2019

Mapping of adverse events - identified by a high throughput data mining technology in a diabetes population - to an international consensus terminology

10^{th} Semester Master Project 2019
Fatima Palani

Translational Medicine, Department of Health and Science & Technology, Aalborg University

Project conducted at:
Novo Nordisk Foundation Center for Protein Research
Blegdamsvej 3B, 2200 Copenhagen

AALBORG UNIVERSITET

The Novo Nordisk Foundation
Center for Protein Research

AALBORG UNIVERSITET

Project title	Mapping of adverse events - identified by a high throughput data mining technology in a diabetes population - to an international consensus terminology
Group	18gr9032
Name	Fatima Palani
Study number	20144360
Education	Translational Medicine
Semester	10th
Pages	38
Pages of appendix	22
Extern supervisor	Søren Brunak
Intern supervisor	Meg Duroux

Abstract

Aim: The purpose of this study is to create a quality mapping to MedDRA, by applying text-mining to detect possible adverse events (AE) and thereby be able to map the terms into an international medical dictionary.

Introduction: Polypharmacy is the most common treatment strategy in diabetic patients, which is associated with various complications; however, the most important is drug-drug interactions (DDI). Pharmacovigilance (PV) is significant in drug development after an approved drug. The Medical Dictionary of Regulatory Activity (MedDRA) was developed to share findings and retrieve the latest information on drugs. A five level hierarchy, very general to very specific, structures it. Every term is identified by an 8-digit number, which describes the term of interest and remains the same number regarding of language. This kind of communications allows researcher to share and understand new findings across different languages, leading to better health globally.

Methods: An automatic computational text mining technique was conducted in 12,073 unstructured electronic patient records (EPR). Natural language processing (NLP) algorithm was executed to structure, understand and validate relevant information. Detected unique terms were mapped into MedDRA in combination with the in-house dictionary, established by International Classification of disease (ICH) to achieve an internationally comparable dataset.

Results: Application of text mining in 12,073 EPRs with NLP gave us 3,830 unique terms, which may be possible AEs. All unique terms were feasible to be mapped into MedDRA. Simple words were easier to be mapped compared to terms that were more complex. Moreover, extraction of the phenotypic profiles, as well as the most common prescribed drugs with associated AEs were able to be extracted.

Conclusion: Based on our results, we conclude that the use of MedDRA as our dictionary in combination with the in-house NER-tagger were possible to use and can give a sophisticated analysis on a Scandinavian as well as international level. Moreover, application of text mining detect possible AE, and is a great tool when dealing with large databases, which is the case in most safety trails.

Resume

Formål: Formålet med dette studie er at foretage et kvalitets mapping ind til MedDRA, ved at, udfører tekst mining med formålet om opdage mulige bivirkninger, og dermed være i stand til at mappe termerne ind til en medicinsk terminologi.

Introduktion: Polyfarmaci er den mest almindelige behandlingsstrategi hos diabetespatienter, som er forbundet med mange følgekomplikationer; Dog er den vigtigste interaktioner mellem to forskellige lægemidler. Pharmako-overvågelse vigtig efter et godkendt lægemiddel. Medical Dictionary for Regulatory Activity (MedDRA) blev oprettet for at dele og hente nye fund om et lægemiddel på kryds af forskellige lande. Den er opbygget af et 5 hierarkisk niveauer fra meget generelt til meget specifikt. Hvert term er karakteriseret med et 8-cifret nummer, og forbliver det samme tal på alle sprog. Denne form for kommunikation tillader videnskaber at forstå og dele nyt viden på tværs af forskellige sprog, hvilket leder til bedre sundhed globalt.

Metode: En automatisk databehandlingsteknologi blev udført i 12.073 ustrukturerede elektroniske patientjournaler. Naturlig sprogbehandling (NLP) algoritmen blev udført for at strukturere, forstå og validere relevante information fra teksterne. Unikke tegn der blev opdaget blev oversat i MedDRA, der blev oprettet af International Classification of disease (ICD) i kombination med in-house ordbog for at være i stand til at sammenligne databasen på internationalt niveau.

Resultater: Anvendelse af tekst mining i 12.073 EPR med NLP gav os 3.830 unikke tegn, hvilket kan være mulige bivirkninger. Alle termer var i stand til at blive oversat til MedDRA. Simple ord var nemmere at mappe sammenlignet med komplekse ord. Herudover var vi i stand til at trække fænotypes profiler samt de mest hyppige recept skrevne medikamenter samt relaterede bivirkninger.

Konklusion: Baseret på vores resultater kan vi konkludere, at anvendelse af MedDRA som vores medicinske terminologi i kombination med vores in-house ordbog var muligt at udføre og kan give en sofistikeret analyse på et skandinavisk samt internationalt plan. Derudover var det muligt at udføre tekst mining til at opdage mulige unikke tegn, hvilket betyder at det er et godt værktøj til håndtering af store databaser, hvilket er tilfældet i de fleste kliniske sikkerhedsundersøgelser.

Abbreviations

Adverse drug event	ADE
Adverse drug reaction	ADR
Artifactual linguistics	AL
Coding Symbols for a Thesaurus of Adverse Reaction Terms	COSTART
Cytochrome P-450	CYP
Diabetes Mellitus	DM
Drug-drug interaction	DDI
Electronic patient records	EPR
European Union	EU
Food and Drug Administration	FDA
High Level Group Term	HLGT
High Level Term	HLT
Information extraction	IE
Information retrieval	IR
International Classification of Disease	ICD
International council	ICH
Lowest level term	LLT
Machine learning	ML
Medical dictionary for regulatory activity	MedDRA
Maintenance and support services organization	MSSO
Named entity recognition	NER
Natural language processing	NLP
Pharmacovigilance	PV
Preferred term	PT
System Organ Classes	SOC
WHO adverse reaction terminology	WHO-ART

Table of Contents

Mapping of adverse events - identified by a high throughput data mining technology in a diabetes population - to an international consensus terminology

1 Introduction

Computational linguistic is a multidisciplinary field that consist of natural language processing (NLP), algorithms and statistics that is in combination with one another, facilitates interaction between humans and computers. Considering that computers work significantly faster than human beings it was thought of to develop a tool that understood human language in texts by the use of software programs to perform a high quality of information extraction. Thus, this field is becoming significantly important in medical investigations due to the difficulties of handling the large amount of information of clinical data, where most documents are saved as text. An estimate of documents stored in companies have shown that 85% of information is recorded as text (Hotho, Nürnberger, & Paaß, 2005). When investigating specific areas of clinical data that contains large datasets, programs that can distinguish between real data and false positives, and at the same time understand natural human language are necessary. Previously designed programs have not always been able to recognize fuzziness or other ambiguous words in the text. Therefore, an innovative tool, text mining was developed to reveal textual information by applying different artificial linguistic (AL) methods, which aims to understand human language as well as extracting information of interest based on the purpose (Hotho et al., 2005). A regular Google search compared to text mining finds aim to provide already known information. It lists documents containing the terms of interest with both relevant and sometimes irrelevant information, and thus these documents have to be sorted. However, text mining differs from that as it finds information of interest and also aims to uncover new patterns and unknown information (Berkeley, 2003). As Hotho et al. (2005) states, the development of text mining allows researchers, the ability to work with a wide range of terms and at the same time be capable to deal with ambiguity and fuzzy relations (Hotho et al., 2005).

Text mining has the ability to work with a large scale of unstructured text such as clinical notes, and extract relevant information of interest from these unstructured texts and convert it to a structured and machine-readable format. Moreover, text mining can be defined as a method to work towards

detecting possible patterns in information. Extracting information from several recordings can do this, and provide an association between the extracted information as well as generating potential hypothesis. Moreover, medical records are rich in phenotypic data as the information that is recorded reflects the interaction between the provider and patient as well as the clinical observation. It is an area that has attracted great interest in recent years due to the importance of phenotypic information that may describe a possible pathway between the phenotype and genetic. Hence, using such method as the above-mentioned serves the medical industry by assisting in estimation and efficiency of a treatment that demonstrates an association between a condition, their symptoms as well as their treatments. Therefore, such information can be examined in further analysis (Berkeley, 2003).

The characteristic structure of text mining includes a cooperation between several techniques such as information retrieval (IR), information extraction (IE), clustering, categorization, visualization, machine learning (ML) and data mining (Tan & others, 1999). However, the two main techniques that are utilized in text mining are rule-based or ML, which will be explain later.
IR extract relevant terms of interest from unstructured data, where IE extract relevant information from both unstructured and structured data. Named entities recognition (NER) is a part of IE and consist of two different NER approaches, a terminology and rule-based NER and corpus-based NER respectively. Both methods aims to identify and classify important named entities (words) from texts and categorizes them into predefined classes such as names, locations, and organizations. The task aims to recognize named entities and make an association between the named entities and occurred event, respectively. Terminology and rule-based NER aims to map important words from the text to terminological resources. Thus, it employs phrases of the dictionary in combination with entities features, whereby corpus-based NER takes advantages of evidence from text corpus, and commonly used with ML algorithms (Shaalan, 2014). A rule-based approach identifies terms based on a set of rules that describe the information of interest. In addition, in the clustering process data are grouped based on their similarity into clusters. Categorization of the data however classifies the terms into groups based on known terms. Lastly, the findings are presented, where similarities and dissimilarities are shown (Guo & Cao, 2015). Subsequently, beside the above-mentioned algorithms, other areas of intelligence algorithms that can be used for information extractions include ML. As shown in figure 1, areas of intelligence ML to execute data mining consist of several algorithms, which are structured based on information of interest. These algorithms can be used for any interest, however in clinical trials it is used to determine clinical decisions and

diagnosis as well as possible new findings. For some researcher the method is used to find information involving relationships between two different events where other investigators try to extract explicit information from text such as named entities that are mentioned or relations suchlike how A leads to B. In general, ML can be divided into three broad algorithms: supervised learning that includes task such as categorization and regression, unsupervised learning that aims to cluster and discover the inherent structure of the data by excluding the use of explicit labels and semi supervised learning, which is a hybrid between the supervised- and unsupervised learning (figure 1). Each algorithm can be subdivided in further algorithms. Regardless of information of interest text mining as a method has shown to save humans for a lot of time and effort as well as executing a high quality of information extraction (Eftimov, Seljak, & Korošec, 2017). These methods allow to collectively act from assumptions to valuations (Cerrito, 2001).

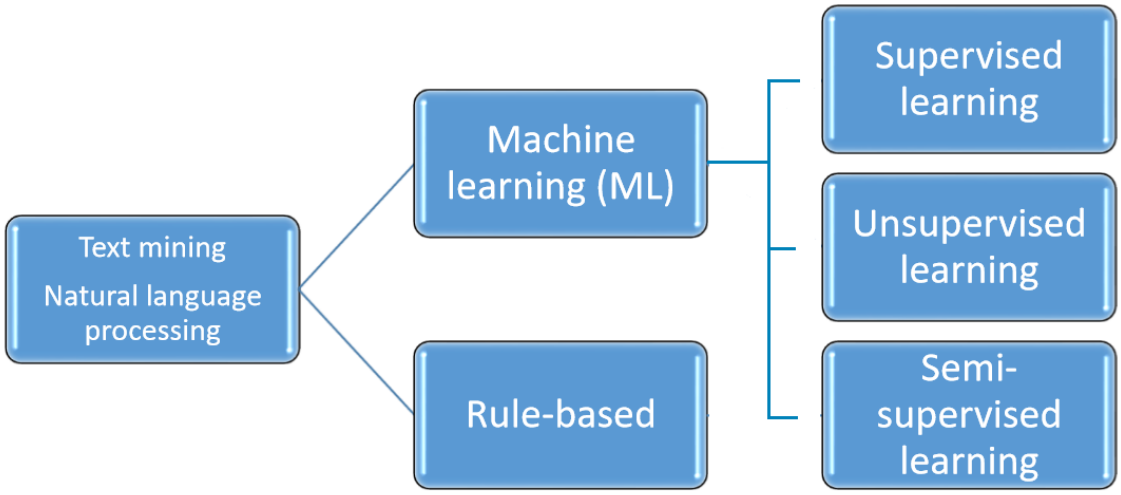

Figure 1: A schematic illustration of algorithms in text mining. Text mining in combination with natural language processing (NLP) can be divided into two main algorithms used for information extraction in clinical trials, which includes rule-based and machine learning (ML). A rule-based approach is executed based on a list of rules, whereby ML based on information of interest can be divided into three algorithms: supervised learning, unsupervised learning and semi supervised learning.

In addition to text mining a distinct form of an analytic process is data mining, which is a part of text mining. The difference between text mining and data mining lies in the name. Text mining works with unstructured natural language text, also known as computational linguistic, whereas data mining extracts information from structured textual sources. With this being said, text is one of the most common forms to store information thus, making text mining potentially more successful (Berkeley, 2003). In any case, text mining and data mining are both methods used to investigate

defined objectives. For example, both are used to examine potential medical issues by establishing and validating a hypothesis (Berkeley, 2003). The significance of formulating and validating a hypothesis is to find similar patterns and associations such as patient findings, hallmarks and therapies. Furthermore, pattern validation gives rise to new unperceived investigation area that can improve patient care by discovering adverse events (AEs), including those caused by drug interactions.

Historically, text mining has been used in clinical data to examine unfound and possible adverse drug reactions (ADRs) in cohorts such as patients with diabetes mellitus (DM) or other groups at risk. Diabetes is associated with several complications due to age and physiological changes leading to the need of multiple drugs to either treat or reduce the symptoms linked to the disease. Multiple drugs results in a higher risk of ADRs and in some instances the treatment can give rise to severe ADRs which, in many cases is difficult to identify in clinical notes, which is why text mining in these certain circumstances serves as a compliment (Cerrito, 2001)**.**

1.1 Named entities recognition tagger in Danish clinical narratives

Previously, the use of IE has been challenging in Danish clinical narratives, as no dictionary has been available in Danish. This lead to a new approach that were established by the in-house research group at Novo Nordic Foundation Center For Protein Research were they developed a NER tagger with the use of a rule and terminology-based system in combination with NLP algorithm (Eriksson, Jensen, Frankild, Jensen, & Brunak, 2013). They founded it to be an advanced technique that has the ability to detect more than 4×10^{12} unique ways of describing AEs. It is used to enable a computer to process and recognize human language as well as extract information from unstructured text by a set of rules and categorize the extracted entities into groups based on their names, locations and disorders. Thus, it is able to determine events from clinical narratives and assemble an association to a dictionary of drug-related AEs. The NER tagger however, also consider different spelling type, synonyms and inflectional variants. Furthermore, anatomical structure groups are created so both cell types and tissues are grouped into one category.

