Chapter 10
Applications of Clinical Text Mining

This chapter will present the state of the art for a number of applications of clinical text mining such as detection and prediction of healthcare associated infections (HAI), detection of adverse drug events (ADE), followed by resources for adverse drug event detection and continuing with an application of automatic assignment and validation of ICD-10 diagnosis codes. An application of automatic mapping of ICD-10 diagnosis codes to SNOMED CT will also be presented. This chapter then continues with automatic summarisation of patient record, simplification of patient records text and natural language generation of patient record text. Techniques for searching and retrieving patients from patient records for cohort studies are described, and finally some classic systems in medical decision support such as MYCIN are reviewed. Finally an overview of IBM Watson Health is provided.

10.1 Detection and Prediction of Healthcare Associated Infections (HAIs)

This section explains what a healthcare associated infection (HAI) is, why it is important to monitor and predict them and how to do it automatically using the information from the electronic patient records. The performance of systems for detecting HAIs will be compared. This section will also describe in what extent and where such surveillance systems for detecting HAIs are used in practice and commercially.

10.1.1 Healthcare Associated Infections (HAIs)

Healthcare associated infections (HAIs) are plaguing healthcare with suffering patients and heavy costs for society. Healthcare associated infections are also called *hospital associated infections* or *nosocomial infections*.

H. Dalianis, *Clinical Text Mining*, https://doi.org/10.1007/978-3-319-78503-5_10

HAI is defined in Ducel et al. (2002) as:

> An infection acquired in hospital by a patient who was admitted for a reason other than that infection (1). An infection occurring in a patient in a hospital or other health care facility in whom the infection was not present or incubating at the time of admission. This includes infections acquired in the hospital but appearing after discharge, and also occupational infections among staff of the facility (2).

It is estimated that 10% of all in-patients will obtain a healthcare associated infection while treated for another disorder (Humphreys and Smyth 2006). It is estimated that in Europe there are three million affected patients per year, of which about 50,000 die. The lethal outcome seems dramatic, but many of these patients had severe disorders, or multiple disorders, were old and had been treated at the hospital for a longer period of time and would have died of the disease, but the HAI hastened their death.

In Sweden with a population of 10 million the expected yearly cost for HAIs is estimated as 6.5 billion SEK or approximately 800 million USD. This cost corresponds to 750,000 extra days in hospital (SALAR 2014).

An important goal in defeating HAIs is to collect statistics by detecting and measure the prevalence of HAIs, but also to predict and warn if a particular patient has a high risk of obtaining HAI. HAIs can encompass, for example, pneumonia, urinary tract infection, sepsis or various wound infections but also norovirus (winter vomiting disease).

Part of the definition of HAI is that the patient must have been admitted to the hospital for more than 48 h before an infection can be defined as a HAI. The patient can also have been admitted and discharged then re-admitted to the same hospital (or to another hospital) within 24 h, as long as the patient entire healthcare period episode lasted more than 48 h the definition of HAI remains valid, see Fig. 10.1.

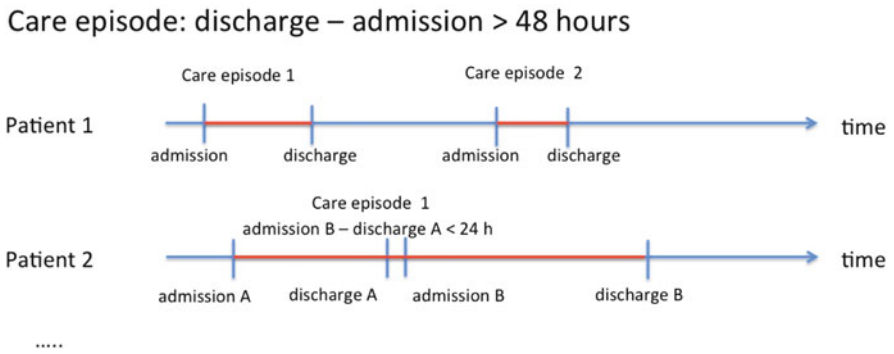

Fig. 10.1 If a patient (Patient 2) is discharged from one clinical unit and admitted to another within 24 h and the whole period is more than 48 h then that patient is considered to be admitted for the whole care episode and can therefore be analysed for HAIs. This figure is courtesy of Hideyuki Tanushi (© 2014 Springer International Publishing Switzerland—reprinted with permission. Published in Dalianis (2014))

To detect and predict HAI, the *healthcare episodes* of the patient record are used. Healthcare (or care) episodes are the daily notes and data that are entered into the patient record regarding the treatment and status of the patient. By observing the information in the care episodes a picture can be obtained whether the patient has a high risk of obtaining a HAI or not, or if the patient has obtained a HAI.

In Sweden a HAI must be reported by law, unfortunately, that is not always the case. Sometimes the reporting is not carried out because of ignorance and sometimes because physicians treating a patient for lethal diseases such as cancer or rheumatism are focusing on curing the patient and not on reporting the HAI. The physician believes that the HAI is just a small side effect or obstacle on the path to cure the patient of a much more serious disease.

One issue is how underreported are HAIs? To solve this problem, the National Board of Health and Welfare in Sweden (Socialstyrelsen) requires that all hospitals in Sweden report how many in-patients in each clinical unit have HAIs on one specific day. This measurement is carried out twice a year, once in the fall and once in the spring. This measurement is called the Point Prevalence Measurement (PPM). By performing the PPM the authorities can obtain an indication of how many patients have a HAI at each clinical unit, on each hospital, 24 h a day throughout Sweden; however, since the PPM is only carried out twice a year and does not give all the information needed, the hospital management would like to have continuous measurement with more measurement points every day and year round.

If specific measure are taken to avoid HAIs, for example, the use of new antibacterial catheters, special equipment in the ward, new cleaning routines for the ward, or single rooms for the patients, then it would be valuable to automatically measure the effect by using the daily nursing notes and doctor notes as input to a HAI monitoring system.

10.1.2 Detecting and Predicting HAI

An automatic HAI detection system analyses the text and the structured information in the patient record for specific terms that may point to a HAI. Such terms may be specific antibiotics for treating HAIs, particular microbiological tests for typical HAI bacteria as well as high body temperatures indicating infection. Risk prone patients are patients with catheters that are present in the body for a long time, patients that have been operated on and patients with open wounds along with generally older patients who stay on a ward for a long time.

There have been many approaches for detecting HAIs using only the structured information in the patient records, or the clinical free text, or both the structured information and the clinical text. Both rule-based methods and machine learning-based methods have been used. The best system obtained up to 97% precision and recall. Usually high recall is more important than high precision in detecting HAI. For a nice review of different HAI detecting systems or HAI surveillance systems see Freeman et al. (2013).

Table 10.1 Statistics for the Stockholm EPR Detect-HAI Corpus

	HAI	non-HAI
Number of records	128	85
Patient ages [years]	2–93	2–92
Total number of tokens	1,034,760	230,226
Time in hospital [days]	2–144	3–93
Total time in hospital [days]	3975	941

Tokens refer to space separated sequences of characters

Proux et al. (2011) describe a rule-based system for detection of HAIs in French patient records called *Assitant de Lutte Automatise et de Detection des Infections Nosocomiales a partir de Documents Textuels Hospitaliers (ALADIN-DTH)*, or simply ALADIN.

Ehrentraut et al. (2014) present a machine learning-based system called Detect-HAI for Swedish patient records using the *Stockholm EPR Detect-HAI Corpus*, see Table 10.1.

Two machine learning algorithms Support Vector Machine (SVM) and Random Forest (RF) in the Weka toolkit were applied on the annotated Stockholm EPR Detect-HAI Corpus. The corpus were preprocessed using nine different preprocessing steps were carried out including NegEx to detect negated expressions in the clinical text, however the preprocessing step that gave the best result was lemmatisation jointly with the Random Forest algorithm that gave 87% recall and 83% precision and an F-score of 0.85 (Ehrentraut et al. 2014).

The results using the same corpus were improved by utilising the Gradient Tree Boosting (GTB) algorithm in the Scikit-learn toolkit, obtaining 93.7% recall and 79.7% precision and F-score of 85.7 (Ehrentraut et al. 2016).

An approach using deep learning to detect HAI also studied the same data, the Stockholm EPR Detect-HAI Corpus, was carried out in (Jacobson and Dalianis 2016). The best results were obtained using the stacked restricted Boltzmann machines with a recall of 88% and with a precision of 79%. The results were comparable to those obtained in Ehrentraut et al. (2014); however, they were slightly lower than the results presented in Ehrentraut et al. (2016). Probably the annotated used data in the study was too sparse to make full use of the power of deep learning algorithms that require a large amount of annotated data.

