HEALTH BANK – A Workbench for Data Science Applications in Healthcare

Hercules Dalianis[1], Aron Henriksson[1], Maria Kvist[1,2], Sumithra Velupillai[1], Rebecka Weegar[1]

[1] Department of Computer and Systems Sciences, (DSV)
Stockholm University, Sweden
[2] Department of Learning, Informatics, Management and Ethics (LIME)
Karolinska Institutet, Stockholm, Sweden
hercules@dsv.su.se, aronhen@dsv.su.se, maria.kvist@karolinska.se,
sumithra@dsv.su.se, rebeckaw@dsv.su.se

Abstract. The enormous 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 readily available to researchers and developers. A relatively rare exception is the large EHR database, the Stockholm EPR Corpus, comprising data from more than two million patients, that has been been made available to a limited group of researchers at Stockholm University. Here, we describe a number of data science applications that have been developed using this database, demonstrating the potential reuse of EHR data to support healthcare and public health activities, as well as facilitate medical research. However, in order to realize the full potential of this resource, it needs to be made available to a larger community of researchers, as well as to industry actors. To that end, we envision the provision of an infrastructure around this database called HEALTH BANK – the Swedish Health Record Research Bank. It will function both as a workbench for the development of data science applications and as a data exploration tool, allowing epidemiologists, pharmacologists and other medical researchers to generate and evaluate hypotheses. Aggregated data will be fed into a pipeline for open e-access, while non-aggregated data will be provided to researchers within an ethical permission framework. We believe that HEALTH BANK has the potential to promote a growing industry around the development of data science applications that will ultimately increase the efficiency and effectiveness of healthcare.

Keywords: electronic health record, data science, health intelligence, infrastructure, data mining, text mining, predictive modeling, clinical text, health bank, health record research

1 Introduction

Data produced in the healthcare setting is very valuable for further analysis and development of improved healthcare processes, such as real-time monitoring, decision support, and predictive analytics. Electronic health record (EHR) systems are used in almost all healthcare institutions in the Nordic countries, providing an invaluable opportunity for secondary data use and development of systems to aid clinicians in their daily work, hospital managements in their work on process and healthcare delivery improvements, and researchers in their work.

Resources for health and medical research are currently available through biobanks and national registers such as cancer registers and cause of death registers for researchers with appropriate ethical permission. However, in the Nordic countries, there are no easily available health record resources that describe health processes, diagnoses and treatments of a real clinical population [33, 40].

There has been an intense development of tools and techniques in the last twenty years to automatically process a variety of data sources because of the digitisation of the world, to enable further analysis and tool development. For instance, as is widely known, the Internet contains information in various formats, and a number of systems have been developed to make this information readily available for easy access, such as search engines and information extraction tools. The move to digitized solutions has also taken place in healthcare. The tools have, however, not been developed at the same pace. One important reason is that the health data has not been openly available for the research community and industry in order to construct such tools, primarily because health record data contains sensitive information about individuals – an aspect that is extremely important and that requires particular considerations.

To address these issues, we propose to develop an infrastructure that enables access to de-identified EHR data for further analysis and system development. This infrastructure will include a workbench with various preprocessing tools, and will consist of two pipelines: one providing access to structured, aggregated and completely de-identified data, and one requiring ethical permission before access to original data is provided.

This infrastructure will be based on a large clinical database, the Stockholm EPR (Electronic Patient Record) Corpus, which has been collected and refined during eight years [8, 7]. The Stockholm EPR Corpus contains over two million patients from all medical and surgical departments from the entire hospital (excluding only psychiatry and venereology), both inpatient and outpatient records written by several different professionals at Karolinska University Hospital. The records encompass the period 2006-2014. The corpus is de-identified with regard to names of patients and personal identity numbers. The personal 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 laboratory values, admission and discharge dates, timestamps – and unstructured data (free text), e.g. daily notes by clinicians and discharges summaries.

The infrastructure and workbench, called the Swedish Health Record Research Bank (HEALTH BANK), will be unique in that it provides access to authentic EHR data from the largest populated area in Sweden, from several clinical departments and clinical professions. It will also provide language technology tools for preprocessing and structuring the clinical narratives. Moreover, it provides complementary data to available biobanks and registries, enabling large-scale population studies for a variety of use-cases.

2 Electronic Health Record Resources for Research and 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 restricted 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 complex and not easy to extract data from. In particular, the parts of the EHRs that are written in free text, such as discharge summaries and daily notes, are often most difficult to obtain access to given their sensitive nature, but constitute a large part of the healthcare documentation.

Some large patient record databases or corpora (text collections) are available for research, including the

- i2b2[1] corpus contains of several clinical sub corpora in English that has been used in several shared challenges.
- CMC[2] corpus, containing 2,216 patient records in English
- MIMIC II database[3], which consists of 30 000 intensive care patient records written in English [44]
- A Finnish clinical corpus[4], containing 2,800 sentences from nursing notes and finally
- THIN database, containing 11 million English patient records from general practices [34]

Both academia and industry have developed methods within computer science, statistics, computational linguistics and machine learning. This is an evolving research area also called e-science, or (big) data science - to process abundant data and produce meaningful information [37, 29, 6].