Subsequently, a negation filter, filters out events that have occurred before drug indication and other unqualified events as AEs that is present in clinical narratives not always reflects the patient history. It can be mentioned as prescriber has to notify patients about ADEs that can be related to the drug or the AEs can be misinterpreted, as in some cases a patient receives a certain drug due to the mentioned AE. Therefore, to be able to extract AEs that may be a potential side effect from the clinical narratives a variety of filtering methods have to be used to avoid such complications

(Eriksson et al., 2013). Moreover, terms with similar meaning allow to be united into one word, e.g. kidney failure and elevated creatinine. All unique terms will be extracted, including those in Danish, Latin and English in both upper- and lower-case letters, like *mrsa*, *MRSA* and *Mrsa*. Duplicates of any term will be removed. Moreover, terms in English that had similar meaning as in Danish or Latin will also be removed. The remaining unique terms are validated and mapped to a medical dictionary (Gupta & Lehal, 2009).

1.2 Medical Dictionary for Regulatory Activity (MedDRA)

Medical dictionaries allows researchers to understand, develop and communicate successfully when investigating potential AEs in clinical trials, which is why extracted unique terms that occurs in clinical notes has to be coded into a medical dictionary weather it is before or after an approved drug. Currently, surveillance of ADEs starts during the initial drug cycle process, however, post-marketing surveillance is also required and important to implement after an authorized medical product becomes available on the market. This is due to the possibility of any ADR that may occur, and that has been unnoticed during the drug development cycle. Spontaneous reporting assist to monitor drug surveillance, by reporting identified ADEs to medical product agency or the pharmaceutical manufacturer where all AEs are gathered into the databases such as the FDA Adverse Event Reporting System (FAERS) in the USA, EudraVigilance in the EU and the WHO Vigibase. Any ADR that occurs should be reported to the authorities, but only if the reaction is suspected to be linked to the drug (Cerrito, 2001).

Medical Dictionary for Regulatory Activity (MedDRA) is the latest dictionary used for coding medical events and were established in 1994 by the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) to empower communication on an international level between industries and regulatory authorizes aimed to improve health globally (Cerrito, 2001). Hence, it is imperative for sharing relevant information on presumed adverse reactions after registration in the EU. It is a standardized terminology used for recording relevant information in all stages in drug development from pre-marketing stages (phase 0 to 3) to post market stage, including pharmacovigilance (PV) in phase 4 (Cerrito, 2001).

Previously, other terminologies and dictionaries for reporting ADEs included WHO´s Adverse Reaction Terminology (WHO-ART), which in contrast to MedDRA only consist of three levels: high-level terms, preferred terms and included terms and Coding Symbols for a Thesaurus of Adverse Reaction Terms (COSTART). This terminology is designed to be used along with the International Classification of Diseases (ICD). The idea behind MedDRA was to be able to make a

standardized forum for researcher as well as companies and additionally overcome issues that were linked to the above-mentioned older dictionaries. Thus, MedDRA is more powerful due to the greater granularity of terms and the ability to make communication on an international level possible (Edwards & Aronson, 2000)(SAS, 2009)**.** Moreover, formerly there has been a lack of medical dictionary on Danish. Neither the WHO-ART, COSTART nor MedDRA has been translated into Danish. Therefore, the NER-tagger were developed and founded to serve the Danish languages as well as other Scandinavian languages (Eriksson et al., 2013).

MedDRA is approved by the European Union (EU), USA and Japan and is highly recommended by the ICH for monitoring drugs in any stages of the drug cycles (Edwards & Aronson, 2000). It is organized by a hierarchical structure that comprises of various terms of morbidities the associated marks, symptoms and diagnosis including laboratory results and social conditions. The terminology consists of five levels as illustrated in figure 2, from the broadest to the most detailed it consists of the following levels: System organ classes (SOC) with 26 classes, 332 high-level group term (HLGTs), 1682 high-level term (HLTs), 16293 preferred term (PTs) and 60518 lowest level terms (LLTs). LLTs gives highest specificity and thereby reflects how an observation is reported in practice. Any LLTs is linked to one PT, and is associated to their parent PT by either being synonyms, lexical variants, quasi-synonyms, sub-concepts or identical. Moreover, LLTs may be identified by “current” or “non-current” labels whereby the last mentioned describes terms that are unclear, ambiguous, abbreviated, out-dated or misspelled. PTs is a diverse descriptor, it demonstrates terms that reflects symptoms, signs, disease diagnoses, therapeutic indications, investigations, surgical or medical procedures, and medical, social or family history characteristics. Adding an additional PT in the dictionary is followed by an identical LLTs that is linked to the PT. as mentioned above each LLT must be linked to at least one PT, however more than one LLT can be linked to a PT. There is no maximum of how many LLTs that can be linked to a PT. PTs that are linked together based on their anatomy, pathology, physiology, etiology or function are clustered into HLTs. HLTs however, are grouped into HLGTs. To categorize a HLT to a SOC, it must be linked to a HLGT. Lastly, HLGTs are categorized into 26 SOCs, based on etiology (e.g., Infections and infestations), manifestation site (e.g., Gastrointestinal disorders) or purpose (e.g., Surgical and medical procedures). Thus, one SOC differs from the other as it describes social circumstances

(SAS, 2009).

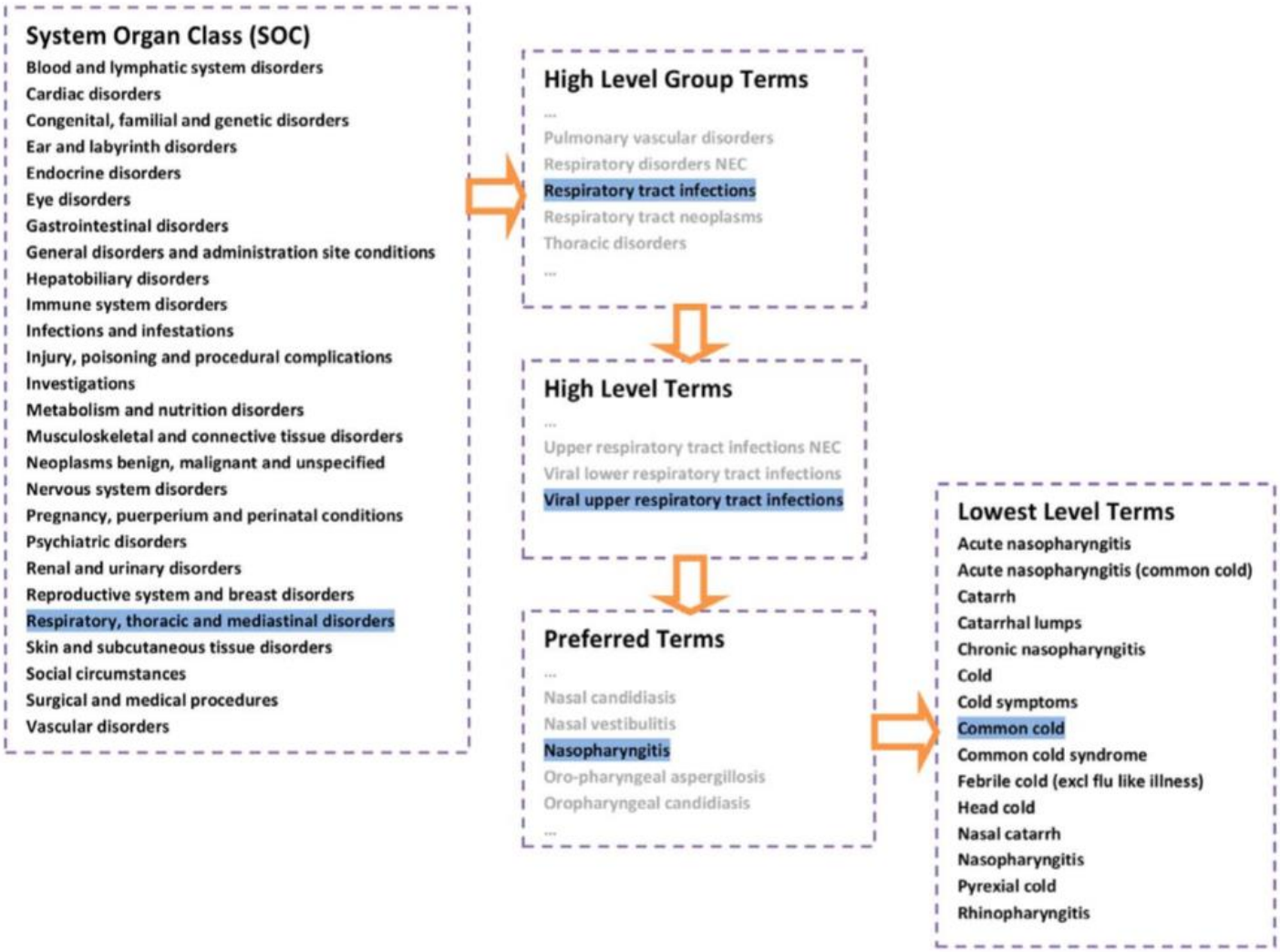

Figure 2: An example of the hierarchical structure of MedDRA. The term "common cold" is demonstrated from a more general level to a specific level (Schroll, Maund, & Gøtzsche, 2012).

Single concepts in PTs may be investigated from different perspectives, as some terms may be present in more than one SOC, a phenome called multiaxial, thus this creates some issue when making a conclusion (figure 3). However, all PTs are designed to be linked to one primary SOC to avoid over counting of the terms. Each PT is only linked to one HLT, HLGT and SOC. Every term is characterized by an 8-digit number as an identification number, which describes the original term of the ADR (Brown, 2004) (SAS, 2009).

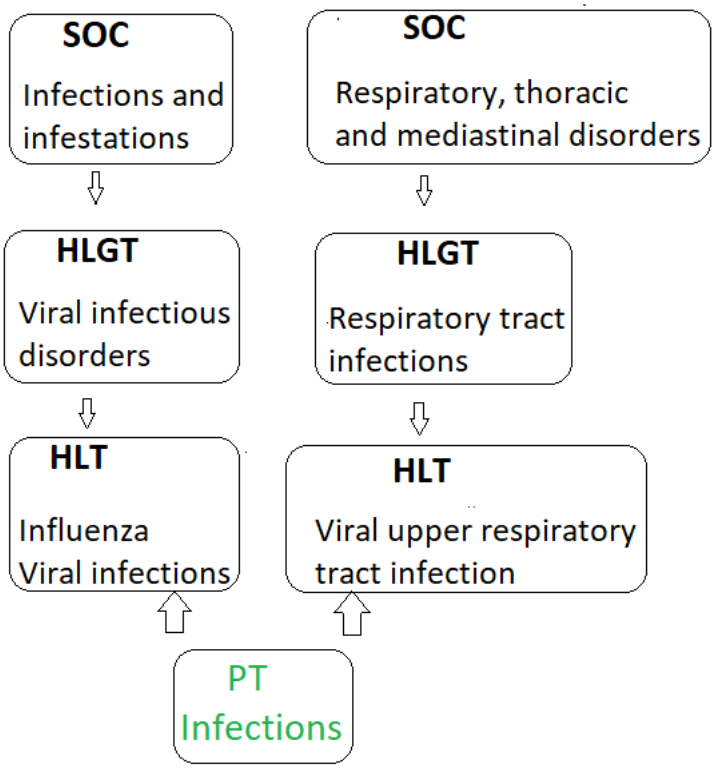

Figure 3: An example of a multiaxial term is illustrated. The term "Influenza" is a symptom that is present in two different System Organ Classes (SOC). Therefore, each preferred term (PT) is linked to a primary SOC, which in this case will be "infection and infestations" and to a secondary SOC "respiratory thoracic and mediastinal disorders" (SAS, 2009).

Detected ADEs in safety trials are analyzed and mapped into MedDRA codes. This kind of communication allows researcher to understand one another across different countries and languages as well as in an intern communication. Originally, MedDRA was available only in English and Japanese. However, it is now multilingual, as it has been translated in Chinese, Czech, Dutch, French, German, Hungarian, Italian, Portuguese, Russian, and Spanish but not in Danish (figure 4). The numerical code each term is defined by remains the same regardless of language. So, a Dutch user can understand a Chinese user based on the codes. This kind of communication allows researcher to act in their mother language, which leads to a more correct and precise translation of the terms into codes. Hence, interwork is of great benefit and gives rise to easily share data on a multinational level

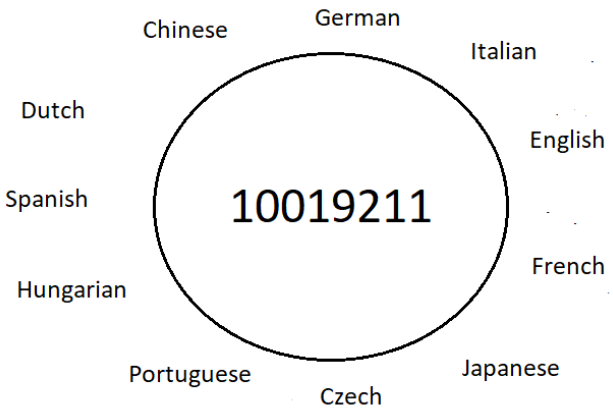

Figure 4: As shown in figure 4, MedDRA is a multilingual terminology. It has been translated into several languages that aims to make communication on an international level possible (SAS, 2009).

(SAS, 2009).

Due to increase medical knowledge and changes in regulatory aspects as well as in medical and scientific aspects, an updated version of MedDRA has to be done regularly. ICH has established a highly skilled team under the name, maintenance and support services organization (MSSO), which maintenances and updates MedDRA twice a year (Brown, 2004)(SAS, 2009).