Another general problem with the Stockholm EPR Detect-HAI Corpus presented in Table 10.1 is that the normal distribution of HAI and non-HAI data is 10 to 90 (around 10% of all patients have a HAI) while the distribution in the Stockholm EPR Detect-HAI Corpus is 60 to 40; hence, it does not reflect the real situation.

A rule-based approach using Swedish clinical text to detect urinary tract infections was investigated by Tanushi et al. (2014), both the data and text of the patient record were used. The data was divided in care episodes according to Fig. 10.2. (The same input data was used for the machine learning approaches of Ehrentraut et al. (2016) and Jacobson and Dalianis (2016)).

For the development of these systems, two development sets were used; one that had been used for the PPM containing 100 patient records and their corresponding

```
123 H - IVA 322916614D 2007-08-21 9:12 1944 Woman
Anamnesis
Got a urine catheter two days ago. Done a lab test
on the urine and gave antibiotics.
<ICD-10 code>
I110 Pneumonia.
I509 Heart failure, unspecified.

<Current medication>
Penomax

<Body temperature>
38
38
38.5

123 H - IVA 322916614D 2007-08-22 16:12 1944 Woman
<Body temperature>
37
36.8
36.9

<Blood culture>
pseudomonas
```

Fig. 10.2 An example of an electronic patient record text translated to English. The text is in a format prepared for processing by a computer program for detecting HAIs. The important features are extracted from the patient record and the program can check the status of a patient day by day. This particular patient got HAI (© 2014 Springer International Publishing Switzerland—reprinted with permission. Published in Dalianis (2014))

215 care episodes from different clinical units, and the other one containing 66 patient records from a rheumatology unit and their corresponding 134 care episodes. Both development sets were assessed for HAI and non-HAI by two physicians, one of them a specialist in infectious diseases.

For the evaluation 1195 patient records from oncological and surgical clinical units corresponding to 1867 care episodes with a duration ranging from 2 to 14 days were used. The resulting rule-based system obtained 98% precision and 60% recall. Observe that the evaluated results are from a different domain than the development of the system.

10.1.3 Commercial HAI Surveillance Systems and Systems in Practical Use

There are many different HAI surveillance systems in use worldwide, most of them use the structured information in the patient record, however, some of them use the clinical free text in addition to the structured data. Most systems are in-house developed systems and not commercial systems (Freeman et al. 2013).

In Sweden, the Anti-Infection tool, (Infektionsverktyget)[1] is used at all hospitals, it is integrated with each electronic patient record system. The Anti-Infection tool forces the prescribing physician at the hospital to state whether antibiotics are prescribed to a patient for prophylactic reasons, for an infection acquired outside the healthcare facility or for a healthcare associated infection. This information is collected nationwide and statistics are produced; however, it only tracks healthcare associated infections that are known and infections that can be treated by antibiotics.

One third of the hospitals in California, USA, use automated surveillance technology (AST) to monitor HAI. Some of the systems use information from the text in the patient record as well. The AST systems are a mix of in-house produced parts and commercially delivered parts (Halpin et al. 2011). Halpin et al. (2011) conclude that there is a strong and statistically significant association of the use of AST and evidence-based practice[2] for the prevention of HAI.

Monitoring of Nosocomial Infections in Intensive Care Units (MONI-ICU) is a surveillance system for electronic patient records written in German, using both text and data. MONI-ICU is used at the Vienna General Hospital in Austria as well as several other hospitals in Germany (Blacky et al. 2011). MONI-ICU has obtained a sensitivity (recall) of 90% and a specificity of 100% for intensive care unit patient records.

Testing of the French ALADIN system has been carried out in four French University hospitals (Lille, Lyon, Nice and Rouen), but has not yet implemented in practice (Proux et al. 2011; Metzger et al. 2012).

10.2 Detection of Adverse Drug Events (ADEs)

Adverse drug events (ADEs) are a major public health problem, around 5% of all hospital admissions in the world are due to ADEs (Beijer and de Blaey 2002). In Sweden, the seventh most common cause of death is an ADE (Wester et al. 2008). The domain of detection of adverse drug events (ADEs), is a complicated area. The relation between the properties of the body of a particular person, the disorder the person has and the pharmacological properties of the drug need to be understood. All drugs are poisonous in some sense but given in the correct amount they may cure a disease.

[1]The Anti-Infection Tool, http://www.inera.se/Documents/TJANSTER_PROJEKT/Infektionsverktyget/The_AntiInfection_tool.pdf. Accessed 2018-01-11.

[2]Evidence-based medicine is a concept within medicine meaning that the treatment of the patient should follow scientific results from well-designed research.

10.2.1 Adverse Drug Events (ADEs)

The science concerning drug safety is called pharmacovigilance or sometimes drug safety surveillance, and is related to the collection, detection, assessment, monitoring and prevention of adverse effects.

According to the World Health Organisation (WHO) the definition of an adverse drug reaction (ADR) is *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 and Aronson 2000).

An adverse effect can be in a range of severity from very mild to very strong, even lethal, and of various types. Usually an *adverse drug effect* is seen from the view of the drug while an *adverse reaction* is seen from the view of the patient.

There are six types of adverse effects (Edwards and Aronson 2000):

(a) *Dose-related*, for example giving toxic effect.
(b) *Non-dose related*, for example penicillin hypersensitivity.
(c) *Dose-related and time-related*, related to the cumulative dose.
(d) *Time-related*, becomes apparent some time after the use of the drug.
(e) *Withdrawal*, occurs after the withdrawal of the drug.
(f) *Unexpected*, often caused by drug interactions.

When the pharmaceutical industry is developing new drugs, at an early stage in the development the drug is tested on humans, usually on healthy young men to study how the human body is affected by the drug and also to detect possible adverse drug effects, while most patients taking a drug are older and suffering from some disorder and, of course, half of them are women. Drugs are usually not tested on women at this stage since women may become pregnant during the tests and the fetus may be affected by the drug.

In other words the sample size for the test is small, the duration of the test is short.

At a second stage the drug is tested on a small group of patients in a so-called *clinical trial* to test the effect of the drug before the drug is approved for the market.

During the clinical trial two test groups of patients are usually created, one is given the drug and the other group, the control group, is given a substitute which does not contain any active ingredients, a *placebo*.

Postmarketing surveillance (PMS) or post market surveillance is carried out to monitor the effect of drugs after they have been released to the market. PMS covers a larger sample over a longer period than a clinical trial.

An adverse drug effect or adverse effect is a general term for all types of unwanted effects of drugs. An adverse drug effect is hopefully mild and probably known by the physician, it is also related to the pharmacological properties of the drug. An adverse drug reaction may also be fatal and probably not known by the physician (Edwards and Aronson 2000; Läkemedelsverket 2012).

In the text mining terminology of adverse drug events the following concepts are used: an *adverse drug event (ADE)* may occur when a drug is taken to treat a

disease. The disease may be indicated by a number of symptoms or findings. This is an *indication* for prescribing a drug. A *contraindication* is a factor or reason to cease or withhold the treatment of the patient with the drug, since the it may harm the patient.

The drug treats the disease but the drug may also give the patient an undesired side effect or adverse drug event, a so-called *ADE cause*, in terms of a new finding or a new disorder, this drug reaction is called an *ADE Cue* in our clinical text mining processing vocabulary (Henriksson et al. 2015). Another article describing the terminology in a nice way is written by Nebeker et al. (2004).

Some of ADEs are known, others are not known and hence are called unknown drug events or unknown side effects. Some drugs also interact with other drugs and give undesired effects, so-called *drug-drug interactions* or just *drug interaction*, (some drug-drug interactions can also give a positive effect).

Sometimes a drug is prescribed to remove the adverse drug effect or symptoms of a drug. For example, when a drug is given as chemotherapy to a patient to treat cancer tumors and the patient feels nausea as an adverse drug effect, then another drug may be given to the patient to mitigate the nausea.

The state of the art in drug development is personalised medicine, meaning that each patient should, for example, obtain a customised drug adapted specifically for that individual's genetic properties and corresponding disorder.

10.2.2 Resources for Adverse Drug Event Detection

To detect adverse drug events expressed in textual form and in structured data in the patient record a number of resources are needed.

First of all, ICD-10 diagnosis codes related to adverse drug events that are assigned to the patient records need to be studied, see Table 10.2, for examples. These ICD-10 diagnosis codes are assigned to the discharge summary by the physician if a patient has experienced an ADE, but in many cases in no code assigned even if an ADE has occurred.