3 Data Science Applications for Healthcare

It has been estimated that at least ten percent of all patients treated at hospitals in Europe suffer from an adverse event (AE), including adverse drug events

[1] https://www.i2b2.org/NLP/HeartDisease/PreviousChallenges.php
[2] http://computationalmedicine.org/catalog
[3] http://www.physionet.org/physiotools/deid
[4] http://bionlp.utu.fi/clinicalcorpus.html

(ADE), healthcare associated infections (HAI), fall injuries and bedsores – in total three million patients yearly [27]. Such AEs prolong the treatment of the patient, cause suffering for the patient, and is costly for society. In Sweden, with its ten million inhabitants, it is estimated that AEs are responsible for 750,000 extra healthcare days at the hospital, costing an additional of 700 million euros yearly, without taking into account the suffering of the patients [45]. Therefore, detecting AEs is a cricital issue in healthcare.

The Stockholm EPR Corpus at Stockholm University has been used for several research projects that are of practical importance for healthcare. These projects have included work on HAI detection, detection of ADEs in a post-marketing setting, text simplification of the EHRs for laypeople, 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 difficult 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 Swedish. These include factuality level classification [58, 61], negation detection [46], spelling error detection [10], abbreviation normalization, [28, 32, 57], named entity recognition [48, 17], as well as tools for expanding medical vocabularies [16, 24, 47, 23].

We have also initiated studies on characterizing the domain-specific language in this type of text [49], and performed studies on how well general language tools and techniques work on clinical notes, such as syntactic parsers [14] and distributional semantic models [15] – 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 de-identification (of protected health information), factuality levels of diagnostic expressions, clinical named entities, indications and ADE relations, cervical cancer symptoms, classifications of HAI (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.

3.1 Automatic surveillance of healthcare-associated infections

A healthcare-associated infection (HAI) is an infection obtained by a patient during healthcare treatment. There is a requirement to report annually the number of HAIs in each hospital, which is currently carried out in one of two ways: by compulsory reporting of HAI cases, but also through so called Point Prevelance Measurements (PPMs), which are carried out twice a year at all hospitals in Sweden. PPMs are conducted manually by assessing all the patients admitted on one particular day and deciding whether those patients have suffered from a HAI or not. The estimates obtained through PPMs 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 insitutions

a better instrument for surveillance, as well as facilitate the evaluation of actions performed to reduce the number of HAIs.

We have developed several prototype tools for detecting HAIs in EHRs. One machine learning based tool, Detect-HAI, analyzes the clinical notes in a patient's health records automatically deterimines if the patient has potentially suffered a HAI or not. 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 [11]. In another approach, rule- or knowledge-based systems are developed for for specific HAI diagnoses, initially focusing on urinary tract infections [56] and bacteriemia [31].

In Figure 1 a tentative system for HAI surveillance is depicted. The system 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.

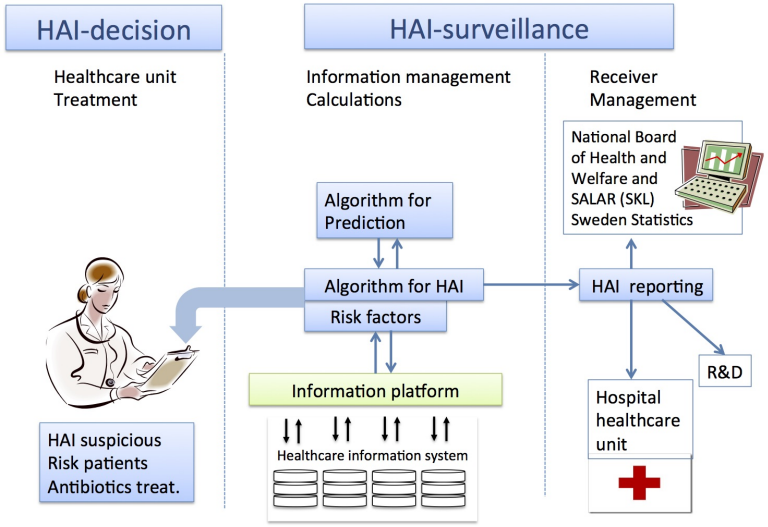

Fig. 1. A tentative system for monitoring and calculating Healthcare-Associated Infections (HAI) - HAI-Surveillance, but also for predicting patients with possible HAIs - HAI-Decision. Information is collected to produce statistics, but also to produce warnings and alerts to clinicians treating patients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

3.2 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 [26], and one of the most common causes of death: in Sweden, they have been identified as the seventh most common cause of death [64]. 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 pharmacovigilance, 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 [12]. In recent years, alternative sources for pharmacovigilance have emerged, including EHRs, which have the distinct advantage of containing longitudinal observations of the treatment of patients, including their drug use. To address the underreporting of ADEs and thereby support pharmacovigilance, predictive modeling can be leveraged to create systems that can detect ADEs on the basis of patient-specific EHR data [30, 67, 66].