2 Diabetes Mellitus

In this study, we applied a rule-based NER-tagger algorithm in a diabetes population to find possible unique terms from clinical notes. Due to their increase risk of polypharmacy as a treatment, unnoticed ADRs may be found. Therefore, to get a better understanding of why some patient groups are more prone to develop possible ADRs and how text mining can be used in such a group, an introduction of DM, the associated complications and treatment will be introduced.

DM is a global health issue and is one of the leading causes of death worldwide. In line with an increasing prevalence of obesity and unhealthy lifestyles, the number of DM through the years has increased from 108 million in 1980 to 422 million in 2014 and this number is expected to raise to 642 in 2040. In addition to the high number of estimated DM, unnoticed DM has also shown a high rate with 193 million people that express the disease, which means that the actual number may be higher than estimated. The development of DM depends on different factors such as age, sex, ethnicity and geographical variations as some populations are more susceptible to develop DM due to cultural eating habits (WHO (2008), n.d.) (WHO, 2004).

DM is a metabolic disease characterized by insulin deficiency in either secretion and action, or both. It is marked by hyperglycemia, which in a chronic stage leads to dysfunction in several organs such as eyes, kidney, heart, nerves and blood vessels due to elevated sugar level, which gives a broad range of effects, such as hypertension, microangiopathy and other various vascular events (Canivell & Gomis, 2014,)(Ozougwu, 2014).

The development of DM occurs in pancreas, which is an organ located behind the stomach with a length of 12.5-15 cm. The pancreas is connected to the first part of the small intestine, which assist the digestion of food intake with enzymes and thereby absorb glucose from the food. Pancreas serves the body with two main functions, an exocrine and endocrine. The exocrine function helps the digestive system by secreting digestive enzymes into the duodenum whereas the endocrine function helps to regulate the blood sugar level by specific types of cells. These cells are located in

small islets in the pancreas, defined as islets of Langerhans, which contains alpha, beta and delta cells. Each cell type differs from their characteristic components by their function, morphology and staining. The alpha cells are responsible for secreting glucagon and is stimulated when the blood sugar is low whereas the beta cells is responsible for secreting insulin and is stimulated when the blood sugar level is high. Insulin is an important hormone as it ensures glucose absorption in muscle -, liver and fat cells to use it as energy. In addition, delta cells secrete somatostatin whose function is to suppress secretion of insulin and glucagon (Nielsen & Bojsen-Møller, 2016).

Insulin is a hormone that is normally present in low amount in the blood, however, when we eat a meal, the concentrations of blood glucose increases, which leads to stimulation of increased insulin secretion from the pancreas. The majority of glucose is absorbed in muscle- and lipid tissue and used as adenosine triphosphate (ATP), where excess glucose is stored in the liver as glucagon. The absorption of glucose is mediated through glucose transport protein, mainly via GLUT4, by a concentration gradient from a high glucose concentration extracellular to a low glucose concentration intracellular. This process is also known as GLUT4-translocation. The liver however, absorbs glucose through GLUT2 and store the glucose as glucagon for future use. Under fasting, the blood glucose level is low, however the liver contains the enzyme glucose-6-phosphatase, which dephosphorylates of glucagon to glucose. This conversion is necessary before releasing it to the blood (Nielsen & Bojsen-Møller, 2016).

So, when an abnormality in the blood sugar level occurs, the cells sense it and execute their function. Alpha cells sense when the blood sugar level is low and secrete glucagon, whereas beta cells sense when the blood sugar level is high, and secrete insulin. However, in diabetic patients' damage to beta cells lead to a decreased insulin level, and therefore absorption of surplus sugar from solid and liquid food cannot be absorbed, resulting in an increased blood sugar level.

The underlying mechanism for developing insulin dysfunction can be classified into two different types referred as type 1 and type 2. The first, also called the insulin-dependent diabetes, occurs due to either a genetic or an environmental factor, which mediates an autoimmune destruction of the pancreatic beta cells. In the vast majority of cases type 1 diabetes is caused by an autoimmune mediated destruction of the cells, which means the immune system produce autoantibodies to destroy the cells (Levy, 2016). Histologically in the diabetes debut pancreas develops insulitis, which is a mononuclear cell infiltrate that mainly contains T-lymphocytes and to a lesser extent macrophages and B-lymphocytes. The T-lymphocytes are divided into Tc-lymphocytes (CD-8-positive T-lymphocytes) that is toxic aiming to kill cells, and Th-lymphocytes (CD-4-positive T-

lymphocytes) that helps the activation of Tc-cells. The development of insulitis may occur due to an unknown environmental factor or gene variation that cause beta-cell destruction in islet of Langerhans. This releases beta cell specific antigen that are presented for macrophages and antigen presented cells, which then migrates into the pancreatic knot, activate the Th-lymphocytes where a clonal expansions activation is executed. Activated Th-lymphocytes migrates into the islets of Langerhans, releases the cytokine IFN-y that attracts additional macrophages, which are stimulated to release IL-ab and TNF-a resulting in destruction of beta cells (Yoon & Jun, 2005)(Ozougwu, 2014). An illustration of the above mentioned is shown below in figure 5.

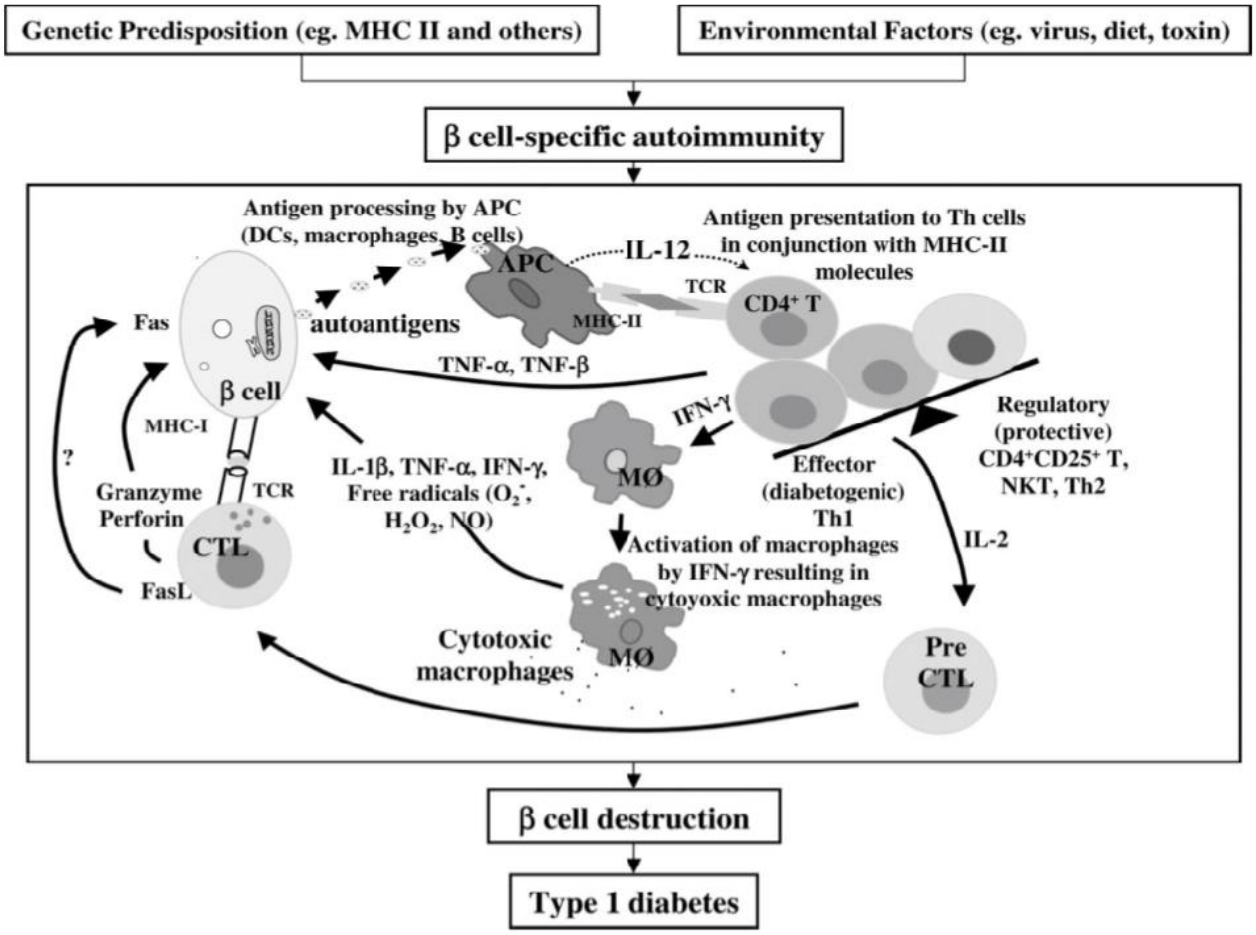

Figure 5: Demonstration of the etiology of type 1 diabetes. Genetic predisposition or environmental factor is the cause of the activity of an autoimmune response. This leads to destructions of the pancreatic beta cell resulting in the onset of diabetes type 1 (Yoon & Jun, 2005).

This type however, occurs in only 5-10% of the diabetic population and is mainly present in younger patients and leads to an absolute elimination of insulin secretion (Levy, 2016)(Atkinson, Eisenbarth, & Michels, 2014).

The second large group of diabetes, also called non-insulin dependent is the most prevalent type accounting for 90-95% of the diabetic population. This type is caused by a dysfunctional insulin action and/or secretion and is often seen in patients with obesity, sedentary lifestyle and increased age (Canivell & Gomis, 2014)(Ozougwu, 2014)(Chatterjee, Khunti, & Davies, 2017). The insulin resistant in type 2 diabetics is primary located in the skeletal muscle but also in the liver and as a result, action of the hormone in the muscles and the liver is not sufficient. Moreover, an increase of free fatty acid (FFA) has shown to be present, due to a decrease absorption in fatty tissues. Resulting from an increased FFA the gluconeogenesis rises as well as the production of triglyceride. Accumulation of triglyceride has been detected to disturb the insulin signal cascade in both the skeletal muscle and liver (Nielsen & Bojsen-Møller, 2016). A demonstration of the two different diabetes types as well as a healthy pancreas is shown below in figure 6.

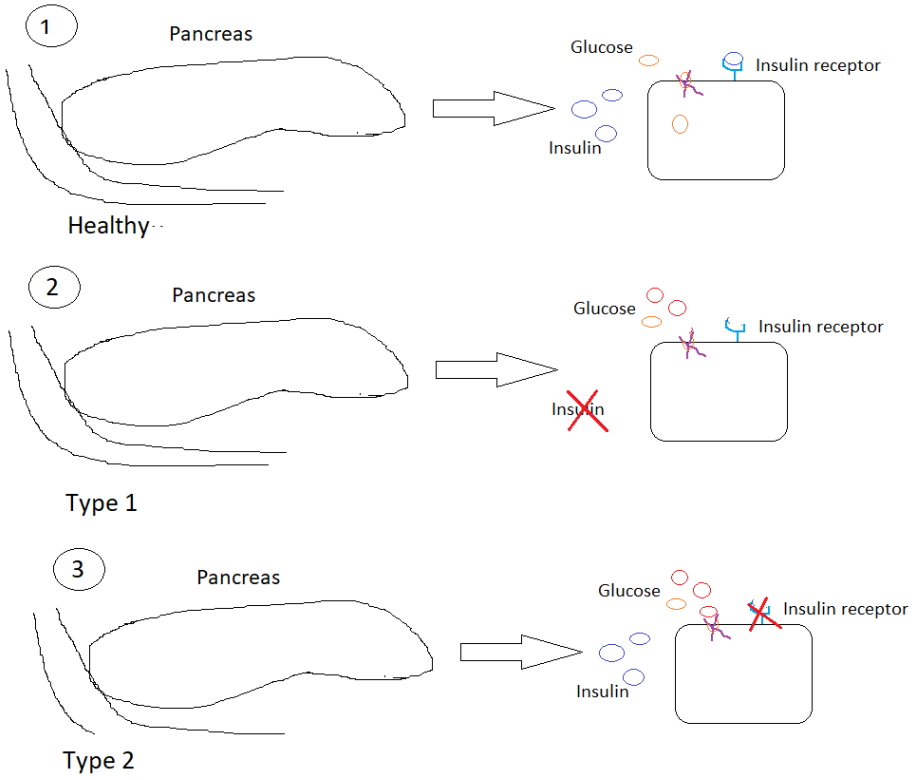

Figure 6: 1: Illustrates a healthy pancreas that secrete insulin, which then binds to the insulin receptor were it absorb the glucose.2: Illustrates type 1 diabetes, were pancreas is not able to secrete insulin, which leads to increase blood sugar level. In addition, 3 illustrates type 2 diabetes were the pancreas is able to secrete insulin, however the insulin receptors fails to response to the insulin leading to high blood sugar level.

2.1 Complication due to long-term hyperglycemia

The associated complications due to long-term hyperglycemia can be divided into microvascular and macrovascular complications. The microvascular complications include retinopathy, nephropathy and neuropathy, which occurs due to damage of the small blood vessel caused by hypertension and high blood sugar levels. In addition, researcher also think there is a direct link to the neurons, where uncontrolled high blood sugar damages the nerves and interferes with their ability to send signals, which gives rise to neuropathy and retinopathy (Canivell & Gomis, 2014). The macrovascular complication includes cerebrovascular disease, cardiovascular disease and diseases in the peripheral. Due to an increase blood sugar level, diabetic patients have a higher chance of developing hypertension and therefore a higher incidence of cardiovascular events compared to non-diabetics, leading to a higher mortality. In addition to hypertension, changes in lipoprotein metabolism are also common (Laing et al., 2003)(Intervention & Disease, 2003)(Canivell & Gomis, 2014).

2.2 Treatment/prevention of diabetes

Diabetes patients are often treated with polypharmacy, which is a multidrug regimen, to control the blood sugar level and lower the associated risk factors. Reducing the risk associated with diabetes is important and beneficial for the patient as reducing future complications leads to a lower need of drugs, but unfortunately, there is a high rate of polypharmacy needed to achieve this. This tends to result in a higher risk of interaction between the drugs given for diabetes and drugs given to prevent cardiovascular events, which may give harmful side effects (Services, 2014).