Secondly, natural language expressions, indicating ADEs in the clinical text can be studied. In a clinical text when a patient is suffering some form of adverse drug reaction it is expressed by a phrase such as: *drug-induced reaction in form of drug X*, or just *reaction, hypersensitivity* or *adverse reaction*, it can also be vague or uncertain expressions such as *suspected hypersensitivity*, or *possible reaction on drug*. Expressions similar to these have been extracted from Swedish clinical text by human annotators (Friedrich and Dalianis 2015; Henriksson et al. 2015). The adverse event cues in Swedish used in Friedrich and Dalianis (2015) can be found on the DSV website.[3]

[3] Adverse event cues in Swedish from Friedrich and Dalianis (2015), see under *Swedish ADE word lists* in http://dsv.su.se/health/tools. Accessed 2018-01-11.

Table 10.2 ICD-10 diagnosis codes for adverse drug events

ICD-10 code	Description
E27.3	Drug-induced adrenocortical insufficiency
G24.0	Drug induced dystonia
G25.1	Drug-induced tremor
G44.4	Drug-induced headache, not elsewhere classified
G62.0	Drug-induced polyneuropathy
I42.7	Cardiomyopathy due to drug and external agent
I95.2	Hypotension due to drugs
L27.0	Generalized skin eruption due to drugs and medicaments taken internally
L27.1	Localized skin eruption due to drugs and medicaments taken internally
N14.1	Nephropathy induced by other drugs, medicaments and biological substances
T78.2	Anaphylactic shock, unspecified
T78.3	Angioneurotic edema
T80.8	Other complications following infusion, transfusion and therapeutic injection
T88.7	Unspecified adverse effect of drug or medicament
T88.6	Anaphylactic reaction due to adverse effect of correct drug or medicament properly administered

The table is taken from Table 1 in Henriksson et al. (2015). Licensed under Creative Commons

Another way to find typical expressions is to use *Farmaceutiska Specialiteter i Sverige (FASS)* , the Swedish corresponding list of the American *PDR, Physician's Desk Reference*. FASS contains all drugs on the market in Sweden and is written for physicians, nurses, dentists and all clinicians allowed to prescribe drugs in Sweden. FASS listed each drug name, pharmacological property, ATC-code and which type of adverse reaction a drug may give, along with the probability of this. To use these terms, it is necessary to download them.

The same information can be extracted from the search tool at *Läkemedelsverket*[4] and, the API[5] and the open data for drugs can be found online.[6] The data in XML format is also available.[7]

The terms used for side effects are often very general and can not really be connected to a particular drug. Resource for performing text mining research for detecting adverse drug events is the *WHO Adverse Reactions Terminology (WHO-ART)*.[8]

[4]Läkemedelsverket, the Swedish Medical Products Agency, drug search tool (in Swedish), https://lakemedelsverket.se/LMF/. Accessed 2018-01-11.

[5]API, http://lakemedelsboken.se/api/. Accessed 2018-01-11.

[6]Open data (in Swedish), https://lakemedelsverket.se/psidata. Accessed 2018-01-11.

[7]XML, https://npl.mpa.se/mpa.npl.services/home2.aspx. Accessed 2018-01-11.

[8]WHO-ART, http://www.umc-products.com/DynPage.aspx?id=73589&mn1=1107&mn2=1664. Accessed 2018-01-11.

If a more fine grained terminology is needed the *Medical Dictionary for Regulatory Activities (MedDRA)* can be used.[9]

10.2.3 Passive Surveillance of ADEs

Postmarketing adverse drug surveillance is important, but is mainly carried out spontaneously by sending a report to the *Uppsala Monitoring Centre (UMC)*[10] when an adverse event has occurred. (UMC is a WHO collaborating centre for international drug monitoring).

Postmarketing adverse drug surveillance is also called *passive surveillance*. However, many physicians consider some ADEs as well known and unavoidable while treating the patient for a more serious, may be lethal disease. Therefore, new methods are needed to perform postmarketing adverse drug surveillance, one method is, for example, to use the large repositories of electronic patient records, both the structured and the unstructured part. The patient records may contain typical patterns of ADEs and undertaking a statistical and register-based data analysis may reveal new or unknown events (Jensen et al. 2012; Harpaz et al. 2012).

10.2.4 Active Surveillance of ADEs

The concept of *adverse event alerting systems*, is using patient data and to analyse adverse event has occurred, this method is also called *Active surveillance*. In a nice review article by Forster et al. (2012), over 48 systems were compared, the earliest from 1988 and the latest from 2008; however, it is not clear exactly how the alerting systems work. The review article mentions rules and a gold standard but it is not clear how the input data is formatted, whether the input data is structured or unstructured. The authors express a wish, in the article, that the systems could also use free-text for the analysis of adverse drug effects. Generally, the different systems presented in the review article, gave rather poor results. The average sensitivity (recall) is around 60% and the average specificity is around 60%, the best system had 94% sensitivity and 71% specificity. The systems are also difficult to compare because they use different input data, but this is a common problem in many applications.

In Bailey et al. (2016) 108 adverse event reporting systems were compared and most of the systems used free-text in the analysis. One finding was that half of the systems had qualitative scores between 60% and 80%, meaning no system had top

[9]MedDRA, http://www.meddra.org. Accessed 2018-01-11.

[10]Uppsala Monitoring Centre, http://who-umc.org. Accessed 2018-01-11.

results and it was also difficult to compare the systems since the input data was different for each system.

In the review article by Warrer et al. (2012) the authors have focused an text mining-based adverse drug alerting systems. Over 200 articles, published from 2001 to 2010, but mainly between 2009 and 2010, were investigated. Seven articles were selected, however the articles presented low specificity and positive value, or precision (PPV), values for the different systems. Warrer et al. (2012) conclude that text mining can only be a supplement to manual chart review for screening large amounts data.

Stausberg and Hasford (2011) carried out a registry study in Germany over the years 2003–2007 using 505 specific ICD-10 codes as definitions for ADEs, the codes were in turn categorised into seven groups depending on the certainty. The finding was that 5% of the hospital episodes were either caused or complicated by an ADE, this is based on approximately 48 million hospital episodes equating to approximately 12 million hospital episodes per year. Similar results were found for the Stockholm EPR Corpus for the years 2009–2010, with 6.6% of the patients having an ADE, calculated on 703,173 patient records (Lagos 2016).

In an extensive article of text and data mining techniques by Karimi et al. (2015b), the authors review the area of ADEs, and also various tools for both data and text mining applied on clinical records and social media to detect ADE.

Another interesting review article in the area of text mining ADEs is by Luo et al. (2017).

10.2.5 Approaches for ADE Detection

To find adverse events, the clinical entities have to be detected, such as *symptom, diagnosis, body part* and *drug*, as well as the relations between the entities.

As previously described, the early approaches in detecting ADEs were rule-based approaches, and the earliest systems did not even analyse the free text but only the structured data.

Here follow some approaches using the textual information as input and a rule-based approach to detect the ADEs.

Eriksson et al. (2013) performed a rule- and dictionary-based approach to detect ADEs in 6011 Danish psychiatric patients' hospital records. The system identified 35,477 unique ADEs with a precision of 0.89 and a recall of 0.75.

Wang et al. (2009) developed a rule-based system to detect the relationship between the drug and the ADEs for seven different types of drugs (ibuprofen, morphine, warfarin, bupropion, paroxetine, rosiglitazone and ACE inhibitors). 25,074 discharge summaries in English were used to evaluate the system. The authors obtained a recall and precision of 0.75 and 0.31 respectively for known ADEs.

Hazlehurst et al. (2009) used the Kaiser Permanente Northwest (KPNW) database containing records for more than 450,000 people to detect vaccine

ADEs. Two automatic methods were compared, the MediClass and the code-based detection. MediClass method obtained better results, than code-based method with 0.74 versus 0.31 PPV (positive predictive value, or precision).

Here follows some machine learning-based approaches for detecting adverse drug events.

Gurulingappa et al. (2012) manually annotated an English corpus containing 3000 medical case reports (i.e. published scientific reports for selected patients with regard to drugs and side effects). The annotation concepts used were *drugs, drug dosage, adverse effect* and their relationship. The corpus was annotated jointly by three annotators, two experienced and one novice annotator in text mining related topics. All of the annotators had an M.Sc. degree in biomedicine.