EHR data can also be used for data exploration and testing hypotheses with respect to, for instance, ADEs [22]. aDEX is an example of an exploratory data analysis tool for investigating ADEs, currently using health records over a two-year period (2009-2010). With the tool, one can create case and control groups to compare, e.g., patients who have experienced a specific ADE to patients who have not. Using disproportionality analysis methods, which calculate how much an event deviates from what is expected, one can identify drugs that seem to have the largest risk of causing the ADE. Figure 2 displays a screenshot of aDEX.

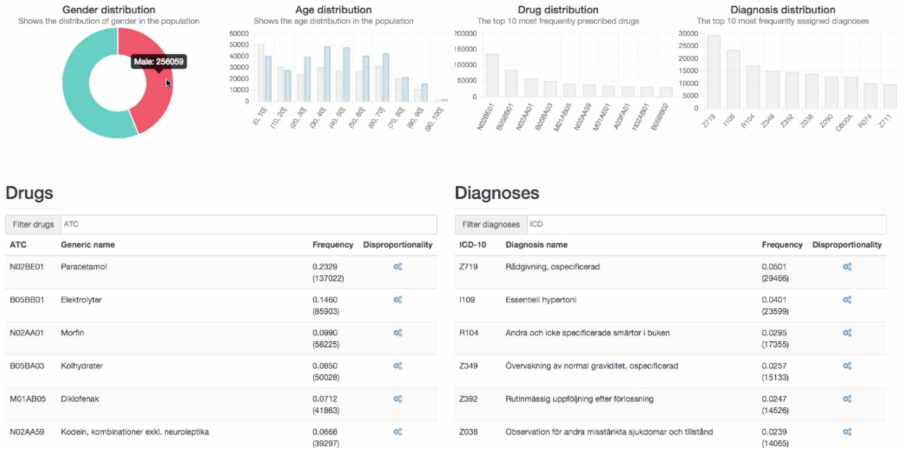

Fig. 2. Screenshot of aDEX - an exploratory data analysis tool for investigating adverse drug events, (for more information see http://people.dsv.su.se/~isak-kar/adex/).

3.3 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),[65], 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 [41]. The Swedish National Board of Health and Welfare also estimates that 20 percent of the assigned ICD-10 diagnosis codes are erroneous [50].

It is not surprising, then, that efforts have long been made to provide computer-aided diagnostic coding [52, 41]. Using the Stockholm EPR Corpus, 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 [5] – for the task of recommending diagnosis codes to assign to a care episode [21, 18–20]. 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.

3.4 Text mining in the cancer domain

Cervic al 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, causing around 180 deaths in Sweden yearly [3].

An infection with a human papiloma virus (HPV) is necessary for the development of cervical cancer [62], and as vaccines against HPV types 6, 11, 16 and 18 provides a high degree of protection against infection, vaccination programs are belived to reduce the cases of cervical cancer [38]. Since screening with pap smears, where women are investigated for pre-cancerous changes, have been implemented, the number of cervical cancer cases has nearly halved in Sweden [3]. 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 described in the health records of patients with a cervical cancer diagnosis from the Stockholm EPR Corpus, by performing named entity recognition and negation detection [63].

Another area in the cancer domain where text mining can be of value is the transferral of free text information in pathology reports into structured databases. Pathology reports describe tissue samples and can contain both macroscopic and microscopic observations and a possible diagnosis for a patient with known or suspected cancer [39]. Several studies have been performed on text mining of pathology reports, where the aim has been to transfer the free text data into structured format [51]. Manual transferal of pathology reports can be expensive and time consuming, for example, at Kreftregistret in Oslo (Cancer Registry of Norway), 20-25 human coders are working with manually transferring the pathology reports produced in Norway to a database. Text mining techniques can be used to automatize the transferral, completely or partly, to the data database.

3.5 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 occured 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. To be able to extract time information from narratives, usually three steps are required:

1. extracting temporal expressions denoting specific points in time (*today, two years ago, a while back, at 6 AM*)
2. extracting the clinically relevant events (*infection, antibiotics, surgery*)
3. ordering these in time (*infection* **before** *surgery*).

This is a challenging natural language processing task that has been subject of several research studies on English clinical text [53, 4, 35, 54, 42]. Work on creating systems for temporal information extraction for Swedish clinical text is ongoing [59]. After successful temporal modeling of information in clinical notes, patient trajectories and visualized timelines can be created, to be further used in applications such as summarization tools [25] or for enriched predictive analysis.