Prevention of diabetes can be either primary or secondary. In type 1, primary treatment includes insulin analogue as no other preventions or treatment will improve the amount of insulin these patient's needs (Atkinson et al., 2014). In type 2, evidence points towards patient can control the disease by altering their lifestyle for example, by increasing their physical activity level with a low fat diet that results in weight loss and better glycemic control (Atkinson et al., 2014).

3 Polypharmacy

The word polypharmacy derives from the Greek and consist of two words, *"poly"* that refers to many and *"pharma"* that refers to medication. A standard definition is established in the literature, as a phenomenon often referred to the use of five or more medications on daily basis. Polypharmacy in treatments are common in elderly adults due to age-related physiological, biological, functional and social changes (Bushardt, Massey, Simpson, Ariail, & Simpson, 2008). Consequently, both pharmacokinetics and pharmacodynamics of drug disposition and action may be disturbed (College & Doddi, 2016).

Previously studies has shown that about 80% of older patients suffer from one chronic disease, whereas half of them suffer from two chronic diseases meaning these patients are likely prescribed more than one drug to either reduce the disease or prevent future potential complications (Shah & Hajjar, 2012). Various age-related conditions such as kidney failure, heart failure, diabetes and arthritis require more than one drug to cure or reduce the symptoms (Strehl, 2013) (Shah & Hajjar, 2012). This is supported by a study, where they measured the prevalence of prescribed medication and found that elderly patients have 31 prescription per year, which is double the number compared to other age populations (Rollason & Vogt, 2003). Other studies have investigated the incidence of polypharmacy in elderly patients. Based on two studies conducted in France, both revealed that polypharmacy, as a treatment is common in elderly patients. The first study observed prescribed medications and found 49% of women and 45% of men were using minimum five or more prescribed medications resulting in drug-drug interaction (DDI) with a number of 21% and 12%, respectively. The second study examined elderly patients living at home where it was shown that 38% of elderly were taking five to ten medications (Rollason & Vogt, 2003).

It is however, worth mentioning that polypharmacy, as a treatment is not always having a negative impact on a patient. Although evidence supports the complication associated with polypharmacy, in some cases patients have a beneficial effect of the treatment to cure, reduce or decrease the progression of disease. However, it is difficult to balance in-between risks and benefits within elderly patients (Bushardt et al., 2008).

3.1 Complication of polypharmacy

Unfortunately, polypharmacy as a treatment regimen is associated with numerous complications such as inappropriate prescribing, drug-to-drug interaction (DDI), non-adherence, increase hospitalization and cost, ADRs and geriatric syndromes, which will be describe below.

3.2 Inappropriate prescribing

Inappropriate prescription is defined as a medication that does not fit the disease and therefore considered as an unnecessary use of it. Patients with polypharmacy treatment are prone to get an inappropriate medication. Hospital visits due to multiple drug intakes lead to exposure to different physicians, which can lead to a falsely interpretation of the AEs. Thus, it will either be interpreted as a symptom from an existing disease or the patient will be diagnosed with a new illness leading to additional prescription. This statement is backed up by a study where it was demonstrated that 80% of AEs occurring in patients were detected as an illness and therefore a new drug was added to their list (Bushardt et al., 2008). Further, patients by themselves also contribute to inappropriate medications due to self-medication. The mindset of elderly patients often refers to, as "*every ill needs a pill*". Moreover, elderly patients are more inclined to borrow medication from friends and family (Bushardt et al., 2008).

3.3 Drug-to-drug interaction

DDI is one of the most important areas in polypharmacy treatment. The risk of DDI increases when larger amounts of drugs are utilized in patients as one drug can inhibit the action and efficiency of another drug. A study conducted by Richard et al. (1996) support this statement, where a correlation between the number of drugs and the risk of interactions was found. Patients who received two medications had a 13% risk of adverse drug interaction (ADI) where those who received seven medications had an 82% risk of ADI (Goldberg, Mabee, Chan, & Wong, 1996).
Interactions between drugs can either be beneficial by increasing the effect of the drug and reduce possible side effects, or give a negative response by decreasing the effectiveness of the drug or increase the adverse effects. The last mentioned develop when more than one drug share the same metabolic pathway as well as absorption and excretion. Moreover, interactions between drugs can either disturb the pharmacokinetic, which include what the body does to the drug, or pharmacodynamics, including what the drug does to the body. The pharmacokinetic covers the effect of one drug on the absorption, distribution, metabolism and excretion of another drug, which among other reasons leads to changes in serum drug concentration and often one specific

cytochrome exhibits high affinity for a target enzyme, which means action of the enzyme has a great influence on the metabolism and pharmacokinetic of the drug. However, overlap with two enzymes may occur (GR, 2005). Thus, the result of pharmacodynamic interactions may show effects of the drug, including potential side effects (Goldberg et al., 1996). Due to multifactorial differences as well as genetic difference among patients pharmacokinetic and pharmacodynamics differs. A challenging aspect caused by genetic differences, as a specific dosage to one patient may not have the same effect on another patient (Goldberg et al., 1996).

In accordance to evidence, drugs that share same metabolic pathways are metabolized through similar cytochrome P-450 (CYP) interactions as well as systems. CYP is a large enzyme family accounting for 57 enzymes and is mainly present in the liver, located in the epithelium, but is also present in the intestinal mucosa. The enzymes are classified based on their family number (CYP1, CYP2), subfamily letters (CYP1A, CYP2D) and number of their isoform (CYP1A1, CYP2D6). These enzymes are important for metabolizing nutrients and drugs, were the drugs can act on the CYP activity as an inhibitor, either inducers or substrates, which means the process either increases or reduces the action of a drug and promotes the elimination. However, it is also demonstrated in the literature that not every CYP enzyme promotes drug activity, thus only a few CYP have been identified to contribute to the metabolism of drugs. These specific CYP families include CYP1, CYP2 that is present in the liver, and CYP3 families that is present in both the intestinal epithelium and liver. Administration of a drug reaches the stomach, which then passes through the intestine where it dissolves and absorbs into the blood. Through the blood, it reaches the liver where CYP perform its function in conjunction with cytochrome p-450 reductase (NADH) and the associated cofactors such as oxygen and NADPH. The different cytochromes metabolize different drugs, and some drugs however, have a low bioavailability, which is the fraction of a drug reaching the systematic circulation. This is caused by a heavy first pass metabolism, which is where a large proportion of the drug is metabolized in the liver by p-450 enzymes so only a little of the drug reaches the bloodstream (Cytochromes_P450_2002.pdf, n.d.) (Brøsen, Simonsen, P.kampmann, & Thirstrup, 2014). As mentioned before, a great majority of drugs are metabolized in the liver by the CYP system. The process they may undergo consist of two phases, a chemical process including oxidation, reduction or hydrolysis, also known as phase I, followed by a conjugation process, also known as phase II. As a result, metabolites are produced, by converting lipophilic compounds to a more hydrophilic product so it can facilitate the excretion (Brøsen et al., 2014).
A majority of drugs however are metabolized by phase I reactions through oxidation were the

enzymes introduce or incorporate polar groups such as –OHs and –O´s atoms on the drug leading to a more water-soluble product. Further, a phase II reaction may occur where it aims to bind a drug or metabolite to a molecule for instance glucuronic acid, sulfuric acid, acetic acid or glutathione. Lastly, the drug will be excreted due to a much more hydrophilic product by urination (figure 7). It is however, worth mentioned that not every drug elimination undergoes both phases (Brøsen et al., 2014).

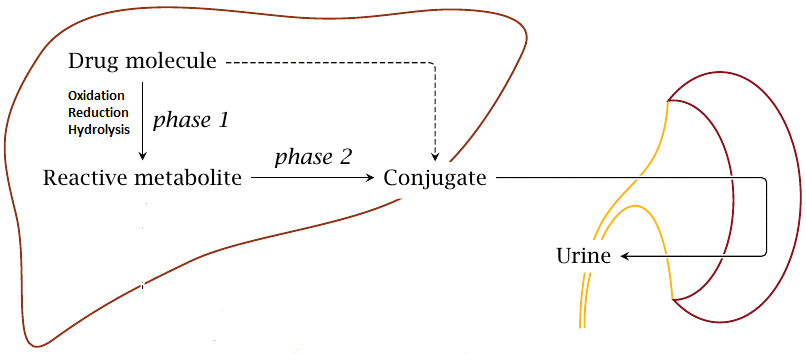

Figure 7: Drug metabolism through cytochrome p-450 system. The drugs may undergo a phase I, phase II or both respectively. In phase I, the drugs can either be metabolize by oxidation, reduction or hydrolysis, thus a majority of drug is metabolized by oxidation were polar groups binds to the drug. In phase II, the drug undergoes conjugation were a metabolite binds to the drug. This leads to a more hydrophilic elimination by the urine ("Drug Metabolism," 2013).

The most important CYP enzyme in drug metabolism is the CYP3A subfamily, as it is the most significantly enzyme expressed in the liver and small intestine, accounting for 50% of all CYP enzymes. CYP3A have three isoforms: CYP3A4, CYP3A5 and CYP3A7 and through these isoforms the most frequently drugs are metabolized. Another important aspect of CYP3A is the capability to metabolize means of distinct enzymes, however this CYP family has the capability to metabolize almost all drug classes (GR, 2005). In addition, another important CYP enzymes is CYP2D6 that belongs to the CYP2 family and metabolize up to 65 of the most common drugs (GR, 2005).

Classifications of patients are based on how they individually metabolize administrated drugs. Patients have either low ability to metabolize or the opposite with a high ability to metabolize. Therefore, it may give rise to potential adverse reaction (GR, 2005). Suppression of an enzymatic CYP pathway caused by a certain drug perchance greater concentration of another drug that share similar CYP pathway, which leads to drug toxicity. Similarly, if a drug persuade the enzymatic pathway of a drug it decrease the concentration of another drug that share similar pathway, which gives rise to treatment failure (GR, 2005)(Cytochromes_P450_2002.pdf, n.d.).

In addition to the above-mentioned, a correlation between drug-disease interactions has also been detected. A few studies have reviewed the most common drug-disease interactions and observed that interactions between aspirin and peptic ulcer disease, calcium channel blockers and heart failure and beta-blockers and diabetes were common. These interactions are of interest, as elderly patients often develop the conditions mentioned above and apart from that, they might have more than one (Tatonetti et al., 2011)(Kadam, Mahadik, & Bothara, 2007). In addition, Goldberg et.al (1996) also found an interaction between drug and disease with a connection in 22% of cases, indicating a high risk of developing a drug-disease reaction is possible (Goldberg et al., 1996).

Various other studies have investigated the relationship between medication and potential ADI. According to Beers et al. (2013), 89% of ADI were associated with drugs such as: Narcotic, analgesic, nonsteroidal anti-inflammatory drugs, benzodiazepines, antacids and diuretics (Beers, 2013). In addition, Karas et al. (1981) identified ten drugs to be 90% associated with ADIs, including aspirin, steroids, digoxin, propranolol, aminophylline, prochlorperazine, quinidine, penicillin, acetaminophen and chlorzoxazone (Karas, 1981). Additionally, a third study conducted by Herr et al. (1992) revealed 90% of ADI were related to antihypertensives, digoxin, theophylline and carbamazepine (Herr, Caravati, Tyler, Iorg, & Linscott, 1992). Lastly, Goldberg et al. (1996) showed the high percentage of ADIs compared to the other above-mentioned studies. Based on the study 98% of detected ADIs were associated with 11 drugs (Goldberg et al., 1996).
In general, based on evidence, it can be assumed that there is a correlation between AEs and specific drug classes (Karas, 1981)(Goldberg et al., 1996).

3.4 Non-adherence

Non-adherence is defined as a patient not following the treatment plan prescribed from medical providers and is a phenomenon often seen in patients receiving a multidrug treatment due to a more complicated process (Rollason & Vogt, 2003). Based on a study, non-adherence in elderly patients ranges from 43.7% to 100% (Vik, Maxwell, & Hogan, 2004). Moreover, a report revealed that patients who took one medication had a 15% risk of a drug error, a higher number with patient taking two to three (25%) and even higher with four (35%) medications, indicating that a higher number of drugs lead to higher risk of drug error (Rollason & Vogt, 2003). The patient can cause the reason for non-adherence, which can either be intentional or non-intentional. Intentional is when the recommendations of the drug are not followed for example, not buying the medication, use more or less of the drug or they may intake the drug at the wrong time. Moreover, it is also believed

that non-adherence is more linked to the number of drugs rather than the age of the recipient. Consequently, non-adherence increases morbidity, mortality, disease progression, treatment failure, hospitalization, AEs and overall economic cost (Rollason & Vogt, 2003).

3.5 Increase hospitalization and cost

Polypharmacy contributes to an increase in hospitalization and economic pressure for both the patient and healthcare system due to the associated complications. The most prevalent population is patients from 65 years and older. An approximation that was made showed an estimate of 2.8% were hospitalized due to ADEs, which will be 245,280 visit per year, cost estimated to 1.3 billion (Demiology, 1998).

3.6 Adverse drug reaction

Investigators have found new methods and other beneficial strategies to reduce or cure diseases. However, ADRs are still a concerning issue in the real-world clinical setting, signifying the importance of PV after medications have been approved to enter the market. PV, is known as a drug safety process defined by WHO as "*The science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other drug-related problem*" (Härmark & Van Grootheest, 2008). Where the aim is to minimize the hazards associated with the treatment and thereby enhance the patient care and safety, however, in every drug cycle a post-approval surveillance is needed to detect and discover potential ADEs (Härmark & Van Grootheest, 2008).

ADEs, together with DDI is one of the most important areas of the daily clinical environment. The consequence of polypharmacy is an increased risk of ADEs. According to WHO, ADEs is defined as "*a response to a drug that is noxious and unintended and occurs at doses normally used in man for the prophylaxis, diagnosis or therapy of disease, or for modification of physiological function*" (Edwards & Aronson, 2000). ADRs is classified into six different types dependent on different factors. Type A, is referred as a dose-related reaction (Augmented) and type B is referred to non-dose related reactions, (Bizarre), respectively. Other types of ADRs include a dose-related and time-related, time-related, withdrawal, and failure of therapy (table 1). It is however, also worth mentioning that the term ADE and ADR are different, due to the causality. ADE is any occurrence due to a treatment that not necessarily causally related, where ADR is due to the treatment resulting in a noxious or unintended reaction. Thus, this part is challenging as many patients receive more than one drug, which makes it more difficult to distinguish between them and their actions (Edwards & Aronson, 2000).