Each annotator annotated 2000 case reports, and 1000 case reports were annotated by all three annotators. Inter-annotator agreement (IAA) for drugs obtained F-scores for partial match ranging from 0.90 down to 0.38. IAA for relations for drugs-adverse effect obtained F-scores ranging from 0.79 down to 0.37. The ADE-corpus is publicly available.[11]

Naïve-Bayes and Maximum Entropy (MaxEnt) classifiers from the MALLET toolkit were used for training, when evaluated they obtained at best 0.75 precision and 0.64 recall for MaxEnt.

Santiso et al. (2014) studied Spanish clinical text and used 6100 concepts and 4700 adverse drug reactions (ADRs) relations for training using the Random Forest algorithm. 2100 concepts and 1600 ADR relations were used for evaluation, and their Random Forest approach obtained 0.93 precision and 0.85 recall.

Aramaki et al. (2010) used 3012 Japanese discharge summaries containing 1045 drugs and 3601 possible adverse drug effects that were manually annotated. The finding was that 7.7% of the discharge summaries contained ADEs, 59% of these could be extracted automatically. Both support vector machine (SVM) and pattern matching (PTN) methods were used. Marginally better results were obtained using PTN. PTN gave a precision of 0.41 and a recall of 0.92, whereas SVM gave a precision of 0.58 and a recall of 0.62.

Roller and Stevenson (2014) used UMLS to identify concepts, such as drugs and contraindications, and relations, such as ADE drug relations, in millions of biomedical scientific articles. A Naïve-Bayes classifier was trained on the data and 25% precision and 100% recall was obtained.

In another type of approach that is not connected to clinical text mining, but to social media analytics Karimi et al. (2015a) used social media textual posts of patient-reported ADEs to find adverse drug events. The annotated corpus is called the CSIRO Adverse Drug Event Corpus (CADEC).

[11]Benchmark corpus to support information extraction for adverse drug effects, https://sites.google.com/site/adecorpus/. Accessed 2018-01-11.

An Approach for Swedish Clinical Text

The *Clinical Entity Finder (CEF)* (Skeppstedt et al. 2014) was used by Henriksson et al. (2015) to speed up annotation by utilising pre-annotation. *Pre-annotation* means a text is annotated by a machine to assist the human annotator. The model for pre-annotation was trained on annotated records from an internal medicine emergency unit obtained from an earlier annotation study. The records were annotated with the entities *finding, disorder, body structure* and *pharmaceutical drug*.

The clinical text to be annotated was extracted based on the ICD-10 codes indicating ADEs, see Table 10.2 for the specific codes.

The annotations were carried out by three annotators, one junior and one senior computer scientist and one physician. The three annotators corrected both the pre-annotated entities and added missing ones. They also added an annotation for the new entity *ADE cue*. They added optional attributes to the clinical entities *finding, disorder, body part, drug* and *ADE*. Attributes such as *negation, speculation, past* and *future*. Then commenced the difficult part of the annotation work, and which was to draw arches to annotate the indications and also the ADEs. The following semantic relations between named entities were annotated: *indication adverse drug event, ADE outcome* and *ADE cause*.

The agreement for the annotators versus the pre-annotations was high, giving a macro-averaged F-score of 0.825, indicating the pre-annotation model generalises well to a different medical domain.

The inter-annotator agreement was fairly high, above an F-score of 0.8. Lower agreement was found for the class ADE cue, may be due to not being defined properly or because it was mixed up with the classes disorder and finding. In total 3789 named entities, 1642 attributes and 2266 relations were annotated. Bag of Words (BOW), Semantic Vectors (SV) and Multiple Semantic Vectors (MSV) were all used for the text mining; however, the relation mining part produce rather low F-score, the best for SV, being below 0.30. This is probably because the relations cross over several sentences and are difficult to detect.

The PhD thesis by Henriksson (2015) contains a nice approach for using unsupervised learning and ensembles of semantic spaces within random indexing, to produce features used in machine learning to detect ADEs in Swedish clinical text.

In Fig. 6.1 we can see part of the annotation work carried out by Henriksson et al. (2015), which is part of Henriksson's PhD thesis (Henriksson 2015).

An Approach for Spanish Clinical Text

Casillas et al. (2016) developed a system to detect adverse drug reactions (ADRs), for Spanish electronic health records. The patient records were obtained from Galdakao-Usansolo Hospital in Usansolo, Bizkaia, Spain. In total they used 194 patient records containing 101,685 words. The documents were annotated by four

experts in pharmacovigilance from the same hospital; however, no inter-annotator agreement results were reported. In total 3084 entities were annotated jointly with 4994 events. The events involve drug-disease pairs where the drug caused the disease, i.e. an adverse drug event including the causal relation.

One remark was that the majority of the relations are inter-sentential, stretching over more than 10 sentences from the disease entity to the drug entity, in the positive ADR event cases. Distance between entities in terms of the number of sentences from the disease entity to the drug entity in the case of positive ADR events was one of the features, other features used were the symptoms lexicon, the drugs lexicon, and keywords between the two lexicons.

One other obvious remark was that there was an imbalance between negative and positive ADR events, where only 6% of the drug-disease entity pairs triggered a positive drug reaction.

The annotated corpus was divided into a training set and a test set. Random Forest, (RF) and Support Vector Machine (SVM) algorithms were used for the training.

In parallel, a rule-based system was also developed as well as a hybrid system using a combination of the rule and machine learning-based approaches.

The rule-based system obtained a rather high precision of 0.89 but a low recall of 0.12, and both the machine learning and the hybrid systems produced poor results, with precision and a recall of less than 0.54.

In the article it also states that the "hospital found the system well worth using, not as an automatic ADR event extraction system but as a decision support system" (Casillas et al. 2016).

A Joint Approach for Spanish and Swedish Clinical Text

Pérez et al. (2017) carried out a joint approach for Spanish and Swedish clinical text to extract disease and drug in Spanish and body part disorder and finding in Swedish using the same methods on different corpora. The methods were maximum probability, CRF, perceptron and SVM. CRF gave the best result on the development sets for both languages and the perceptron the best result on the Spanish test set.

10.3 Suicide Prevention by Mining Electronic Patient Records

Adverse event detection in electronic patient records for suicide risk is an important application for clinical text mining. According to Leonard Westgate et al. (2015) suicide was, in 2010, the tenth leading cause of death in the United States and among the top four leading causes of death for Americans between the ages of 10 and 54.

There have been some early applications in the area of clinical text and data mining in the domain, most approaches have been to perform data mining on the structured information in patient records such as ICD-10 or ICD-9 coding for suicide attempts and self harming (Tran et al. 2014; Barak-Corren et al. 2016).

However, Leonard Westgate et al. (2015) used simple linguistic analysis and Gkotsis et al. (2016) developed a negation detection solution using psychiatric electronic health records for detecting suicide risk.

Haerian et al. (2012) used a combination of ICD-9 coding and UMLS Concept Unique Identifiers for text mining of suicidal expressions in the electronic patient records. The best results they obtained were a positive predictive value (PPV) or precision of 97% with a 90% confidence interval of 0.92–0.99.

Metzger et al. (2016) studied electronic patient records in French from an emergency unit to carry out an epidemiological surveillance study of suicide attempts. The authors used 444 patient records with both recorded suicide attempts and suicidal ideations and 292 patient records as control group, they used the Random Forest and Naïve Bayes machine learning methods obtaining F-scores ranging from 70.4% to 95.3%.

Their study shows that the national manual coding made at the Croix-Rousse Hospital (Red Cross Hospital) were massively underestimated, as the number of events, only 15 of total 98 emergency department visits related to suicide attempts in 2012, were recorded in the national manual coding, while their text mining approach gave adequate results with much less work load for the physicians in the emergency department.

In another study by Downs et al. (2017) 230,465 British English anonymised clinical records were used containing 500 patients with autism spectrum disorders (ASD). The aim of the study was to detect suicide risk in ASD adolescents. The authors used a rule-based NLP approach and obtained high system performance with precision, recall and F-scores of over 0.85 for detecting positive and negative suicidality. Each patient record encompasses a large number of notes over time, it is, therefore, difficult for the annotators to classify whether something is suicide-related or not from only the information in one note.

Moreover, suicide-related terms are very rare in the notes. less that 3%. Another difficulty is that the terms might be references to the past, about other family members, relatives or friends. Suicidal expressions can be very vague, e.g. “took an excessive amount of pills”, or very abstract, it can also be as part of a behavioural change of the patient without even explicitly mentioning suicide-related information. These are problems a NLP tool has to overcome as well as the human annotator.