3.6 Text simplication of clinical narratives

An area of increasing importance is also patient engagement and involvement. In the future, patients themselves will most likely take a more active role in their own healthcare process. This is already the case in some areas, through, for instance, systems for self-monitoring of measurement values and self-treatment guided by remote healthcare contact. There is also political incentives and legislation in Sweden that describe how healthcare is to be transparent and understandable for patients. One aspect with healthcare documentation is that it

is very specialized and complicated – an aspect that is necessary for the communication among healthcare professionals in order to ensure preciseness and detail. However, this means that the documentation is difficult to understand for a layperson – e.g., a patient wanting to read her own medical records. One way of bridging this gap would be to provide patients with a simplified version of the medical records, where technical jargon and domain-specific vocabulary is translated, or converted, to language that does not require medical expert knowledge. We have performed several studies in this area, in particular in the radiology domain. In collaboration with the Center for Easy-to-Read (Centrum för Lättläst, in Sweden), we have analyzed which aspects of clinical documentation are central to target for the creation of simplified "translations", we have also studied and identified linguistic features that are characteristic for this type of documentation [49]. Moreover, we have developed a pilot tool for handling medical abbreviations [28, 32, 36], initiated work on lexical simplification [13], and conducted interview studies with patients to identify which aspects of clinical documentation are difficult to understand from their perspective [1].

3.7 Comorbidity analysis

Comorbidity is the presence of one or more additional disorders (or diseases) co-occurring with a primary disease or disorder. In the current prototype Comorbidity view[5], researchers can inspect what comorbidities, based on assigned ICD-10 diagnosis codes from the Stockholm EPR corpus, a group of patients have. The case group can be selected based on, for instance, gender and age.

Figure 3 shows a screenshot of the current Comorbidity View demonstrator, displaying a group of patients from a subset of the database (2006-2008) who have at least two ICD-10 diagnosis codes. An early version of the demonstrator is described in Tanushi et al. [55].

4 HEALTH BANK – 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 – the Swedish Health Record Research Bank (HEALTH BANK) – will provide a workbench for data science application development in the healthcare domain. We believe that HEALTH BANK will attract researchers and IT entrepreneurs from around the world to

[5] http://www2.dsv.su.se/comorbidityview-demo/

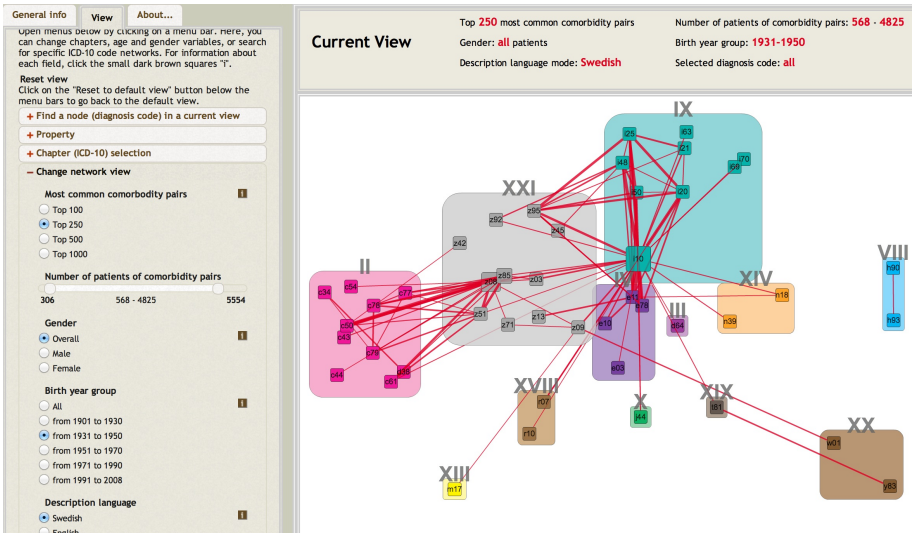

Fig. 3. Screenshot of the Comorbidity View demonstrator applied on 605,587 patient records encompassing the years 2006-2008. The rectangles represent the number of patients that have the same group of ICD-10 codes. The lines, together with the thickness of the lines, connect the number of patients that have the same pairs of diagnosis codes.

promote the growth of the industry around data-intensive IT solutions in healthcare. Making this valuable resoure readily available will moreover give Sweden a competetitive advantage, while hopefully leading to more countries following suit in taking similar iniatiatives in the endeavor of improving healthcare.

4.1 Technical solutions

The HEALTH BANK infrastructure 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 (see 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 from HEALTH BANK:

1. Through standard web-based access, where users without ethical permission can analyze the data from different views and/or download aggregated data at levels encompassing at least one hundred patients. This will allow us to provide users with secure access to de-identified data. For these purposes,

we plan to make the previously described aDEX and Comorbidity View[6] available.
2. Through an encrypted e-connection, where users with ethical permission can download non-aggregated data, including sensitive text[7].