Type of ADE	Cue	Features
Dose-related	Augmented	• Common • Related to pharmacological action of the drug • Predictable • Low mortality
Non-dose related	Bizarre	• Uncommon • Not related to a pharmacological action of the drug • Unpredictable • High mortality
Dose-related and time-related	Chronic	• Uncommon • Related to the cumulative dose
Time-related	Delayed	• Uncommon • Usually dose-related • Occurs or becomes apparent sometime after the use of the drug
Withdrawal	End of use	• Uncommon • Occurs soon after withdrawal of the drug
Unexpected failure of therapy	Failure	• Common • Dose-related • Often caused by drug-interactions

Table 1: The classification of adverse events is grouped into six classes as shown above which is related to either dosage or time (Edwards & Aronson, 2000).

The assumption about polypharmacy increasing the risk of ADRs has been supported by a study where it was found that patients who received two drugs had a 13% risk of ADR, and those taking five drugs had a 58% risk of ADR, and up to 82% when receiving seven or more drugs. That indicates the risk of ADRs increases with a larger number of drugs. Typically, these are drug classes often prescribed to elderly patients, which are linked to the most common conditions in this group

population such as drugs for cardiac treatment, renal treatment, antibiotics and anticoagulants nonsteroidal anti-inflammatory drugs (Shah & Hajjar, 2012).

3.7 Sensitive population

Elderly patients are a sensitive population towards polypharmacy. Syndromes such as cognitive impairment, falls, urinary incontinence and nutrition disturbance is present in such a group. Delirium and dementia are often seen in elderly patients. Some drugs have shown an association between a progressive delirium and some specific drugs including opioids, benzodiazepine, anticonvulsant and anticholinergic. Moreover, almost same drug classes aggravate dementia, which includes benzodiazepine, anticonvulsant and tricyclic antidepressant. In conclusion, reduced mental health is related to the need of multiple drugs (Kojima et al., 2011).

Another concerning problem in the elders is an increased risk of falls leading to higher potential of comorbidity and mortality. In the study, they compared three groups, including a group that has never had a fall episode before, a group that had one fall episode and a third group that had multiple falls. The results showed a significant association between falls and the number of prescribed medications, another study conducted in Amsterdam showed that patient who received more than four medication had a higher risk of falls. Furthermore, psychotropic and cardiovascular drugs are also in concern as combination with these kinds of drugs also increases the risk of falls (Kojima et al., 2011).

A third symptom elderly patient are prone develop is urinary incontinence. Patients receiving a drug that causes polyuria also had a decrease in sensory input and bladder contractility. Out of 126 patients, about 60% had urinary incontinence and these patients were on at least four drugs indicating a higher number of drugs leading to higher potential of urinary incontinence (Shah & Hajjar, 2012).

Lastly, elderly patients are also susceptible to nutritional disturbances, due to imbalanced intake of cholesterol, glucose, sodium, soluble, vitamins and minerals (Shah & Hajjar, 2012).

4 Purpose of the study

To investigate undesirable effects caused by polypharmacy, information regarding AEs has to be transferred to a structured format and this is the core of the project. The main goal in this study is to lay framework to enable text-mining investigations into potential ADEs conducted on electronic patient records (EPRs) acquired from a cohort of diabetic patients. Moreover, we sought to identify the most common drugs used in this population and detect the associated potential ADEs in each drug, if possible.Since no dictionary on Danish have been available, the in-house NER-tagger has shown to be of great interest. It has been extended to be able to detect more than $4x10^{12}$ unique ways of describing AEs. More than 90% of these are already mapped to MedDRA but to perform high quality text mining the mapping needs to be comprehensive in terms of AEs identified. Therefore, I manually inspect and annotate AE descriptions derived from the text mining. Moreover, an evaluation of the method will be done and the associated challenges linked to the use of it.

The aim of the second part of the project was to investigate the frequencies of the mapped terms, once these have been mapped to MedDRA.

This study adds a possible new approach to Danish clinical narrative to be able to be mapped into a standardized medical dictionary, MedDRA, by the use of the in-house NER-tagger. Moreover, it uses a non-hypothetical data driven approach to extract valuable information regarding out from its corpus.

5 Method

The collection of the data in this study was based on automatic computational text mining in combination with NLP algorithm, which aims to convert unstructured data into structured data that can be used in further analysis. A rule-based NER approach is used to extract any potential AEs in combination with the in-house dictionary. To perform a high-quality output and optimize the method, another corpus is executed to capture other ways of writing, which I did for two months. Subsequently, analyzation of clinical records from Steno Diabetes Center was performed through seven months.

5.1 Data collection

The collected and analyzed data I worked with was based on EPRs from Steno Diabetes Center. The Steno Diabetes Center consist of a large database, thus I was able to extract and detect potential AEs from these patient records, whereas each patient had at least one prescribed medication. The gender distribution in the study population was 68.9% men and 31.1% women between 17-90 years, with an average age of 61± 15.6 SD.

5.2 Information extraction

Extraction of any potential AEs as well as phenotypic profiles from unstructured or semi-structured Danish clinical narrative, were obtained by the use of a rule-based algorithm in combination with the in-house NER tagger, which has been described further in the introduction section 1.1 (Eriksson et al., 2013)

5.3 Mapping

Extracted unique terms from the EPRs were stored in a file. Each term was characterized by codes. A software was written by a group member to retrieve 15 random narrative texts for each term by their codes, to validate the extracted terms (figure 8, step 2). The term was evaluated to either be a potential AE or removed. The validation was executed by looking into each random text and determine if the extracted term were a false positive term due to a computational artifact or a positive result that could be a potential AEs. A false positive result may be “sygdomsforløb”, which does not describe an AE but more how the course of the illness would be or “patologisk CTG”, which is a part of the patient history and not an actual symptom. In some cases I found terms that was interpreted wrongly by the computer, such as “*kar%sten*”, which probably could be a symptom if you look at the word separately, as “*kar*” can be interpreted as “vessels” and “*sten*” as “*stones*”.

However, I found it to be a name of a person "*Karsten*", which is a computational error, and such words were removed. However, terms that were validated to be a positive outcome were mapped into ID codes meaning that the term may be a potential AE suchlike "brystkræft", which were translated into "breast cancer" and thereby mapped into MedDRA. MedDRA is not translated into Danish, therefore each word were translated from Danish to English to be able to map the term into MedDRA (figure 8, step 3). An 8-digit number identified the extracted terms that were validated to be a potential AEs. This were done by looking up each extracted word in MedDRA were there was an identification number that described the term (figure 8, step 4). The goal was to map the terms into LLTs as it shows high specificity suchlike "pain" were mapped into the ID number "10033371", which is at the LLT level and many of the terms achieved that level of the MedDRA hierarchy. However, in some cases, some terms were mapped to PTs due to a more correct or meaningful translation of the word. As an example the extracted term "influenza" were mapped to "10022000" at the PT level, thus the challenging part was that this term is a multiaxial terms meaning that it is present in more than one SOC suchlike in "Respiratory, thoracic and mediastinal disorders" and in the "Infections and infestations" group. Moreover, terms that could not be translated into an ID code or a correctly synonym that possible could describe the term were removed suchlike "graderet" as I found in my dataset. Although I could find a potential synonym for the term an ID code were not possible to find. Moreover, MedDRA consist of synonyms, which means I was able to map a single term into two different codes, but with similar meaning, as an example I mapped "skin" as "derm". In addition, I mapped terms such as "hyperprolactinemia" as "increased prolactin" based on my assessment, as both terms has similar meaning but is described in different ways. Terms that could not be mapped into an ID code additional word were added to be able to map the term into an ID code. As an example, it could be "Epidural", thus this term cannot be mapped to an ID code due to an insufficient translation therefore I added "Epidural blockage" were an ID code were achieved. This was done by adding the additional word into the file after I validated the term in the software program, were it then was saved. Or in some cases I changed the word to get an better translation suchlike "smerter kolik" I was not able to map the term to a ID code, however, I changed the term to "abdominal pain" as the medical notes I retrieved contained similar content. Another example were "kontusion lunge" were I change the word "lunge" to "pulmonal" so the term were changed to "pulmonal contusion" and thereby I was able to map the term to a ID code. Moreover, I did not consider the order of a term, for example, the unique term was "*pain in the head*" I mapped the term as "headache" as they have a similar meaning, or the

word could be “biopsy tunge” and the term were mapped to “tongue biopsy”. Other terms may consist of two words, but only one of them was mapped into a MedDRA ID code. As an example, it could be “*increased depression*”, but only “*depression*” was mapped into an ID number.
Due to exposure of different terms some terms were more time consuming than others. Simple words such as “pain, fatigue, vomiting” were found easily and took only minutes to be mapped, however more complex words suchlike the above-mentioned “kontusion lunge” were more time consuming and took around five minutes to be mapped. An average measure of time was between 5-10 minutes for each term. An illustration of how the extracted data was obtained are show below in figure 8.

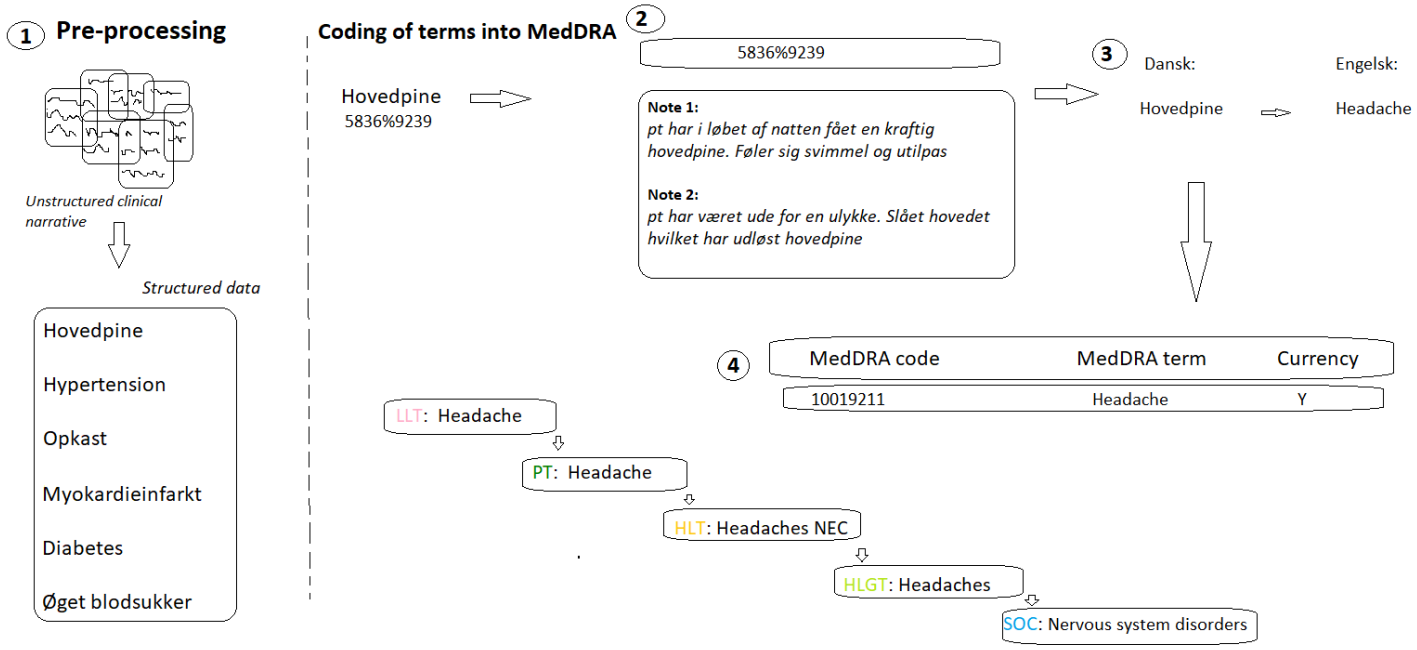

Figure 8: Demonstration of coding of the extracted terms into MedDRA. In the pre-processing phase, unstructured clinical narratives from Steno Diabetes Center undergoes a computational text mining by the use of natural language processing (NLP) algorithm, based on an rule and terminology NER system, which aims to understand and structure the data (step 1). The structured data are stored in a file, and each term is defined with codes. The terms are validated by looking up each term based on their codes in a software program that retrieve 15 random clinical text (step 2). Terms that are validated to be a potential adverse event (AE) are translated from Danish to English (step 3). The English term are mapped into MedDRA´s lowest level hierarchy (step 4).

5.4 Extracted drugs

All prescribed drugs were stored in the EPRs. After converting the unstructured EPRs to a structured dataset I was able to extract the drugs based on our rule-based approach. All drugs with the same active substance were considered as a single drug. The Anatomical Therapeutic Chemical (ATC) were used to group drugs into classes. The most common drugs that were collected, were obtained based on a data driven approach, were we looked into the numbers of patients receiving

the medication. Only drugs prescribed to more than 1000 patients were considered. Further, I examined counts of ADRs associated with each drug, by comparing the number of unique terms that were discovered in the EPRs with the total number of patients receiving the drugs. Thus, these were further calculated in percentages.

6 Results

Integration of a rule-based NER-tagger text mining in cooperation with mapping gave us the ability to extract unique terms and thereby detect the most common drugs prescribed in our population. Moreover, the number of potential ADRs associated with the drugs used was also identified.

6.1 Mapping

Out of 12,073 patient records, we identified 3,830 unique terms based on the in-house NER dictionary in combination med MedDRA that could be a potential ADE, all of which were mapped into MedDRA ID numbers, which can be used in further analysis (See appendix). We stopped at this number due to lack of time and could most likely find several other unique terms as the data set we worked with were huge. It was possible to use MedDRA as my dictionary and translate the extracted terms into ID-numbers (see appendix). Moreover, it was evidence that terms such as "nausea, vomiting, pain" were easier to map into ID codes compare to more complex and multiple terms, which were also more time consuming, such as "increased depression" as the term "increased" cannot be translated and thereby you have to validate if the translation is acceptable. Adding an additional word were also possible and gave a more meaningful translation. However, the validation based on the 15 random EPRs were important to detect any possible computational errors, which I find in-between my terms.