There are still relatively few articles in this domain, using free text in the patient records for suicide risk detection, but hopefully research in this area will increase soon.

10.4 Mining Pathology Reports for Diagnostic Tests

There have been attempts to automatise the process of interpreting, the unstructured text of pathology reports and automatically enter it into the database of the cancer registry, mostly using various text mining tools to extract the diagnostic tests from the pathology text. Part of this work has also involved encouraging pathologists to enter structured information into the pathology report hence making it easier for the text mining tools. A description and overview of various tools is available in Scharber (2007).

The text mining tools for pathology reports are mostly rule-based but there are also some machine learning-based tools, for a review article on the topic, see Spasić et al. (2014) and also Weegar and Dalianis (2015).

Currie et al. (2006) describe a rule-based approach to extract concepts from 5826 breast cancer and 2838 prostate cancer pathology reports. Their system extracted 80 fields and obtained 90–95% accuracy. The system was evaluated by domain experts.

Coden et al. (2009) have written an extensive article on how to extract nine different classes from the pathology free text describing colon cancer using both machine learning and rule-based approaches. Hybrid methods were used in the work by Ou and Patrick (2014). The authors studied pathology reports concerning primary cutaneous melanomas (skin cancer) and extracted 28 different concepts.

Martinez and Li (2011) used machine learning to classify pathology reports for colorectal cancer according to the TNM staging scale. The authors used the Weka toolkit and the algorithms Naïve-Bayes and SVM for the best results. In Nguyen et al. (2011) a rule-based method for mining pathology reports for lung cancer is described.

A study with very good results was carried out by Buckley et al. (2012) on 76,000 breast pathology reports, where they obtained results with 99.1% sensitivity (recall) and 96.5% specificity; however, it is not clear what method they used to extract the information apart from the commercial tool developed by *Clearforest, Waltham, MA*, nor is the method for evaluation of their results explained.

Another study using machine learning on surgical pathology reports for cancer also gave good results with a maximum of 99.4% accuracy using the perceptron algorithm with uneven margins (the PAUM-algorithm). The reports were manually annotated pathology reports in British English. The training set and test set contained 635 and 163 reports respectively. The evaluation used was fivefold cross-validation (Napolitano et al. 2016).

Two studies using machine learning, both for breast cancer pathology reports one for English and one for Chinese were reported in Yala et al. (2017) and in Tang et al. (2018) respectively. Both studies obtained promising results.

10.4.1 The Case of the Cancer Registry of Norway

At the Cancer Registry of Norway (Kreftregisteret) in Oslo over 25 full time experts are manually coding 180,000 pathology reports annually from the whole of Norway which has a population of 5.3 million inhabitants. The pathology reports are written in Norwegian. The experts read the free text in the pathology reports and produce structured coding that is stored jointly with the pathology reports in XML-format. The coding is carried out for research and statistical purposes, and stored in a database.

Three studies will be described here that use the same set of pathology reports from the Cancer Registry of Norway as input data. All three studies use rule-based approaches. One of the studies used a larger data set than the other two.

The first study is by Singh et al. (2015) using 25 pathology reports for prostate cancer written in free text in Norwegian. The authors used the SAS Institute software to extract fields. Their system obtained 76% correctly extracted fields for number of biopsies and 24% for number of biopsies containing tumor tissue, and 100% for Gleason score; however, the study focuses on system development and the evaluation of the system is not described.

The second study is by Dahl et al. (2016) using the same 25 pathology reports for prostate cancer in free text used by Singh et al. (2015). Half of the reports were used as a development set and the other half as a test set. The developed system obtained an F-score of 0.94 on four data points *total malign, primary gleason, secondary gleason* and *total gleason*.

The third study is by Weegar et al. (2017) that used a much larger subset of pathology reports for prostate cancer than used by both Singh et al. (2015) and Dahl et al. (2016). Weegar et al. (2017) built a rule-based system which extracted structured information from over 554 pathology reports for prostate cancer written in Norwegian. The authors divided the 554 pathology reports in 388 documents for the development set and 176 documents for the test set. The system extracts structured information from the free text describing biopsies and Gleason grades. The system obtained an F-score of 0.91. The most interesting part of the system is a flagging mechanism that identifies reports that contain ambiguities or other problems, and therefore need manual review. The system shows the possibility to automatise encoding and make it faster but still with high quality. See Fig. 10.3 for an example of a pathology report for prostate cancer.

There is also a study on Norwegian pathology reports for breast cancer, also from the Cancer Registry of Norway carried out by Weegar and Dalianis (2015). In total there were 40 pathology reports, of these 30 reports were used for developing a rule-based system and 10 reports for testing the system. An F-score of 0.86 was achieved. See Fig. 10.4 for an example of a pathology report for breast cancer written in Norwegian.

Generally, one conclusion is that most systems give around 80% precision and recall on average.

```
Biopsier fra venstre prostatalapp.
2: Prostatakarsinom, Gleason score 3+4=7(4/13 mm)
4: Prostatakarsinom, Gleason, score 3+3=6(0,5/12
mm)
1,3:Ikke påvist malignitet
Biopsier fra høyre prostatalapp:
5-7,9: HPIN og Prostatakarsinom, Gleason score
3+3=6(1/13 mm)
8: Prostatakarsinom, Gleason score 4+3=7(5/15,4/15
mm)
Perinevral infiltrasjon: ikke påvist
```

Translated to English:

```
Biopsies from the left lobe.
2: Prostate carcinoma, Gleason score 3+4=7(4/13 mm)
4: Prostate carcinoma, Gleason score 3+3=6(0.5/12
mm)
1,3:No signs of malignancy
Biopsies from the right lobe:
5-7,9: HPIN and adenocarcinoma, Gleason score
3+3=6(1/13 mm)
8: Prostate carcinoma, Gleason score 3+4=7(5/15 mm)
Perineural invasion: no signs
```

Fig. 10.3 A pseudonymised example of a Norwegian pathology report describing prostate biopsies. The text contains descriptions of 9 biopsies, four from the left side and five from the right side. Figure published in Weegar et al. (2017)

10.4.2 The Medical Text Extraction (Medtex) System

The Australian e-Health Research Centre (AeHRC) at CSIRO in Brisbane in Australia is closely connected with the Royal Brisbane and Women's Hospital. Nguyen et al. (2015) at the AeHRC constructed the Medical Text Extraction (Medtex) system, built in the Java programming language, which automatically processes a trickle feed of incoming pathology reports in HL7 format from the whole state of Queensland in Australia. The State of Queensland has almost 5 million inhabitants hence is almost as populated as Norway. Medtex uses NLP techniques, and the external resource of SNOMED CT for mapping and identifying medical concepts and abbreviations. A set of business rules was constructed for finding the structured data in the unstructured pathology text. The system was evaluated using a set of 220 unseen pathology reports and obtained an F-measure of 0.80 over seven categories.

In Fig. 10.5 we can see the interface of the CIPAR annotation system, which is part of the Medtex system.

```
Mammaresektat (ve. side) med infiltrerende duktalt
karsinom, histologisk grad 3
Tumordiameter 15 mm
Lavgradig DCIS med utstrekning 4 mm i kranial
retning fra tumor
Frie reseksjonsrender for infiltrerende tumor (3 mm
kranialt)
Lavgradig DCIS under 2 mm fra kraniale
reseksjonsrand

ER: ca 65 % av cellene positive
PGR: negativ
Ki-67: Hot-spot 23% positive celler. Cold spot 8%.
Gjennomsnitt 15%
HER-2: negativ
Tidl. BU 13:

3 sentinelle lymfeknuter uten påviste patologiske
forandringer
```

Translated to English:

```
Mamma specimen (le. side) with infiltrating ductal
carcinoma, histological grade 3
Tumor diameter 15 mm
Low-grade DCIS extending 4 mm in cranial direction
from the tumor
Free resection margins for infiltrating tumor (3 mm
cranially)
Low-grade DCIS less than 2 mm from the cranial
resection margin

ER: ca 65 % of the cells are positive
PGR: negative
Ki-67: Hot-spot 23% positive cells. Cold spot 8%.
Average 15%
HER-2: negative
Prev. BU 13:

3 sentinel lymph nodes without proven pathological
changes
```

Found concepts by a mockup system

```
Progesteronreseptorer (PGR): 1 (1 is a table value
that corresponds to "negative" in the text)
Samtidig Sentinell Node: 0
Østrogenresepttorer (ER): 4 (4 is a table value
that corresponds to "65 %" in the text)
KI67 Hotspot: 23
Tumors histologiske grad (Histological grade): 3
KI67 Gjennomsnitt hot/cold (Average): 15
Tumordiameter (Tumor diameter): 15
(In some cases the data is not found in the text
but in sketch attached to the pathology report).
```