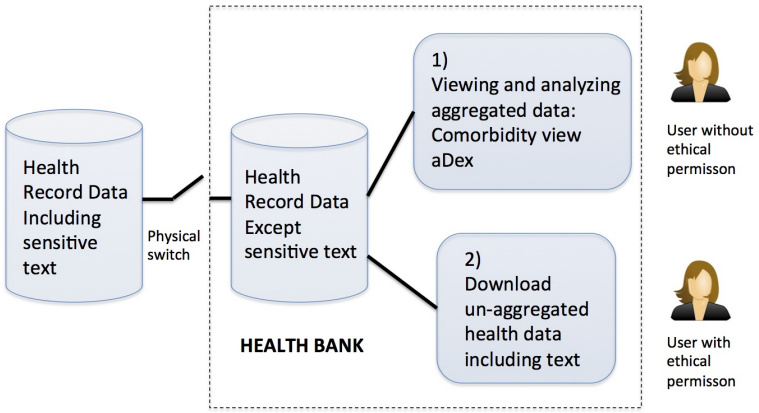

Fig. 4. The proposed technical solution that provides access to HEALTH BANK in a pipeline fashion. The physical switch shows that there is no Internet connection to the sensitive EHR text except for when it is downloaded by a user with ethical permission.

As there is a great demand to link different health registers, such as biobanks and cancer registers, with healthcare data[8], we envision linking these data sources to create added value. We also plan to add primary care data to the already acquired hospital care data. This will allow researchers to follow patients throughout the, sometimes elaborate, healthcare process.

4.2 Ethical considerations

We are aware of the profound ethical challenges involved in having access to a large repository that contains information that, if it ends up in the wrong hands,

[6] We plan to extend the Comorbidity view demonstrator to encompass the entire database and to include more functionality, e.g., by adding diagnosis expressions mined from clinical notes (similar to Roques et al. [43]).

[7] Although the data has already been de-identified in the sense that social security numbers and names in structured fields have been removed/replaced, the clinical text may contain names of, e.g., relatives to the patient or phone numbers. Regarding the sensitive nature of clinical text in the Stockholm EPR Corpus, several studies have been carried out [60, 9]

[8] http://www.nordforsk.org/en/news/report-on-nordic-registers-and-biobanks-launched

can cause suffering for the individual patients. For this reason, it is vital that we continue, as we have been, to communicate with the ethical review board (Regionala etikprövningsnämnden i Stockholm) regarding our research initiatives. Approval from the ethical review board needs to be obtained before carrying out new research or giving access to the Swedish Health Record Research Bank.

We have hitherto obtained seven ethical permissions from the regional ethical board for five different research projects that have been carried out both internally, together with other Swedish universities, and externally, in a research network (HEXAnord) and a center of excellence (NIASC), both in a Nordic context. One of the ethical permissions has an amendment, allowing us to share one hundred de-identified and pseudonymised health records, in the framework of a shared task, with other researchers affiliated to an academic institution. These hundred records are described in Alfalahi et al. [2]. We are moreover in continuous contact with the chief medical officer of Karolinska University Hospital on these matters.

When providing access to sensitive data to a larger group of people, as HEALTH BANK is intended to do, it is important to have guidelines that describe how to conduct research with EHR data. These guidelines should describe various technical details, known problems and solutions, and, perhaps most importantly, how to write applications to the ethical review board: what the contents of such applications should be and a description of the required steps for applying for ethical permission.

HEALTH BANK will moreover comply with applicable legal requirements and generally accepted standards. For information security and protection of patient data, such as:

- Patientdatalag (2008:355), in applicable parts
- Personuppgiftslag (1998:204)

For information security standards, such as:

- ISO/IEC 27001, requirements for information security management system
- ISO/IEC 27002, information security standard

The security of the infrastructure's technical components will be designed in accordance with internal and external security requirements with respect to the risks involved. The security of the infrastructure will moreover be audited on a regular basis. In addition to addressing information security concerns, there will be a reference group that will discuss any issues that may arise in relation to how data is made available through HEALTH BANK. This reference group will consist of medical experts, researchers, system developers, suppliers of health management systems and patient organizations.

4.3 Potential users

We believe that interest in HEALTH BANK would be substantial and the number of potential users large. In the eight years (2007-2015) that we have had access to EHR data, albeit in a significantly more limited setting than is intended

for HEALTH BANK, we have collaborated with numerous academic institutions, hospitals and health organizations, pharmaceutical companies, healthcare management system developers and patient organizations:

- Academic institutions: Stockholm University, Karolinska Institutet, Karolinska University Hospital, Uppsala University, Gothenburg University, University of Borås, University of Turku, University of Copenhagen, DTU-Danmarks Tekniske Universitet, NTNU-Trondheim, Vytautas Magnus University, Lithuania, UC San Diego and University of Utah, USA.
- We have also collaborated with several hospitals and organizations: National Board of Health, (Socialstyrelsen), The Swedish Association of Local Authorities and Regions (Sveriges Kommuner och Landsting), Stockholm County Council (Stockholms Läns Landsting), Östergötland County Council (Landstinget i Östergötland), Uppsala Monitoring Center (UMC).
- Moreover with several several companies: Astra Zeneca (pharmaceutical company), Capish Knowledge (database and software company), Pygargus (clinical trials company), TakeCare Compugroup Medical (Electronic patient records system company).
- Patient organizations as The Swedish Heart and Lung Association (Hjärt- och Lungsjukas Riksförbund) och the Swedish Rheumatism association (Svenska Reumatikerförbundet) and Swedish Patient Insurance (Landstingens Ömsesidiga Försäkringsbolag).