6.2 Frequency of terms:

Based on the rule-based approach in combination with NLP algorithmic it was possible to obtain the phenotypic profiles from the EPRs. The most common terms that occur are estimated, and it is notable to mention that the same term may occur multiple times in the same patient. Terms that had a frequency from 1000 and above that number were taking into account to be able to give an overview of the most common phenotypic description and thereby minimize the amount of terms. The phenotypic extraction describes the characteristic features of the population that is present in this study. As pictured in table 2, the phenotypic terms occurs in different level of granularity. The

most frequently term “glycosylated haemoglobin decreased” is much more detailed than the second frequent term “pain”. However, as shown below, there is a good mixture of both descriptors.

Frequency of term	MedDRA ID code	Term
42501	10018482	Glycosylated haemoglobin decreased
29654	10033371	Pain
26417	10029331	Neuropathy peripheral
24802	10020993	Hypoglycaemia
22742	10052428	Wound
15167	10054805	Macroangiopathy
15148	10020772	Hypertension
14618	10016256	Fatigue
13366	10049803	Blood glucose fluctuation
8332	10038923	Retinopathy
7543	10021789	Infection
6960	10042209	Stress
6171	10047895	Weight decreased
6078	10042674	Swelling
5337	10005614	Blood insulin increased
4765	10015150	Erythema
4208	10012378	Depression
4127	10013082	Discomfort
4120	10033775	Paraesthesia
4032	10057430	Retinal injury
3804	10061666	Autonomic neuropathy
3660	10022562	Intermittent claudication
3584	10033425	Pain in extremity
3509	10047700	Vomiting
3466	10062315	Lipohypertrophy
3193	10046571	Urinary tract infection
3163	10008190	Cerebrovascular accident
2957	10002383	Angina pectoris
2951	10011224	Cough
2937	10031161	Osteoarthritis
2866	10020649	Hyperkeratosis
2732	10005734	Blood pressure decreased
2725	10061428	Decreased appetite
2629	10010774	Constipation
2574	10023379	Ketoacidosis
2500	10020635	Hyperglycaemia
2349	10055798	Haemorrhage
2326	10002855	Anxiety
2255	10000059	Abdominal discomfort
2184	10005191	Blister
2152	10062237	Renal impairment
2082	10057593	Blood ketone body
1994	10008479	Chest pain
1987	10037844	Rash
1980	10037660	Pyrexia
1875	10021518	Impaired gastric emptying
1823	10028851	necrosis
1697	10043607	Thrombosis
1689	10022000	influenza
1653	10018884	Haemoglobin decreased
1639	10011953	Decreased activity
1585	10000081	Abdominal pain
1580	10020642	Hyperhidrosis
1530	10028411	Myalgia
1509	10020937	Hypoesthesia
1450	10014062	Eating disorder
1438	10052340	Decreased insulin requirement
1399	10007649	Cardiovascular disorder
1391	10039906	Seizure
1391	10028997	Neoplasm malignant
1323	10017944	Gastrointestinal disorder
1292	10037087	Pruritus
1276	10044565	Tremor
1261	10003553	Asthma
1259	10007586	Cardiac murmur
1224	10061224	Limb discomfort
1213	10027175	Memory impairment
1201	10061284	Mental disorder
1194	10012374	Depressed mood
1105	10017577	Gait disturbance
1088	10049848	Balance disorder
1088	10022489	Insulin resistance
1069	10029202	Nervous system disorder
1061	10040026	Sensory disturbance
1057	10053156	Musculoskeletal discomfort
1050	10034620	Peripheral sensory neuropathy
1050	10017076	Fracture
1049	10077753	Frustration tolerance decreased
1015	10011781	Cystitis
1013	10050296	Intervertebral disc protrusion
1004	10074300	Therapy change

Table 2: The most frequently terms in the electronic patient records (EPRs).

6.3 Extraction of common drugs

Based on our data driven approach I was able to extract the most frequently used drugs among the population. The 12 most common drugs are illustrated below, with insulin being the highest number of patients that receives the drug (Figure 9), which is also expected as I was working with a diabetic population that requires insulin to reduce the symptoms. Moreover, I extracted several cardiovascular drugs suchlike simvastatin as the second common drug, and thereby aspirin, metformin and enalapril (Figure 9). It is likely that patients with diabetes are more disposed to adverse reactions due to increased risk of heart conditions, which leads to the need of multiple medications. Simvastatin is often prescribed to diabetic patients as statins attend to lower glycemic control and therefor the estimation of simvastatin being the second common drug would be expected. The remaining drugs of the list included bendroflumethiazide, amlodipine, furosemide, potassium chloride, glimepiride, metoprolol and irbesartan, which is also demonstrated in figure 9 below.

Drug	Function
Insulin	Increases insulin level in the bloodstream
Simvastatin	Lower cholesterol level Used in prevention of cardiovascular diseases
Aspirin	Analgesic Blood thinning
Metformin	Perioral antidiabetic Lower the production of glucose in the liver
Enalapril	Ace-inhibitor Used for hypertension and heart failure
Bendroflumethiazide	Thiazide diuretic Used for hypertension and edema
Amlodipine	Calcium channel blocker Used for hypertension and angina pectoris
Furosemide	Loop diuretic Used for edema
Potassium chloride	Used for calcium deficiency
Glimepiride	Decrease blood sugar level
Metoprolol	Adrenergic receptor antagonist Used for abnormal heart rhythm
Irbesartan	Angiotensin II receptor antagonist Primary used for hypertension

Table 3: Demonstration of the extracted drugs and their functions is shown.

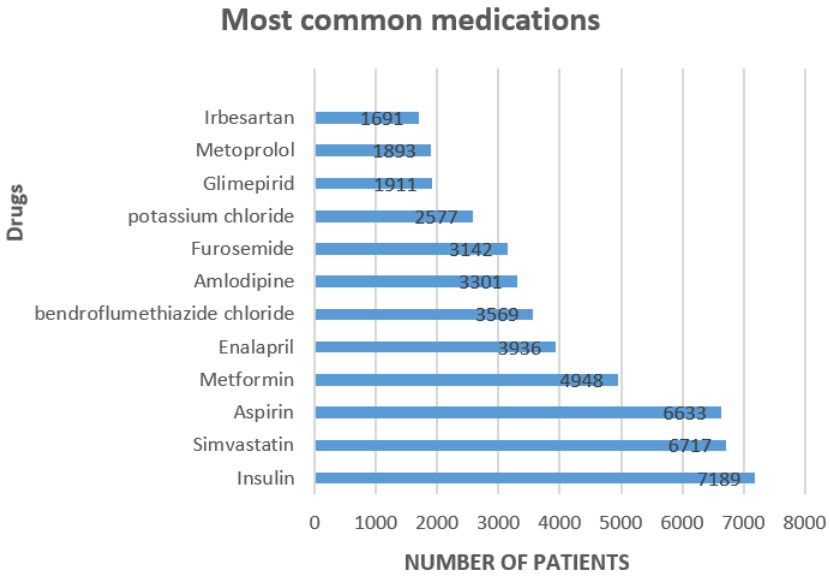

Figure 9: An overview of the extracted drugs are shown. The twelve most common drugs patients received and the number of patients are illustrated. Insulin shows to be the most common prescribed drug with a total number of 7189 patients receives, which will be expected. Subsequently several cardiovascular drugs are followed by.

6.4 Adverse drug reactions

All enrolled 12,073 patients were analyzed for potential ADRs. We identified the medications with the most associated potential ADRs. Medications that caused potential ADRs in 100 patients and above that number were analyzed and shown in figure 10. Included medications are pictured below with number of unique potential ADRs that occur in each patient group. Based on the extracted numbers of potential ADR I identified six medications that were above 200. Out of 12,073 EPRs, insulin caused potential ADRs in 354 patients. Simvastatin was the second most common drug patients received with 293 patients experiencing a potential ADR. Aspirin was close up to simvastatin with 282 patients. Other drugs included were metformin with 275 patients experiencing potential ADR, bendroflumethiazide with 211 patients and lastly amlodipine with 208 patients experiencing potential ADR. Drugs that caused potential ADR in patients below 200 patients were furosemide with 187, potassium chloride with 138, irbesartan with 113, metoprolol with 107 and glimepiride with 103 patients. These findings revealed that each drug was linked to a number of potential ADRs. However, some drugs are shown to be more prone to cause ADRs compared to others. Thus, it is worth mentioned that a potential ADR that occurs due to a specific drug may not be caused by the actual drug the patient received, but rather due to DDI or inappropriate prescription.

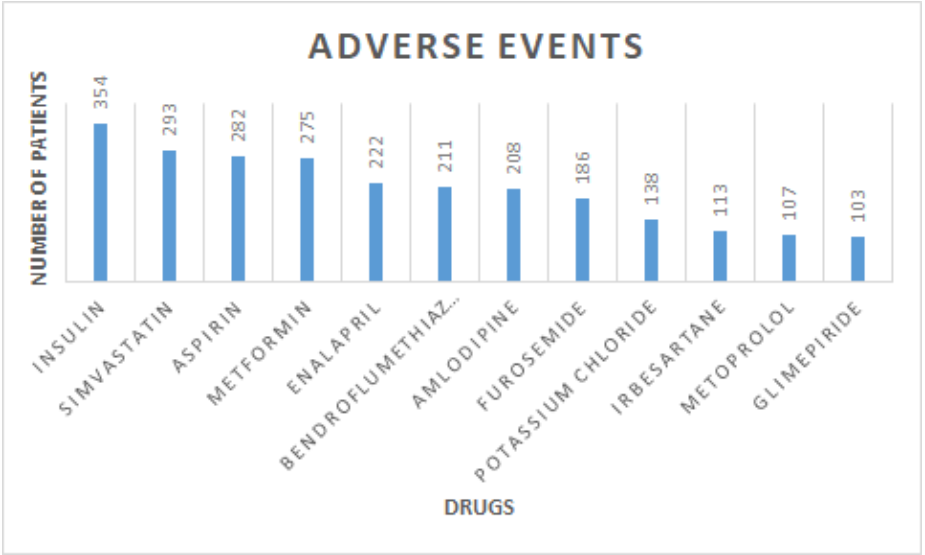

Figure 10: An illustration of the number of potential adverse drug reactions (ADRs) that occurs within patients depending on what kind of drugs they receive are shown.

Considering the findings above (Figure 10), I calculated a percentage of each drug and their potential ADRs that occurred in patients based on drugs they received. I compared all drugs within each other. As our patient population consisted of patients with diabetes, having insulin among the most frequently used drugs was expected however potential ADRs that occurred in patients were not among the highest compared to how many patients that received the drug. Out of 7189 patients that received insulin only 4.9% of the total number experienced a potential ADRs. This also counts for simvastatin and aspirin. Although simvastatin are giving to most patients, and is the second common prescribed drug it was found to be one of the lowest drugs that were associated to potential ADRs with 4.3% out of 6717 patients. Aspirin however were also estimated to be one of the drugs with lowest potential ADRs with 4.2%. Thus, irbesartan had the lowest number of patients that received the drug with 1691 patients, but showed the highest percentages of potential ADRs with 6.6% among the patients, compared to the other extracted drugs. The second highest percentages were followed by amlodipine that were calculated to 6.3% out of 3301 patients. Moreover, some drugs had similar percentages such as bendroflumethiazide and furosemide, both calculated to 5.9%.

MOST COMMON DRUGS	PERCENTAGES OF EACH DRUG
INSULIN	4.9%
SIMVASTATIN	4.3%
ASPIRIN	4.2%
METFORMIN	5.5%
ENALAPRILE	5.6%
BENDROFLUMETHIAZIDE CHLORIDE	5.9%
AMLODIPINE	6.3%
FUROSEMIDE	5.9%
POSSTAISUM CHLORIDE	5.3%
IRBESARTANE	6.6%
METOPROLOL	5.6%
GLIMEPLINIDE	5.4%

Table 4: A percentage of potential adverse drug reactions (ADRs) that occurred compared to the number of patients receiving the drug were calculated in each drug to measure the risk linked to each drug.

7 Discussion

7.1 Text mining

The goal of enabling text mining in the present study based on a rule-based NER approach showed a significantly positive outcome with 3,830 unique terms in the 12,073 patient records. These findings may lead to detection of unknown AEs that has been unnoticed in the patient records. Moreover, I was able to translate extracted potential ADEs into MedDRA to achieve an international comparable dataset (see appendix). Since MedDRA does not contain a translation of the terms in Danish, the in-house tagger NER dictionary in combination with MedDRA were applied in this cohort to perform a high quality of text mining and this found to be a significantly great tool to use to obtain a comparable dataset across different countries. Subsequently, the phenotypic descriptors could be extracted both terms with a high granularity and terms that were broader. Further, the most common prescribed drugs could be extracted and analyzed (Figure 9). Twelve medications were detected, and all of them were associated to potential AEs (Figure 10). These results indicate that computational automatic text mining is a great tool to extract possible ADEs on a large set of data that can be used in research to find unknown facts or information in a certain area of interest. Additionally, the in-house dictionary in combination with MedDRA has also proven to be feasible to use, as a dictionary on Danish is not yet available. However, I did not look into if these detected potential ADEs were associated to the drug or if it was based on coincidence; this can be done in further investments and is a limitation of the study. The approach demonstrated in this study, have opened doors for executing non-hypothetical data driven approach, and thereby be able to extract valuable information regarding of which corpus you deal with.