Fig. 10.4 Extract from the free text part of an anonymised breast cancer pathology report in Norwegian (and its translation to English). This is a small subset of a report, with very few values, breasts cancer reports may have over 80 values. The data in the figure is made up and can not be linked to any individual (© 2015, Association for Computational Linguistics (ACL). All rights reserved. The Norwegian pathology text is reprinted with permission of ACL and the authors. Published in Weegar and Dalianis (2015). The translation to English and extracts from the free text are added in this publication)

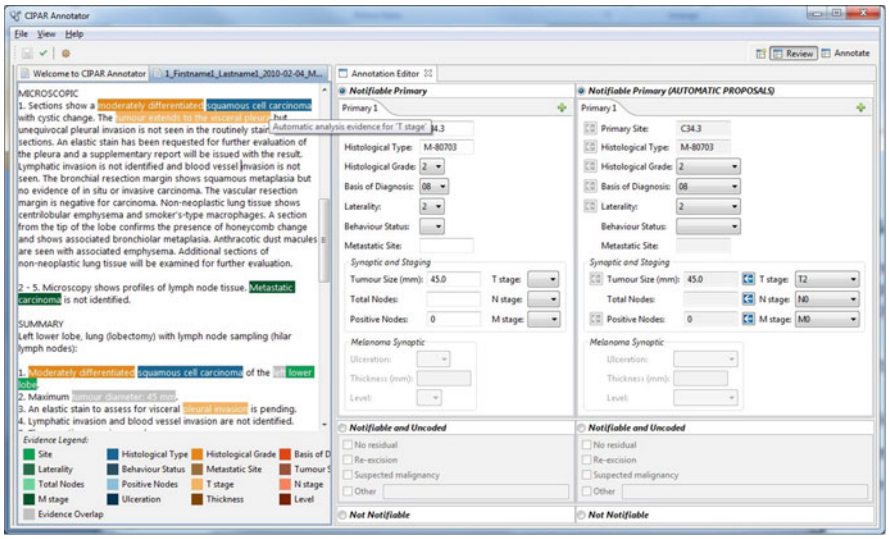

Fig. 10.5 Screenshot of the CIPAR annotation system. The Medtex software processes narrative reports and generates structured data to aid clinical staff in abstraction tasks, in this case lung cancer pathology reports. Taken from Figure 4 in Nguyen et al. (2015) (© 2015 Reprinted with permission from AMIA and the authors. Published in Nguyen et al. (2015))

10.5 Mining for Cancer Symptoms

There is a demand to predict diseases before they occur, or at least as early as possible, so the correct treatment can start early to minimise the suffering of the patient. One area in which to perform prediction is in the cancer domain. Cancer encompasses over 200 cancer types, where the name of the cancer type originates from the affected body organ. Each cancer type has its own and is treated according to each type.

Spasić et al. (2014) have written a nice review of different approaches for performing text mining for different cancer types. The conclusion is that most systems are based on pattern matching and rule-based methods for NER obtaining F-scores in the range of 0.80–0.90. To increase the performance, the systems need to deal with clinical sublanguage and non-standard abbreviations as well as misspellings and grammatical errors.

Jensen et al. (2012) focus on predicting the outcome for patients with a large number of diseases but also specifically in predicting the survival for breast cancer patients. Jensen et al. (2012) propose the use of multivariate models including variables such as age, sex, smoking etc. to make inferences on unknown data. Moreover, they recommend integrating patient record data with genetic data to understand genotype-phenotype relationships, but this requires careful consideration of ethical aspects.

In the domain of cancer text mining we can distinguish two aims and corresponding methods, one is the NER method, to find symptoms, disorders and affected body parts and the other is the classification of text to identify certain features which may indicate a type of cancer.

Weegar et al. (2015) carried out an approach to find symptoms of cervical cancer in Swedish patient records. Early detection of cervical cancer is crucial for the treatment and survival of the patient. The Clinical Entity Finder (CEF) was used, CEF is based on CRF++ and trained on records from Swedish internal medicine emergency units that are annotated for finding, disorder, body part and drug (Skeppstedt et al. 2014). The CEF was extended in the study with the rule-based NegEx negation detection system for Swedish by Weegar et al. (2015).

To evaluate the system two annotators who were also trained physicians, annotated a test set of patient records. The test data consisted of 646 records with patients diagnosed with cervical cancer that had been assigned the ICD-10 diagnosis code C53.

The inter-annotator agreement for finding, disorder and body part was on average an F-score of 0.677. The Clinical Entity Finder extended with NegEx obtained an average F-score of 0.667 (Weegar et al. 2015). The most frequent findings and disorders can be found in Table 10.3.

Zhao and Weng (2011) describe a system called iDiagnosis to predict pancreatic cancer. The system operates by combining PubMed knowledge and electronic health records (EHRs). iDiagnosis extracts 20 risk factors from PubMed abstracts. Keywords were used to classify the risk factors such as positive, negative or neutral associations. Risk factors can be in the classes of demographics, lifestyle, symptoms, co-morbidities and lab test results. Each variable of the model was assigned probabilities for patients with pancreatic cancer based on the information in PubMed, Prior probability for each variable was calculated using the EHRs. A model based on weighted Bayesian Network Inference (BNI) was trained, and pancreatic cancer could be predicted with the sensitivity (recall), specificity and accuracy of around 85%.

10.6 Text Summarisation and Translation of Patient Record

Automatic text summarisation is the technique where a computer program summarises a text. A text is entered into the computer and a summarised text is returned, which is a shorter non-redundant extract from the original text.

Text abstraction is when a completely new text abstract is created, this is similar to what a human does when a human reads the text, comprehends the text and then rephrases it into a completely new text with new word phrasing and new syntactical constructions—an abstract. To perform abstraction, the source text needs to be parsed syntactically and semantically into a formal representation, and then a completely new, shorter and non-redundant text is generated from the

Table 10.3 The most frequent findings, disorders and negations found in the physicians' notes

Most frequent findings and disorders	Nbr of instances	Most frequently negated findings and disorders	Nbr of negated instances	Findings and disorders with highest portion of negation	Portion negated
Cervixcancer (cervical cancer)	873	Besvär (trouble/problem)	338	Gynekologiska besvär (gynecological problems)	1.0
Besvär (trouble/problem)	790	Feber (fever)	243	Palpabla resistenser (palpable resistance)	1.0
Illamående (nausea)	677	Illamående (nausea)	198	Särskilda besvär (particular problems)	1.0
Mår bra (feels well)	662	Blödningar (bleedings)	171	Nytillkomna symtom (new symptoms)	0.96
Smärta (pain)	656	smärta (pain)	150	Nytillkomna besvär (new problems)	0.92
Tumör (tumor)	642	smärtor (pains)	126	Infektionstecken (signs of infection)	0.89
Smärter (pains)	629	Blödning (bleeding)	99	Biljud (murmur)	0.86
Feber (fever)	562	Infektionstecken (signs of infection)	91	Tumörstrukturer (tumour structures)	0.83
Cancer (cancer)	508	Tumör (tumor)	83	Tumörsuspekta förändringar (tumor suspicious changes)	0.82
Blödningar (bleedings)	491	Nytillkomna besvär (new troubles)	79	Subjektiva besvär (subjective problems)	0.8
Blödning (bleeding)	482	Buksmärtor (pain of the abdomen)	72	Tumörsuspekt (tumor suspicion)	0.80
Skivepitelcancer (squamous cell carcinoma)	428	Hydronefros (hydronephrosis)	65	Spridning (spreading)	0.78

Taken from Table 7 in Weegar et al. (2015) (© 2015 Reprinted with permission from AMIA and the authors. Published in Weegar et al. (2015))

formal representation, with new wording and syntactical structure but with the same meaning.

An automatic text summarisation system usually produces summary extracts, so-called extraction-based summarisation systems. An abstraction system is very difficult to construct therefore most systems are extraction based systems.

One of the first systems to create an automatic summary or abstract was described in Luhn (1958). The purpose was to create an informative extract to help researchers and scientist to cope with the growing number of publications. The extract was compiled from the most important parts of the article.

Later text summarisation systems were created to summarise new articles. Multi-document text summarisation systems have also been developed that analyse several similar news articles about the same topic to one non-redundant summary. Mani and Maybury (1999) have carried out a compilation of the state of the art in text summarisation including an historical review.

