes data available in a ready and secure fashion. Supporting users with practical, legal and ethical guidelines, to perform high quality research.

We believe that database, by providing a workbench for system development, will pro-mote a growing industry around the creation of data science applications in healthcare.

References

[1]. Henriksson, A., Skeppstedt, M., Kvist, M., Duneld, M., Conway, M.: Corpus driven terminology development: populating Swedish SNOMED CT with synonyms ex- tracted from electronic health records. In: Proceedings of BioNLP. pp. 36–44. As-sociation for Computational Linguistics (2013)

[2]. Hirsch, J.S., Tanenbaum, J.S., Gorman, S.L., Liu, C., Schmitz, E., Hashorva, D., Ervits, A., Vawdrey, D., Sturm, M., Elhadad, N.: HARVEST, a longitudinal patient record summarizer. Journal of American Medical Informatics Association 22 (2015)

[3]. Howard, R., Avery, A., Slavenburg, S., Royal, S., Pipe, G., Lucassen, P., Pirmo-hamed, M.: Which drugs cause preventable admissions to hospital? a systematic review. British Journal of Clinical Pharmacology 63(2), 136–147 (2007)

[4]. Humphreys, H., Smyth, E.T.M.: Prevalence surveys of healthcare-associated infections: what do they tell us, if anything? Clinical Microbiology and Infection 12(1), 2–4 (2006)

[5]. Isenius, N., Velupillai, S., Kvist, M.: Initial Results in the Development of SCAN a Swedish Clinical Abbreviation Normalizer. In: Proceedings of the CLEF 2012 Workshop on Cross-Language Evaluation of Methods, Applications, and Resources for eHealth Document Analysis - CLEFeHealth2012. CLEF, Rome, Italy (Septem-ber 2012)

[6]. Jensen, P.B., Jensen, L.J., Brunak, S.: Mining electronic health records: towards better research applications and clinical care. Nature Reviews Genetics 13(6), 395– 405 (2012)

[7]. Karlsson, I., Zhao, J., Asker, L., Boström, H.: Predicting adverse drug events by analyzing electronic patient records. In: Artificial Intelligence in Medicine Lecture Notes in Computer Science, pp. 125–129. Springer (2013)

[8]. Kvist, M., Tanushi, H., Sparrelid, E.: Automated detection of Healthcare-Associated Infections in Swedish Electronic Health Records. Manscript in prepa-ration (2015)

[9]. Kvist, M., Velupillai, S.: SCAN: A Swedish Clinical Abbreviation Normalizer. In: Information Access Evaluation. Multilinguality, Multimodality, and Interaction, pp. 62–73. Springer (2014)

[10]. Langseth, H., Luostarinen, T., Bray, F., Dillner, J.: Ensuring quality in studies linking cancer registries and biobanks. Acta Oncologica 49(3), 368–377 (2010)

[11]. Lewis, J.D., Schinnar, R., Bilker, W.B., Wang, X., Strom, B.L.: Validation stud-ies of the health improvement network (thin) database for pharmacoepidemiology research. Pharmacoepidemiology and drug safety 16(4), 393–401 (2007)

[12]. Lin, Y.K., Chen, H., Brown, R.A.: MedTime: A temporal information extraction system for clinical narratives. Journal of Biomedical Informatics 46, 20–28 (2013)

[13]. Lövestam, E., Velupillai, S., Kvist, M.: Abbreviations in Swedish Clinical Text - use by three professions. Studies in Health Technology and Informatics 205, 720–724 (August 2014)

[14]. Meystre, S.M., Savova, G.K., Kipper-Schuler, K.C., Hurdle, J.F.: Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform 35, 128–144.

[15]. Roque, F.S., Jensen, P.B., Schmock, H., Dalgaard, M., Andreatta, M., Hansen, T., Søeby, K., Bredkjær, S., Juul, A., Werge, T., et al.: Using electronic patient records to discover disease correlations and stratify patient cohorts. PLoS computational biology 7(8), e1002141 (2011)

[16]. Saeed, M., Villarroel, M., Reisner, A.T., Clifford, G., Lehman, L.W., Moody, G., Heldt, T., Kyaw, T.H., Moody, B., Mark, R.G.: Multiparameter intelligent moni-toring in intensive care ii (mimic-ii): A public-access intensive care unit database. Critical care medicine 39(5), 952 (2011)

[17]. SALAR: Swedish Association of Local Authorities and Regions: Vårdrelaterade infektioner framgångsfaktorer som förebygger. Stockholm, Sweden. ISBN: 978-91-7585-109-9, http://webbutik.skl.se/bilder/artiklar/pdf/978-91-7585-109-9.pdf, Accessed April 10 (2014)

[18]. Skeppstedt, M.: Negation detection in Swedish clinical text: An adaption of NegEx to Swedish. Journal of Biomedical Semantics 2(Suppl 3), S3 (2011)

[19]. Skeppstedt, M., Ahltorp, M., Henriksson, A.: Vocabulary expansion by semantic extraction of medical terms. In: The 5th International Symposium on Languages in Biology and Medicine (LBM). pp. 63–68 (2013)

[20]. Skeppstedt, M., Kvist, M., Nilsson, G., Dalianis, H.: Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study. In: Journal of Biomedical Informatics, 49. pp. 148–158

[21]. Smith, K., Megyesi, B., Velupillai, S., Kvist, M.: Professional language in Swedish clinical text: Linguistic characterization and comparative studies. Nordic Journal of Linguistics 2, 297–327 (2014)

[22]. Socialstyrelsen: The National Board of Health and Welfare, iagnosgranskningar utförda i Sverige 1997-2005 samt råd inför granskning, (In Swedish). http://www.socialstyrelsen.se/Lists/Artikelkatalog/Attachments/9740/2006-131-30_200613131.pdf (2006)

[23]. Spasić, I., Livsey, J., Keane, J.A., Nenadić, G.: Text mining of cancer-related information: Review of current status and future directions. I. J. Medical Informatics 83(9), 605–623 (2014), http://dx.doi.org/10.1016/j.ijmedinf.2014.06.009

[24]. Sun, W., Rumshisky, A., Uzuner, Ö.: Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. JAMIA 20(5), 806–813 (2013)

[25]. Tanushi, H., Dalianis, H., Nilsson, G.: Calculating prevalence of comorbidity and comorbidity combinations with diabetes in hospital care in sweden using a health care record database volume 744, ISSN: 1613-0073, 59–66 (2011)

[26]. Tengstrand, L., Megyesi, B., Henriksson, A Duneld, M., Kvist, M.: EACL – Ex- pansion of Abbreviations in Clinical text. In: Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR). pp. 94–103. Association for Computational Linguistics (2014)

[27]. Velupillai, S.: Temporal Expressions in Swedish Medical Text – A Pilot Study. In: Proceedings of BioNLP 2014. pp. 88–92. Association for Computational Linguis-tics, Baltimore, Maryland (June 2014)

[28]. Velupillai, S., Dalianis, H., Hassel, M., Nilsson, G.H.: Developing a standard de-identifying electronic patient records written in Swedish: precision, recall and F-measure in a manual and computerized annotation trial. International journal of medical informatics 78(12), e19–e26 (2009)

[29]. Wester, K., Jönsson, A.K., Spigset, O., Druid, H., Hägg, S.: Incidence of fatal adverse drug reactions: a population based study. British Journal of Clinical Pharmacology 65(4), 573–579 (2008)

[30]. Weegar, R., Kvist, M., Sundström, K., Brunak, S., Dalianis, H.: Finding Cervical Cancer Symptoms in Swedish Clinical Text using a Machine Learning Approach and NegEx (2015 submitted)