All of the above organizations are possible users of HEALTH BANK. In addition to these, the following potential users have been identified: SciLifeLab (national center of Science for Life Laboratory), partners of SciLifeLab, BBMRI.se (The Biobanking and Molecular Resource Infrastructure of Sweden), and the Swedish node of the bioinformatics infrastructure ELIXIR, and partners of NIASC (The Nordic Center of Excellence in Health-Related e-Sciences), which aims to connect health records to biobanks and registries. Moreover, the Vinnova funded project IntergrIT needs to develop tools to perform research on EHR data. We also believe that HEALTH BANK has the potential to encourage entrepreneurs to start companies that focus on developing data science applications in the healthcare domain.

5 Conclusions

We have here provided an overview of research conducted using a database of electronic health records – the Stockholm EPR Corpus – demonstrating 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 realize the full potential 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. To that end, we have outlined a vision to

create an infrastructure, HEALTH BANK, around the Stockholm EPR Corpus, effectively 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 information security standards and regulations: HEALTH BANK will make data available in a ready and secure fashion. Supporting users with practical, legal and ethical guidelines, to perform high quality research. We believe that HEALTH BANK, by providing a workbench for system development, will promote a growing industry around the creation of data science applications in healthcare.

Acknowledgements

We thank Karolinska University Hospital for their confidence in giving us access to the Stockholm EPR Corpus to carry out important research. We would also like to thank Vinnova - Swedish Agency for Innovation Systems for initial funding, SSF - Swedish Foundation for Strategic Research, through the project High-Performance Data Mining for Drug Effect Detection under grant IIS11-0053, the Swedish Research Council (project 350-2012-6658), Vårdalstiftelsens Idéprovning, as well as NIASC-Nordic Center of Excellence in Health-Related e-Sciences for partial funding of the research.

References

1. Aanta, K., Wide, C., Kvist, M., Salanterä, S.: Patients interpreting the medical language of discharge summaries. Manscript in preparation (2015)
2. Alfalahi, A., Brissman, S., Dalianis, H.: Pseudonymisation of personal names and other PHIs in an annotated clinical Swedish corpus. In: Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2012) held in conjunction with LREC 2012, May 26, Istanbul. pp. 49–54 (2012)
3. Axelsson, A., Borgfeldt, C.: Cervixcancer (2013), http://www.internetmedicin.se/page.aspx?id=2735
4. Bethard, S., Derczynski, L., Pustejovsky, J., Verhagen, M.: Semeval-2015 task 6: Clinical tempeval. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). Association for Computational Linguistics (2015)
5. Cohen, T., Widdows, D.: Empirical distributional semantics: methods and biomedical applications. Journal of biomedical informatics 42(2), 390–405 (2009)
6. Dalianis, H.: Clinical text retrieval-an overview of basic building blocks and applications. In: Professional Search in the Modern World, pp. 147–165. Springer (2014)
7. Dalianis, H., Hassel, M., Henriksson, A., Skeppstedt, M.: Stockholm EPR Corpus: A clinical database used to improve health care. In: Swedish Language Technology Conference. pp. 17–18 (2012)
8. Dalianis, H., Hassel, M., Velupillai, S.: The Stockholm EPR Corpus-Characteristics and Some Initial Findings. In: Proceedings of ISHIMR 2009, Evaluation and implementation of e-health and health information initiatives: international perspectives. 14th International Symposium for Health Information Management Research. pp. 243–249 (2009)