An extractive text summarisation system works basically by splitting a text into sentences and then finding the most information intensive or meaning bearing words of the texts. The meaning bearing parts are usually title words, initial sentences (especially in news text) and sentences containing named entities, such as personal names, organisations, locations and numerical values, but also sentences containing verbs that are frequent. Stop words, which are not meaning bearing words, are filtered away.

All these measurement points are put together to give each sentence a numerical ranking or scoring. The ranking is in turn normalised depending on the length of the sentence, since longer sentences tend to contain more information than shorter sentences. Finally, the summary is compiled from the highest ranking sentences in consecutive order. The length of the summary is decided beforehand or by a cut-off value. The final summary is shorter, hopefully coherent and almost as information intensive as the original text.

There are many different ways to perform automatic text summarisation, the example just described is a rule-based or heuristic (rule of thumb) approach, but there are also approaches that use statistics from large corpora or pure machine learning methods trained on example texts and their corresponding abstracts. The algorithms based on these methods can then produce a summary that is similar to other learned summaries or abstracts.

There is also a method called query based or keyword based text summarisation that customises the summary around what the user wants to read about. The summary is then focused around the specific keywords the user enters to the system. This approach is slightly similar to the text snippets focused around the search words in the ranked search list obtained when carrying out a regular search in a search engine.

There is a group of text summarisation systems using distributional semantics and word embeddings to construct summaries see Hassel (2007), Hassel and Sjöbergh (2006) and specifically for clinical text by Moen et al. (2016) that will be discussed in the next section.

10.6.1 Summarising the Patient Record

One requirement from the physicians regarding summarisation of patient records is to summarise the whole healthcare episodes or period of patient treatment into a discharge summary, until today there have been few approaches.

Early work on understanding the content of the patient records is described in LifeLines where information is extracted and presented in a timeline (Plaisant et al. 1998), this is described in Sect. 3.3 and specifically in Fig. 3.2.

Pivovarov and Elhadad (2015) review different approaches of summarising a patient record, but none of them really address the challenge of creating a discharge summary from a healthcare episode.

Van Vleck and Elhadad (2010) describe an approach to find the most relevant information in the form of clinical problems to be included in a patient summary. The authors treat this problem as a classification problem. The classifier is trained on a corpus of patient notes and their corresponding problem list using the two machine learning algorithms Naïve-Bayes and J48 (C4.5) implemented in the Weka toolkit. One issue was how to exclude negated events, or events in the summary that had not occurred yet. The classifiers obtained an accuracy of 82% and an F-score of 0.62 on this task.

Aramaki et al. (2009) propose a text-summarising system TEXT2TABLE, which extracts relevant information from the patient record and presents it in a table. The problem is similar to the one in Van Vleck and Elhadad (2010) to filter out non-relevant information from the patient records as negated events, or events that will happen in the future or may happen. The authors trained their algorithms on 435 Japanese discharge summaries in which seven different events types that should be *excluded* in the summary, such as *negation, future, (planned) purpose, S/O (suspected disease), necessity, intend, possible, recommend* (by other doctor) and their modalities were annotated. The SVM classifier was trained to decide whether an event had actually occurred or not. The experimental results obtained an F-score of 0.858.

Liu (2009) used the MEAD text summarisation system made for English text to summarise Finnish nursing narratives. The author adapted the system to Finnish. In total 252 text summaries were created and evaluated; however, no evaluation results are presented.

Moen et al. (2016) did excellent work in comparing a number of summarisation methods to create a discharge summary from a consecutive number of documents for an individual patient. 66,884 care episodes in Finnish were used describing patients with cardiac problems. The constructed automatic summarisation system was an extraction-based, multi-document summarisation system since each patient record contain multiple documents used as inputs to the text summariser. The summarisation system used distributional semantics and specifically the random indexing method, to create a word space model for extracting features for the different summarisation methods used.

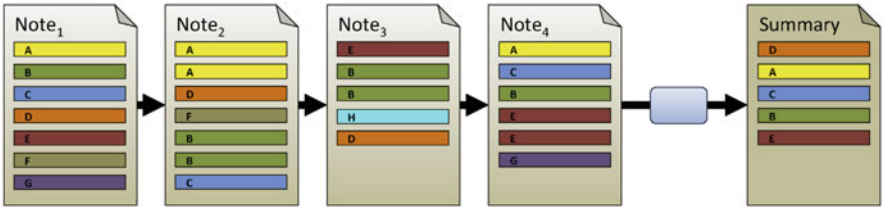

Fig. 10.6 Example of automatic discharge summary creation. Redundant information is removed and high scoring information is added to the beginning of the summary from highest to lowest, low scoring information G, F and H, is excluded. Taken from Figure 3 in Moen et al. (2016), licensed under Creative Commons

Sentences from each document in the care episode are extracted and composed into the automatic discharge summary, see Fig. 10.6.

The authors compare eight different summarisation methods. The summariser uses sentence topic clustering and topic scoring to score the sentences. Topic clustering was used specifically to obtain redundancy reduction. Redundancy can occur both within one document but also across over several documents that are going to be summarised. The similarity measurement for sentence similarity is based on distributional semantics.

One of the best performing summarisation methods was the composite method that combines parts from other methods such as sentence ranking and normalisation of sentence length, the top high ranking sentences are selected for the final summary.

During the experiment each care episode was summarised using eight different summarisation methods producing in total eight different discharge summaries (where one of the "methods" was an original human-made discharge summary). Three domain experts evaluated the discharge summaries produced. In addition, four different evaluation metrics from the ROUGE[12] evaluation package were used. In total 156 care episodes were utilised for the automatic evaluation. This work was also part of Moen's PhD thesis, see Moen (2016).

10.6.2 Other Approaches in Summarising the Patient Record

There are also approaches using *natural language generation (NLG)* techniques to transfer information directly from structured data to text. Portet et al. (2009) used data from a neonatal intensive care data department to create a flow text description of how the baby feels directed towards the parents and relatives.

[12]ROUGE score is a metric that uses unigram co-occurrences between summary pairs: the machine produced one and the gold standard or human produced abstract to calculate the quality of the summary (or machine translation). https://en.wikipedia.org/wiki/ROUGE_(metric). Accessed 2018-01-11.

Torgersson and Falkman (2002) used a mixture of text generation and text summarisation to create an easily navigated and comprehensible patient record called MedView in the area of oral medicine.

Elhadad et al. (2005) present an automatic summarisation system where the discharge summary of the patient is used as an input to search PubMed, to extract and summarise the most relevant information for the patient's characteristics into one article.

Johnson et al. (2008) used a similar approach but took the textual input from the physician, together with other structured information from the patient record, to structure the patient records and create a summary.

10.6.3 Summarising Medical Scientific Text

Sarker et al. (2013) studied an approach to summarise query-focused information from medical scientific text. The text originate from the *Journal of Family Practice* that is aimed at general practitioners. The corpus studied contains almost three thousand medical scientific articles document, of which 1319 were set aside for evaluation, the rest were used for training. The abstracts of the documents are structured and only these are used for training and evaluation. Each document is associated with a clinical query.

Both the query and the content of the document was used to create a summary. Eleven different algorithms were tested, of which five were baseline algorithms (see Sect. 6.3 for the explanation of baseline) and the rest were machine learning algorithms trained on the training corpus. The QSpec (grid search) summarisation algorithm gave the best results using standard sentence scoring rules for text summarisation (each part of the sentence contributes to the total scoring), but where the weights for each sentence score were optimised using grid search. The QSpec system obtained a percentile rank of 96.8% outperforming all the other systems tested in the study.

10.6.4 Simplification of the Patient Record for Laypeople

Related to the summarisation of the patient record is the simplification of the patient record for the layreader. This function has become relevant as individual patient records have become accessible on the Internet for the patients in many countries. The patient can hence read his or hers patient record on-line; however, the patient record is difficult to understand for laypeople since it contains many specialised medical terms and also very domain specific language (Ramesh et al. 2013).