In addition to the present study, several other studies have analyzed how to extract unique events by the use of different informatics methods to detect possible ADE. A study conducted by Murff et al. (2003) examined AE in discharges summaries by trigger words and found trigger words in 251 discharges summaries out of 424 (Urff et al., 2003). Wang et al. (2010) also used discharges summaries, and found ADE by the use of MedLEE (Wang, Chase, Markatou, Hripcsak, & Friedman, 2010). A third study conducted by Visweswaran et al. (2003) examined possible ADE in discharge summaries by the use of Bayers models (Visweswaran, Hanbury, Saul, & Cooper, 2003). Honigman et al. (2011) used computational data programs to identify ADE and found a highly sensitive towards ADE (Rothschild et al., 2011). Another different system was used in Sohn et al. (2011) study where the authors investigated ADE by clinical text Analysis and Knowledge

Extraction System (cTAKES), which showed that this system had the ability to identify ADE (Sohn, Kocher, Chute, & Savova, 2011). Kilbridge et al. (2009) used rule-based computer programs (Kilbridge et al., 2009). Chen et al. (2007) detected ADE by rule mining (Chen, Pedersen, Chu, & Olsen, 2007). Lastly, Melton and Hripcsak (2005) investigated ADE through NLP(Melton & Hripcsak, 2005). All the above-mentioned informatics techniques showed detectable outcomes, meaning these methods are of great influence in safety trials.

7.2 Advantages and disadvantages of text mining

It is clear that the use of computational text mining is an innovative tool that demonstrates excellent outcomes in clinical settings. Researchers deals with large databases, which consumes a lot of time if each document has to be reading in contrast to text mining. Text mining has the advantage of having a fast timeframe, lower cost and at the same time the ability to work with a large data set which goal is to find possible ADE. Traditional ADE detection was achieved through manual chart review, thus this method compared to text mining needs a great deal of time and effort.
However, we cannot ignore the fact that this method also suffers from some disadvantages. In text mining, the method is not critical to what we as investigators order it to do. The human language is complex and peculiar. Each individual person speaks differently in their own way both verbally and, in writing skills. Besides that, misspelling and abbreviations are common in clinical notes. It would be beneficial to have an automated program that collects all slang and abbreviation words into a more standardized panel, so it can help researcher or other physicians. Additionally, it may give rise to an effective panel if the majority of similar words or abbreviations are translated into one word. Human natural languages are a communication forum for humans to interact with one another and computational software is far from understanding this field without advanced informatics techniques. Another limitation is that the method has it easier to find more general terms in our databases as mentioned in the method section compared to more complex or specific terms. This can be an issue as majority of clinical interest is complex areas researchers would like to look further into. This indicates that this area has to improve continuously with increase knowledge through the years (Tan & others, 1999)

7.3 Mapping

As mentioned before using MedDRA in combination with the in-house dictionary gave us the ability to translate extracted unique terms from our EPRs into a standardized international comparable dataset. Despite the fact that the in-house NER were established for a Danish dictionary, it is thought to be able to be used in other Scandinavian languages due to high similarity among the languages and thereby the ability to share findings in-between them. MedDRA is an autonomous dictionary compare to prior dictionaries. Before MedDRA, no dictionary on an internationally level were accessible for medical terminologies. Previously, in Europe researcher utilized two different dictionary such as WHO-ART along with ICD-9. Furthermore, in the US COSTART were used along with ICD-9. Lastly, the Japanese made their own dictionary under the name Japanese Adverse Reaction Terminology (J-ART) and Medical Information System (MEDIS) (SAS, 2009). The use of multiple dictionaries gives rise to several challenges in any stage of drug development as data retrieval as well as data analysis become more complicated. Moreover, use of distinct terminologies across different countries makes is difficult to compare and share new findings on an international level. In addition, it would be necessary to translate from one dictionary to another. Thus, these challenges not only affects researcher across different regions, but also between companies and clinical research organizations. As a result of these challenges it was evidence that a standardized electronic communication forum was required to be able to share and retrieve new findings across different countries. Therefor ICH developed MedDRA, which main advantage is to map extracted unique terms from clinical notes and create a standardized forum between researchers internationally as well as intern, which makes it possible to share findings that includes detection of risks induced by DDI or other important findings that is associated with patient safety and thereby share beneficial control strategies. Other advantages comprise of, transferring computational data, the capability to achieve a more specific translation by the use of LLTs, which endorses a more accuracy statement, a structure that permits compliance in both investments and data retrieval. Lastly, it has an enlargement in terms compared to WHO-ART and COSTART, it is easy to use and cheap. WHO-ART and COSTART only consist of three level, which makes MedDRA more attractive due to its hierarchy structure, which consist of five levels. Moreover, the prior dictionaries are no longer maintained, in contrast to MedDRA that is updated twice a year. These advantages benefits researchers as it enhance the quality, timeliness and availability of data for analysis (SAS, 2009).

However, although MedDRA is a more specific dictionary it is somehow followed by limitation. MedDRA does not cover all medical topics or safety issues (Brown, 2004). In addition, related terms may be mapped into two different SOC dependent on how the term is described. For instance, the term “hepatic function abnormal” is related to “hepatobiliary disorders” while “liver function test abnormal” is related to “investigations”, which demonstrates that two identical terms can occur in two different SOC (Schroll et al., 2012). Furthermore, the complexity of PTs is challenge due to the ability to be in more than one SOC, which I also saw (E.G., L., & S., 1999). This increases the risk of incorrectly interpretation of the data, which is supported by Toneatti et al (2006) were they based on a pilot test found differences in coding that were carried out by two different coders. At the PT level, 12% of the cases were coded differently, however at the broader levels only 5% of the terms were coded differently (Tonéatti et al., 2006). Besides that, the relationship between LLTs and PTs is quite paradoxical because although terms at the LLT level are characterized as the most specific translation with various terms to the same concept, every LLTs is a subordinate term to a PT with the same medical term and with identical ID codes, which means the term appears in both levels. This could challenge the hierarchical structure of MedDRA, as it is described as a fifth level structure, but may actually only consist of four. If we look deeper into the relationship between LLTs and PTs, it is shown that a PT can be related to a LLT by being identical, which means the term is the same in both levels. Another way a PT can be related to LLT is by lexical variants such as different order of the term or abbreviation of a term. In either case distinction of which of the two terms from the levels are more general than the other cannot be made. In addition, PTs can be related to LLTs by synonyms, meaning the terms are two different terms, but with similar meaning, such as “arthritis” and “joint inflammation” leading to a discussion of which of the terms are more specific than the other. Hence, to maintain the characterization of LLTs being more specific and a subordinate term to PT, and subsequently PTs being more general, a clear distinction of the levels has to be done (Merrill, 2008).

7.4 Challenges in my mapping:

Mapping extracted terms gave rise to some challenges in this study. Since I worked with Danish EPRs and MedDRA has not been translated into Danish, I had to translate every extracted term into English. Many terms were easily translated and mapped into ID-codes, some terms however were more complex as translating the term were not in compliance with the term on English , such as "graderet" or "opkørt" as I mentioned in the method section. When translating these words to English, it gave me a distinct word compared to the Danish term. This may be due to the term not being accessible in the English language. Moreover, simple words were much easier to map such as "breast Cancer, nausea, pain" compared to more complex words that consist of two words. This meant that I only had the ability to map one of the words and thereby not able to map both words, which I also did in many cases.

As mentioned above, multiple terms with similar meaning can occur, which could be confusing when mapping a term, as an example at the LLT level I mapped "prostate cancer" as "prostatic cancer" however, at the PT level the term could also be mapped to "prostate cancer". It is exactly the same word with dissimilarity in the spelling but in two different levels or the term can also be mapped to "malign neoplasm of prostate". However, for the reason that the granularity of the terms in MedDRA is much more expanded compared to prior dictionaries, it gave me the ability to choose between synonyms and made it easier for me to map the extracted terms. The granularity is a result from the update MedDRA goes through each year. It is beneficial to be updated and aware of new findings, thus the challenging part is, the more terms that are added to the dictionary, the greater chance is there to add terms with similar meaning and thereby a greater risk to categorize terms with same content in different groups. The increasingly amount of terms has also been investigated by Brown et al. (2004) were the study demonstrated that 315 terms in WHO-ART could be mapped to 943 terms in MedDRA (Brown, 2004). Moreover, Toneatti et al. found that updating MedDRA modifies terms in different levels. Based on 436 LLTs 38 of them changed their PTs or SOC (Tonéatti et al., 2006).

Another limitation of the study were the number of participants, compared to the number of prescribed drugs in the EPRs. The number of individuals that is prescribed a drug in some cases merely reaches the amount of subjects in a phase III clinical trials. This were also the case in this study, however despite that we were able to observe a positive result.

7. 5 Phenotypic extraction

Medical technological in phenotypic research has been of great interest to get a better understanding of the genotype-phenotype relationship in diseases. Studies have shown an excellent correlation between phenotypes and genomic functions (Lussier & Liu, 2007). The present study were able to extract the phenotypic profiles that occurs in different level of specificity, each term were estimated and characterized by MedDRA ID codes. The extracted terms that is less detailed only describes an event, but not were the event occurs. This lack of information may affect the interpretation of the geno-type phenotype relation. Unfortunately, methods to investigate the relationship between genes and phenotypes are limited, and besides that, examining phenotypic data is much more complex compared to biological data. EPRs contain important information that can enhance clinical trials and to have a tool that extract accurate phenotype profiles is necessary. This lead to a new area called phonemics, which aims to take advantages of computational high throughout technology as well as informatics techniques. Recently, a study combined DNA biorepositories to EPR data by the use of NLP algorithm and were able to detect genomics based on this approach. Over the past decades 20 article have been approved to be announced, were they value phenotypic algorithms (Pathak, Kho, & Denny, 2013). The increase number of EPRs and the capability to combine several computational algorithm across different databases builds powerful data to phenotype algorithms. However, this field need further investigation (Lussier & Liu, 2007).

7.6 The use of text mining in drug examination

DDI is becoming one of the most important clinical areas due to an exponentially use of polypharmacy in patients, which leads to higher risk of treatment complications. Considering the study population, diabetic patients are in high risk of DDI due to associated complications such as cardiovascular events, meaning these patients may show possible ADEs. A regular patient diagnosed with diabetes with associated cardiac insufficiency includes drugs to lowering hyperglycemia and four cardiac drugs consisting of digoxin, a loop diuretic, an ACE inhibitor and a beta-blocker. In such case, it increases the risk of drug interactions, but also makes it challenging to distinguish between the drugs and, which of them may cause the ADE, if any occurs (Hospital, Diego, Israel, & Medical, 1998).

However, harmful ADE is difficult to discover in large datasets. An estimation can be if one patient experiences an ADE outside from 1,000 patients, at least five patients will experience such an ADE out of 5,000. In many patient models to investigate ADEs, such group consist of 2,500 patients meaning only one or two patients will experience an ADE and that makes is challenging. However,

it does not lower the importance of it (Cerrito, 2001). In addition, most of them are conducted in healthy participants, as subgroup in most cases are in the list of exclusions criteria. This limits a study because subgroups normally show a higher sensitivity towards drugs. Simultaneously, AEs occurs in subgroups due to their comorbidity. As an example, the use of statins lead to lowering glycemic control and usage of angiotensin-converting enzyme (ACE) inhibitors give rise to hypoglycemia (Cerrito, 2001). Another problem in safety trails is the duration of the trial. In most cases, clinical trials are frequently short term with no follow-up period, which makes it difficult to detect adverse reactions if the event happens after the end of study. Hence, long-term studies are needed and adverse reactions may have a greater risk of being detected over a long period compared to a short-term period. However, the duration of the study can be discussed, a study investigated adverse reactions over 7 years and reported 1 drug that causes an AE, which could indicate that seven years may not be long enough. However, it is notable that not all drug combinations can be investigated in safety trials (Cerrito, 2001) (Tan & others, 1999). Additionally, another challenging aspect can be the dissimilarity in-between patient. Dependent on the patients' different dosages and drug combinations are giving leading to misinterpretation when performing a safety trial as a conclusion cannot be made based on the statistical analyze.

Despite that, large databases can still contain unnoticed ADE. The largest international database developed by the WHO contains almost 2 million reported ADE. These reports are from health care professionals. Bate and co (1998) established a study where they investigated ADEs based on WHO database. Bayesian network were used to analyze the data due to this methods robustness and ability to exhibit great power. They found 12 new undiscovered drugs that caused serious ADE denoting that a computational text mining, and in this case Bayesian network method are great when we want to demonstrate possible ADEs between drugs (Bate et al., 1998).

Every approved drug by the FDA is undergoing post surveillance monitoring due to any unnoticed reactions. If the approved drug exhibit negative response in form of harmful ADE such events are reported to the FDA, which will assess the case and in some cases withdrawal the drug from the market. Through the past years, there have been cases where drugs are being withdrawal. These drugs include rofecoxib as patients receiving the drugs showed a higher risk of developing cardiovascular events (Sibbald, 2004). Moreover, back in 2006 Mangano et al. (2006) performed a study was they found a strong association between aprotinin and death rate, leading to removal of the drug from the market (Mangano, Tudor, & Dietzel, 2006). In addition, rosiglitazone also

demonstrated increase prevalence in cardiac events leading to withdrawal (Härmark & Van Grootheest, 2008)

7.7 Future studies

Based on the results from this study, further investigation can be done. Numerous studies have had their focus on different kinds of cardiovascular drugs individually and with combination with other drugs. A new innovative method could be to look into frequencies of cardiovascular drugs and their combinations. It is evident that some drugs either alone or in combination with another drug will cause a common ADE such as headache; however, the interesting part would be to look at the frequency when two drugs are combined. Would the two-drug combination reduce the ADE, increase the ADE or be constant? Lastly, detailed phenotypes can be investigated by the extracted terms from EPR to see if these are associated with polypharmacy profiles and how many times a specific AE occurs. Investigating phenotypes in a population is valuable to get an insight in which kind of population we are dealing with, weather it is a high-risk group or a low-risk group (Leopold & Loscalzo, 2018).

The extracted outcomes will be compared by the use of one of the standard analytical tools in clinical research. The relative risk (RR) compare the risk of outcome in-between the groups. When mentioning RR, either it refers to risk ratio or odds ratio. If the frequency of the result is below <10%, it means odds ratio is close to risk ratio and verse versa. By using RR we can compare all cardiovascular drugs with one another and estimate the difference, if any occurs (J. & K.F., 1998).

8 Conclusion

Automatic computational text mining is an innovative tool for extracting relevant information in future research. In this study, I was able to extract unique terms that could be a potential ADRs from the EPRs based on a rule-based NER method. Moreover, I was able to map 3,830 unique terms into MedDRA codes, based on the in-house tagger NER, to achieve an internationally comparable dataset. Simultaneously, extraction of the most common drug is possible based on the mappings. We consider that the method used in this study in combination with the dictionaries can give researcher a more sophisticated analysis of clinical results suchlike the relationship between ADEs and prescribed drugs on a Scandinavian as well as on an international level. Furthermore, it is well known that EPRs may contain invaluable information and unnoticed ADEs, which can be used in clinical trials to improve health.