9. Dalianis, H., Velupillai, S.: De-identifying Swedish clinical text-refinement of a gold standard and experiments with Conditional random fields. J. Biomedical Semantics 1, 6 (2010)
10. Dziadek, J.: Improving snomed mapping of clinical texts using context-sensitive spelling correction. Master thesis (2015)
11. Ehrentraut, C., Kvist, M., Sparrelid, E., Dalianis, H.: Detecting healthcare-associated infections in electronic health records: Evaluation of machine learning and preprocessing techniques. In: Sixth International Symposium on Semantic Mining in Biomedicine (SMBM 2014). University of Aveiro (2014)
12. Goldman, S.A.: Limitations and strengths of spontaneous reports data. Clinical Therapeutics 20, C40–C44 (1998)
13. Grigonyte, G., Kvist, M., Velupillai, S., Wirèn, M.: Improving Readability of Swedish Electronic Health Records through Lexical Simplification: First Results, booktitle = Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations – PITR. pp. 74–83. Association for Computational Linguistics, Gothenburg, Sweden (April 2014), http://www.aclweb.org/anthology/W14-1209
14. Hassel, M., Henriksson, A., Velupillai, S.: Something Old, Something New – Applying a Pre-trained Parsing Model to Clinical Swedish. In: Proc. 18th Nordic Conf. on Comp. Ling. - NODALIDA '11 (May 11-13 2011), http://dspace.utlib.ee/dspace/handle/10062/17355
15. Henriksson, A.: Semantic Spaces of Clinical Text: Leveraging Distributional Semantics for Natural Language Processing of Electronic Health Records. Licentiate Thesis, Stockholm University (2013)
16. Henriksson, A., Conway, M., Duneld, M., Chapman, W.W.: Identifying synonymy between SNOMED clinical terms of varying length using distributional analysis of electronic health records. In: AMIA Annual Symposium Proceedings. pp. 600–609. American Medical Informatics Association (2013)
17. Henriksson, A., Dalianis, H., Kowalski, S.: Generating features for named entity recognition by learning prototypes in semantic space: The case of de-identifying health records. In: International Conference on Bioinformatics and Biomedicine (BIBM). pp. 450–457. IEEE (2014)
18. Henriksson, A., Hassel, M.: Election of diagnosis codes: Words as responsible citizens. In: Proceedings of Louhi Workshop on Health Document Text Mining and Information Analysis. pp. 67–74. CEUR-WS (2011)
19. Henriksson, A., Hassel, M.: Exploiting Structured Data, Negation Detection and SNOMED CT Terms in a Random Indexing Approach to Clinical Coding. In: Proceedings of RANLP Workshop on Biomedical Natural Language Processing. pp. 3–10. Association for Computational Linguistics (2011)
20. Henriksson, A., Hassel, M.: Optimizing the dimensionality of clinical term spaces for improved diagnosis coding support. In: Proceedings of Louhi Workshop on Health Document Text Mining and Information Analysis (2013)
21. Henriksson, A., Hassel, M., Kvist, M.: Diagnosis code assignment support using random indexing of patient records – a qualitative feasibility study. In: Proccedings of Artificial Intelligence in Medicine, pp. 348–352. Springer (2011)
22. Henriksson, A., Kvist, M., Hassel, M., Dalianis, H.: Exploration of adverse drug reactions in semantic vector space models of clinical text. In: Proceedings of ICML Workshop on Machine Learning for Clinical Data Analysis (2012)
23. Henriksson, A., Moen, H., Skeppstedt, M., Daudaravicius, V., Duneld, M.: Synonym extraction and abbreviation expansion with ensembles of semantic spaces. J. Biomedical Semantics 5(6) (2014)

24. Henriksson, A., Skeppstedt, M., Kvist, M., Duneld, M., Conway, M.: Corpus-driven terminology development: populating Swedish SNOMED CT with synonyms extracted from electronic health records. In: Proceedings of BioNLP. pp. 36–44. Association for Computational Linguistics (2013)
25. 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)
26. Howard, R., Avery, A., Slavenburg, S., Royal, S., Pipe, G., Lucassen, P., Pirmohamed, M.: Which drugs cause preventable admissions to hospital? a systematic review. British Journal of Clinical Pharmacology 63(2), 136–147 (2007)
27. 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)
28. 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 (September 2012)
29. 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)
30. 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)
31. Kvist, M., Tanushi, H., Sparrelid, E.: Automated detection of Healthcare-Associated Infections in Swedish Electronic Health Records. Manscript in preparation (2015)
32. Kvist, M., Velupillai, S.: SCAN: A Swedish Clinical Abbreviation Normalizer. In: Information Access Evaluation. Multilinguality, Multimodality, and Interaction, pp. 62–73. Springer (2014)
33. Langseth, H., Luostarinen, T., Bray, F., Dillner, J.: Ensuring quality in studies linking cancer registries and biobanks. Acta Oncologica 49(3), 368–377 (2010)
34. Lewis, J.D., Schinnar, R., Bilker, W.B., Wang, X., Strom, B.L.: Validation studies of the health improvement network (thin) database for pharmacoepidemiology research. Pharmacoepidemiology and drug safety 16(4), 393–401 (2007)
35. 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)
36. 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)
37. 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 (2008)
38. Muñoz, N., Kjaer, S.K., Sigurdsson, K.n., Iversen, O.E., Hernandez-Avila, M., Wheeler, C.M., Perez, G., Brown, D.R., Koutsky, L.A., Tay, E.H., Garcia, P.a.J., Ault, K.A., Garland, S.M., Leodolter, S., Olsson, S.E., Tang, G.W.K., Ferris, D.G., Paavonen, J., Steben, M., Bosch, F.X., Dillner, J., Huh, W.K., Joura, E.A., Kurman, R.J., Majewski, S., Myers, E.R., Villa, L.L., Taddeo, F.J., Roberts, C., Tadesse, A., Bryan, J.T., Lupinacci, L.C., Giacoletti, K.E.D., Sings, H.L., James, M.K., Hesley, T.M., Barr, E., Haupt, R.M.: Impact of Human Papillomavirus