9 References

Atkinson, M. A., Eisenbarth, G. S., & Michels, A. W. (2014). Type 1 diabetes. *The Lancet*, *383*(9911), 69–82. https://doi.org/10.1016/S0140-6736(13)60591-7

Bate, A., Lindquist, M., Edwards, I. R., Olsson, S., Orre, R., Lansner, A., & De Freitas, R. M. (1998). A Bayesian neural network method for adverse drug reaction signal generation. *European Journal of Clinical Pharmacology*, *54*(4), 315–321. https://doi.org/10.1007/s002280050466

Beers, M. H. (2013). Potential Adverse Drug Interactions in the Emergency Room. *Annals of Internal Medicine*, *112*(1), 61. https://doi.org/10.7326/0003-4819-112-1-61

Berkeley, U. C. (2003). *Retrieved October 2003 Hearst*. 1–3. Retrieved from papers3://publication/uuid/34051528-5F83-4A7E-BEA9-4197678FA532

Brøsen, K., Simonsen, U., P.kampmann, J., & Thirstrup, S. (2014). *Basal og Klinisk farmakologi*.

Brown, E. G. (2004). Using MedDRA: Implications for risk management. *Drug Safety*, *27*(8), 591–602. https://doi.org/10.2165/00002018-200427080-00010

Bushardt, R. L., Massey, E. B., Simpson, T. W., Ariail, J. C., & Simpson, K. N. (2008). Polypharmacy: misleading, but manageable. *Clinical Interventions in Aging*, *3*(2), 383–389. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/18686760%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2546482

Canivell, S., & Gomis, R. (2014). Diagnosis and classification of autoimmune diabetes mellitus. *Autoimmunity Reviews*, *13*(4–5), 403–407. https://doi.org/10.1016/j.autrev.2014.01.020

Cerrito, P. (2001). Application of data mining for examining polypharmacy and adverse effects in cardiology patients. *Cardiovascular Toxicology*, *1*(3), 177–179. https://doi.org/10.1385/CT:1:3:177

Chatterjee, S., Khunti, K., & Davies, M. J. (2017). Type 2 diabetes. *The Lancet*, *389*(10085), 2239–2251. https://doi.org/10.1016/S0140-6736(17)30058-2

Chen, Y., Pedersen, L. H., Chu, W. W., & Olsen, J. (2007). Drug exposure side effects from mining pregnancy data. *ACM SIGKDD Explorations Newsletter*, *9*(1), 22. https://doi.org/10.1145/1294301.1294308

College, B., & Doddi, K. M. (2016). *Multiple Diseases and Polypharmacy in the Elderly Cardiovascular Disease Patients : Challenges for the Internist With a Need for an*. *6*(10).

Cytochromes_P450_2002.pdf. (n.d.).

Demiology, E. P. I. (1998). *Frequency of Hospitalization after Exposure to Known*.

Drug Metabolism. (2013). *Encyclopedia of Systems Biology*, pp. 618–618. https://doi.org/10.1007/978-1-4419-9863-7_100411

E.G., B., L., W., & S., W. (1999). The Medical Dictionary for Regulatory Activities (MedDRA). *Drug Safety*, *20*(2), 109–117. Retrieved from http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed4&NEWS=N&AN=1999078863

Edwards, I. R., & Aronson, J. K. (2000). Adverse drug reactions Adverse drug reactions : definitions , diagnosis , and management. *The Lancet*, *356*, 1255–1259. https://doi.org/10.1016/S0140-6736(00)02799-9

Eftimov, T., Seljak, B. K., & Korošec, P. (2017). A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations. In *PLoS ONE* (Vol. 12).

https://doi.org/10.1371/journal.pone.0179488

Eriksson, R., Jensen, P. B., Frankild, S., Jensen, L. J., & Brunak, S. (2013). Dictionary construction and identification of possible adverse drug events in danish clinical narrative text. *Journal of the American Medical Informatics Association*, *20*(5), 947–953. https://doi.org/10.1136/amiajnl-2013-001708

Goldberg, R. M., Mabee, J., Chan, L., & Wong, S. (1996). Drug-drug and drug-disease interactions in the ED: Analysis of a high- risk population. *American Journal of Emergency Medicine*, *14*(5), 447–450. https://doi.org/10.1016/S0735-6757(96)90147-3

GR, W. (2005). Drug therapy. Drug metabolism and variability among patients in drug response. *New England Journal of Medicine*, *352*(21), 2211–2260. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&db=c8h&AN=106372327&lang=pt-br&site=ehost-live

Guo, S., & Cao, B. (2015). *Text Mining and Its Applications*. (August 2017). https://doi.org/10.2991/csic-15.2015.17

Gupta, V., & Lehal, G. S. (2009). A Survey of Text Mining Techniques and Applications - Volume 1, No. 1, August 2009 - JETWI. *Journal of Emerging Technologies in Web Intelligence*, *1*(1), 60–76. Retrieved from http://www.jetwi.us/index.php?m=content&c=index&a=show&catid=165&id=969

Härmark, L., & Van Grootheest, A. C. (2008). Pharmacovigilance: Methods, recent developments and future perspectives. *European Journal of Clinical Pharmacology*, *64*(8), 743–752. https://doi.org/10.1007/s00228-008-0475-9

Herr, R. D., Caravati, E. M., Tyler, L. S., Iorg, E., & Linscott, M. S. (1992). Prospective evaluation of adverse drug interactions in the emergency department. *Annals of Emergency Medicine*, *21*(11), 1331–1336. https://doi.org/10.1016/S0196-0644(05)81897-9

Hospital, H. F., Diego, S., Israel, B., & Medical, D. (1998). *The Ne w E n g l a nd Jo u r n a l o f Me d ic i ne A DOSE-DEPENDENT INCREASE IN MORTALITY WITH VESNARINONE AMONG PATIENTS WITH SEVERE HEART FAILURE*.

Hotho, A., Nürnberger, A., & Paaß, G. (2005). A Brief Survey of Text Mining. *Ldv Forum*, *20*(1), 19–62.

Intervention, M., & Disease, C. (2003). New England Journal. *Library*, 383–393.

J., Z., & K.F., Y. (1998). What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. *Journal of the American Medical Association*, *280*(19), 1690–1691.

Kadam, S., Mahadik, K., & Bothara, K. (2007). The challenge of managing drug interactions in elderly people. *Lancet*, *370*(9582), 185–191.

Karas, S. (1981). The potential for drug interactions. *Annals of Emergency Medicine*, *10*(12), 627–630. https://doi.org/10.1016/S0196-0644(81)80085-6

Kilbridge, P. M., Noirot, L. A., Reichley, R. M., Berchelmann, K. M., Schneider, C., Heard, K. M., … Bailey, T. C. (2009). Computerized Surveillance for Adverse Drug Events in a Pediatric Hospital. *Journal of the American Medical Informatics Association*, *16*(5), 607–612. https://doi.org/10.1197/jamia.M3167

Kojima, T., Akishita, M., Nakamura, T., Nomura, K., Ogawa, S., Iijima, K., … Ouchi, Y. (2011). Association of polypharmacy with fall risk among geriatric outpatients. *Geriatrics and Gerontology International*, *11*(4), 438–444. https://doi.org/10.1111/j.1447-0594.2011.00703.x

Laing, S. P., Swerdlow, A. J., Slater, S. D., Burden, A. C., Morris, A., Waugh, N. R., … Patterson, C. C. (2003). Mortality from heart disease in a cohort of 23,000 patients with insulin-treated diabetes.

Diabetologia, *46*(6), 760–765. https://doi.org/10.1007/s00125-003-1116-6

Leopold, J. A., & Loscalzo, J. (2018). Emerging Role of Precision Medicine in Cardiovascular Disease. *Circulation Research*, *122*(9), 1302–1315. https://doi.org/10.1161/CIRCRESAHA.117.310782

Levy, D. (2016). *Type 1 Diabetes*. *1*, 3–6. https://doi.org/10.1093/med/9780198766452.001.0001

Lussier, Y. A., & Liu, Y. (2007). Computational approaches to phenotyping: High-throughput phenomics. *Proceedings of the American Thoracic Society*, *4*(1), 18–25. https://doi.org/10.1513/pats.200607-142JG

Mangano, D. T., Tudor, I. C., & Dietzel, C. (2006). The Risk Associated with Aprotinin in Cardiac Surgery. *New England Journal of Medicine*, *354*(4), 353–365. https://doi.org/10.1056/nejmoa051379

Melton, G. B., & Hripcsak, G. (2005). Automated detection of adverse events using natural language processing of discharge summaries. *Journal of the American Medical Informatics Association*, *12*(4), 448–457. https://doi.org/10.1197/jamia.M1794

Merrill, G. H. (2008). The MedDRA paradox. *AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium*, 470–474.

Nielsen, O. F., & Bojsen-Møller, M. J. (2016). *Anatomi og fysiologi*.

Ozougwu, O. (2014). The pathogenesis and pathophysiology of type 1 and type 2 diabetes mellitus. *Journal of Physiology and Pathophysiology*, *4*(4), 46–57. https://doi.org/10.5897/jpap2013.0001

Pathak, J., Kho, A. N., & Denny, J. C. (2013). Electronic health records-driven phenotyping: Challenges, recent advances, and perspectives. *Journal of the American Medical Informatics Association*, *20*(E2). https://doi.org/10.1136/amiajnl-2013-002428

Rollason, V., & Vogt, N. (2003). Reduction of polypharmacy in the elderly: A systematic review of the role of the pharmacist. *Drugs and Aging*, *20*(11), 817–832. https://doi.org/10.2165/00002512-200320110-00003

Rothschild, J., Pulling, R. M., Honigman, B., Yu, T., Bates, D. W., Light, P., & Lee, J. (2011). Using Computerized Data to Identify Adverse Drug Events in Outpatients. *Journal of the American Medical Informatics Association*, *8*(3), 254–266. https://doi.org/10.1136/jamia.2001.0080254

SAS. (2009). *Introductory Guide JMP8*. (March), 1–20.

Schroll, J. B., Maund, E., & Gøtzsche, P. C. (2012). Challenges in coding adverse events in clinical trials: A systematic review. *PLoS ONE*, *7*(7). https://doi.org/10.1371/journal.pone.0041174

Services, H. (2014). *Polypharmacy and Medication Adherence in Patients With Type 2*. (February).

Shaalan, K. (2014). A Survey of Arabic Named Entity Recognition and Classification. *Computational Linguistics*, *40*(2), 469–510. https://doi.org/10.1162/COLI_a_00178

Shah, B. M., & Hajjar, E. R. (2012). Polypharmacy, Adverse Drug Reactions, and Geriatric Syndromes. *Clinics in Geriatric Medicine*, *28*(2), 173–186. https://doi.org/10.1016/j.cger.2012.01.002

Sibbald, B. (2004). Rofecoxib (Vioxx) voluntarily withdrawn from market. *CMAJ : Canadian Medical Association Journal = Journal de l'Association Medicale Canadienne*, *171*(9), 1027–1028. https://doi.org/10.1503/cmaj.1041606

Sohn, S., Kocher, J. P. A., Chute, C. G., & Savova, G. K. (2011). Drug side effect extraction from clinical narratives of psychiatry and psychology patients. *Journal of the American Medical Informatics Association*, *18*(SUPPL. 1), 144–149. https://doi.org/10.1136/amiajnl-2011-000351

Strehl, V. (2013). *Wavelet Transformationen in der Bildverarbeitung Script*. *13*(1), 1–11. https://doi.org/10.1517/14740338.2013.827660.Clinical

Tan, A.-H., & others. (1999). Text mining: The state of the art and the challenges. *Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*, *8*, 65–70.

Tatonetti, N. P., Denny, J. C., Murphy, S. N., Fernald, G. H., Krishnan, G., Castro, V., … Altman, R. B. (2011). Detecting drug interactions from adverse-event reports: Interaction between paroxetine and pravastatin increases blood glucose levels. *Clinical Pharmacology and Therapeutics*, *90*(1), 133–142. https://doi.org/10.1038/clpt.2011.83

Tonéatti, C., Saïdi, Y., Meiffrédy, V., Tangre, P., Harel, M., Eliette, V., … Pierre Aboulker, J. (2006). Experience using MedDRA for global events coding in HIV clinical trials. *Contemporary Clinical Trials*, *27*(1), 13–22. https://doi.org/10.1016/j.cct.2005.09.009

Urff, H. A. J. M., Orster, A. L. A. N. J. F., C, M. S., Eterson, J. O. S. H. F. P., Iskio, J. U. M. F., Eiman, H. E. L. H., & Ates, D. A. W. B. (2003). Electronically Screening Discharge Summaries for Adverse Medical Events. *Journal of the American Medical Informatics Association*, *10*(4), 12–14. https://doi.org/10.1197/jamia.M1201.Affiliations

Vik, S. A., Maxwell, C. J., & Hogan, D. B. (2004). Measurement, Correlates, and Health Outcomes of Medication Adherence among Seniors. *Annals of Pharmacotherapy*, *38*(2), 303–312. https://doi.org/10.1345/aph.1D252

Visweswaran, S., Hanbury, P., Saul, M., & Cooper, G. F. (2003). Detecting adverse drug events in discharge summaries using variations on the simple bayes model. *Proc AMIA Symp*, (2), 689–693.

Wang, X., Chase, H., Markatou, M., Hripcsak, G., & Friedman, C. (2010). Selecting information in electronic health records for knowledge acquisition. *Journal of Biomedical Informatics*, *43*(4), 595–601. https://doi.org/10.1016/j.jbi.2010.03.011

WHO. (2004). Estimates for the year 2000 and projections for 2030. *World Health*, *27*(5), 1047–1053.

WHO (2008). (n.d.). *Diabetes*.

Yoon, J.-W., & Jun, H.-S. (2005). Autoimmune Destruction of Pancreatic β Cells : American Journal of Therapeutics. *American Journal of Therapeutics*, *591*, 580–591.