(HPV)-6/11/16/18 Vaccine on All HPV-Associated Genital Diseases in Young Women. 102, 325–339 (2010), http://dx.doi.org/10.1093/jnci/djp534
39. National Cancer Institute: Pathology reports (2010), http://www.cancer.gov/cancertopics/diagnosis-staging/diagnosis/pathology-reports-fact-sheet
40. Nordforsk: Joint Nordic Registers and Biobanks - A goldmine for health and welfare research. Nordforsk policy paper 5 (2014), http://www.nordforsk.org/en/news/report-on-nordic-registers-and-biobanks-launched
41. Pestian, J.P., Brew, C., Matykiewicz, P., Hovermale, D., Johnson, N., Cohen, K.B., Duch, W.: A shared task involving multi-label classification of clinical free text. In: Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing. pp. 97–104. Association for Computational Linguistics (2007)
42. Reeves, R.M., Ong, F.R., Matheny, M.E., Denny, J.C., Aronsky, D., Gobbel, G.T., Montella, D., Speroff, T., Brown, S.H.: Detecting temporal expressions in medical narratives. International Journal of Medical Informatics 82, 118–127 (2013)
43. 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)
44. 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 monitoring in intensive care ii (mimic-ii): A public-access intensive care unit database. Critical care medicine 39(5), 952 (2011)
45. 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)
46. Skeppstedt, M.: Negation detection in Swedish clinical text: An adaption of NegEx to Swedish. Journal of Biomedical Semantics 2(Suppl 3), S3 (2011)
47. 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)
48. 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
49. 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)
50. Socialstyrelsen: The National Board of Health and Welfare, Diagnosgranskningar 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)
51. 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
52. Stanfill, M.H., Williams, M., Fenton, S.H., Jenders, R.A., Hersh, W.R.: A systematic literature review of automated clinical coding and classification systems. Journal of the American Medical Informatics Association 17(6), 646–651 (2010)

53. Styler, W.I., Bethard, S., Finan, S., Palmer, M., Pradhan, S., de Groen, P., Erickson, B., Miller, T., Lin, C., Savova, G., Pustejovsky, J.: Temporal Annotation in the Clinical Domain. Transactions of the Association for Computational Linguistics 2, 143–154 (2014), https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/305
54. Sun, W., Rumshisky, A., Uzuner, Ö.: Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. JAMIA 20(5), 806–813 (2013)
55. 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)
56. Tanushi, H., Kvist, M., Sparrelid, E.: Detection of healthcare-associated urinary tract infection in Swedish electronic health records. Studies in health technology and informatics 207, 330–339 (2013)
57. Tengstrand, L., Megyesi, B., Henriksson, A., Duneld, M., Kvist, M.: EACL – Expansion 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)
58. Velupillai, S.: Shades of Certainty: Annotation and Classification of Swedish Medical Records. Ph.D. thesis, Stockholm University (2012)
59. Velupillai, S.: Temporal Expressions in Swedish Medical Text – A Pilot Study. In: Proceedings of BioNLP 2014. pp. 88–92. Association for Computational Linguistics, Baltimore, Maryland (June 2014), http://www.aclweb.org/anthology/W14-3413
60. Velupillai, S., Dalianis, H., Hassel, M., Nilsson, G.H.: Developing a standard for 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)
61. Velupillai, S., Skeppstedt, M., Kvist, M., Mowery, D., Chapman, B.E., Dalianis, H., Chapman, W.W.: Cue-based assertion classification for swedish clinical text—developing a lexicon for pycontextswe. Artificial intelligence in medicine 61(3), 137–144 (2014)
62. Walboomers, J.M.M., Jacobs, M.V., Manos, M.M., Bosch, F.X., Kummer, J.A., Shah, K.V., Snijders, P.J.F., Peto, J., Meijer, C.J.L.M., Muñoz, N.: Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. The Journal of Pathology 189(1), 12–19 (1999)
63. 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)
64. 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)
65. WHO: International Classification of Diseases (ICD)), http://www.who.int/classifications/icd/en/, accessed 2014-04-09
66. Zhao, J., Henriksson, A., Asker, L., Boström, H.: Detecting adverse drug events with multiple representations of clinical measurements. In: IEEE International Conference on Bioinformatics and Biomedicine. pp. 536–543 (2014)
67. Zhao, J., Henriksson, A., Boström, H.: Detecting adverse drug events using concept hierarchies of clinical codes. In: IEEE International Conference on Healthcare Informatics (ICHI). pp. 285–293 (2014)