Ramesh et al. (2013) have shown that their system Clinical Notes Aid (NoteAid) improved comprehension of the patient record among laypeople. NoteAid translates the medical jargon in the patient record to a consumer-oriented lay language, it uses

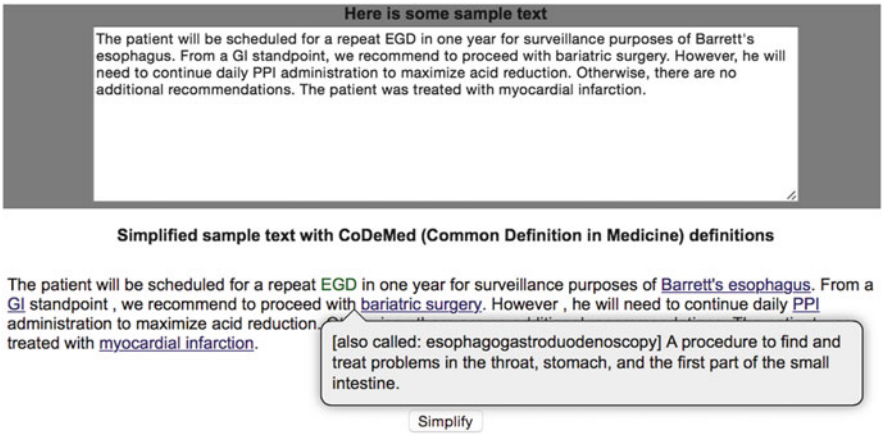

Fig. 10.7 Screenshot of Clinical Notes Aid showing a clinical text where a medical concept is explained in a pop up menu. From the online demo, http://www.clinicalnotesaid.org/emrreadability/notesaid.uwm. Accessed 2018-01-11

Wikipedia, MedlinePlus (a disease dictionary) and UMLS to support the creation of the simplified text. Progress notes are considered to be more difficult to comprehend for laypeople. NoteAid improved comprehensibility most for progress notes using Wikipedia as a dictionary, the results were also statistically significant. See Fig. 10.7 for a screenshot of NoteAid.

Kandula et al. (2010) performed an approach to simplify medical patient record text in English. Texts containing long sentences and difficult words are considered to be hard to read. The authors used a part of speech tagger and a grammar simplifier to replace sentences longer that 10 words with two or more shorter sentences and semantically difficult terms with easier synonyms. The evaluation used the *cloze test* where every fifth word was replaced with a blank and the reviewers were asked to fill in the missing term. The easier it is to replace the missing term, the easier the text is to read. A cloze test on the patient records showed a statistically significant improvement for the simplified patient records from 35.8% to 43.6% in cloze score.

For Swedish there has been one approach for lexical simplification of clinical text with the aim of making it more comprehensible for laypeople. The approach was to resolve unknown words as compounds words, abbreviations or misspellings (Grigonyte et al. 2014). The method used was to first detect and expand all abbreviations, the remaining words may be compounds, unknown words or misspelled words. Finally unknown words are resolved as either as compounded known words, abbreviations or misspellings. The results obtained were: 91.1% precision for abbreviations, 83.5% precision for compound splitting and finally 83.9% precision for spelling corrections.

For Swedish a ground breaking study on grammatical simplification of news text was carried out by Anna Decker in her master's thesis (Decker 2003). She proposed 25 formalised simplification rules. Decker used newspaper articles written in Swedish for immigrants, and compared parallel text, both original and simplified,

to find patterns and create her simplification rules. The simplification rules for Swedish are implemented in the web service SCREAM[13] by Falkenjack et al. (2017). SCREAM also contains a text summariser for Swedish.

10.7 ICD-10 Diagnosis Code Assignment and Validation

There are around 32,000 different ICD-10 diagnosis codes that are used for classifying diseases. The coding is mainly carried out for statistical, administrative and financial reasons but is also useful for the physician to quickly perceive what codes diseases, the patient record is assigned with. The coding is performed by the treating physician but also by specific trained coders. The ICD-10 code is assigned to the discharge letter of the patient.

It is important to have correct and reliable statistics when planning healthcare. However, in an investigation carried out by National Board of Health and Welfare in Sweden (Socialstyrelsen), 4200 patient records were reviewed and it was found that there were 20% errors in the ICD-10 codes. In another investigation encompassing 1.5 million patient records, the National Board found that 1.2% of the main diagnoses were missing (Socialstyrelsen 2010).

Because of these challenges it would be valuable to have a system that automatically assigns ICD-10 diagnosis codes to a discharge summary, or at least proposes probable codes to the treating physician or coder. It would also be valuable to have a system that validates the manually assigned codes and warns when something is wrong.

In 2007 there was a shared task to automatically assign ICD-9 codes to radiology reports. The shared task was called the *Computational Medicine Center's 2007 Medical Natural Language Processing Challenge*. Their data consisted of a training and development set of 978 documents and a test set of 976 documents in the radiology domain. The records were assigned with 45 different ICD-9 labels in the form of 780.6, 786.2 etc, which in this example refer to fever and cough, Pestian et al. (2007).

Over 44 teams participated in the shared task, most approaches used rule-based methods; the second best team consist of Farkas and Szarvas (2008), that used a combined rule-based and machine learning approach (the Maximum Entropy classifier) and obtained an F-score of 0.903 on the training set and an F-score of 0.889 on the test.

Perotte et al. (2014) extended the approach using 22,815 non-empty discharge summaries from the MIMIC II database of which 90%, or 20,533 documents, were used for training and 10%, or 2282 document, were used for evaluation. In total 5030 ICD-9 codes were assigned to the documents, some documents were, of course, assigned multiple codes. The best results for the machine learning approach were

[13]SCREAM Textförenklare (in Swedish), http://www.ida.liu.se/projects/scream/webapp/. Accessed 2018-01-11.

obtained using a hierarchy-based SVM classifier that gave an F-score of 0.40. It assigned an average 6.31 codes per document.

Suominen et al. (2008) used both the Regularized Least Squares (RLS) classifier that is closely related to SVM and the RIPPER rule induction-based learning method. The RIPPER rule induction-based learning method was the best of the two methods and obtained an F-score of 0.877 giving the team a third place in the challenge.

Studies for automatic ICD-10 code assignment for French, Bulgarian, Danish, Swedish and Japanese respectively have been carried out by Lavergne et al. (2016), Boytcheva (2011), Roque et al. (2011a), Henriksson and Hassel (2013) and Aramaki et al. (2014).

Kavuluru et al. (2015) experimented with 93,694 French death certificates and assigned 377,677 ICD-10 codes (3457 unique codes). The data was used for a shared task where the best team obtained an F-score of 0.848.

In Fig. 10.8 we can see the results of building a word space model from 408,144 Swedish annotated patient records and their corresponding 35,185 ICD-10 diagnosis codes using the Random Forest algorithm. The data originate from the Stockholm EPR Corpus's first 5 months of year 2008. This constructed word space model can be used both to find which ICD-10 code corresponds best with a certain word, and which medical term or symptom corresponds best with a certain ICD-10 code.

In Henriksson et al. (2011) the random indexing experiments showed that of the top 10 generated candidates, 20% were correct and 77% were partially correct. The approach was refined in Henriksson and Hassel (2013) where the dimensionality was increased and gave up to 18% better results.

Stanfill et al. (2010) have written an overview of ICD-8, ICD-9 and ICD-10 automatic coding of patient records; almost a historical review is carried out. The best performing systems obtained F-scores of around 0.90 for ICD-9 codes.

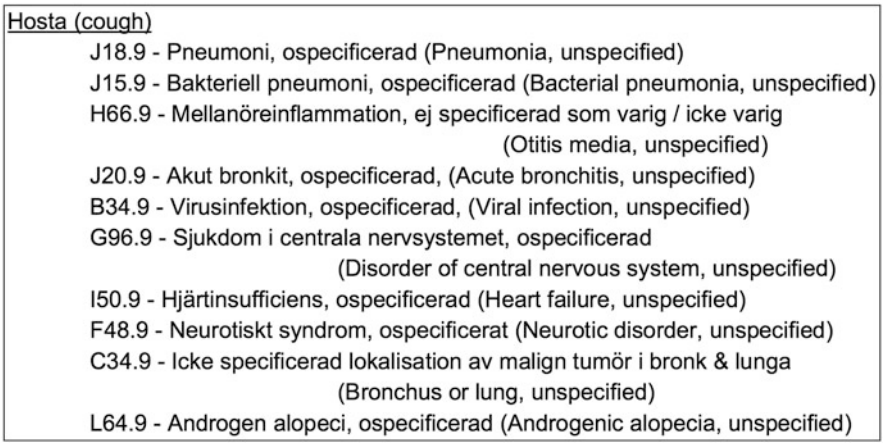

Fig. 10.8 Example of ICD-10 code suggestions (© 2009 The authors—reprinted with permission from the authors. Published in Dalianis et al. (2009))

Koopman et al. (2015a) experimented with death certificates from New South Wales in Australia to classify diabetes, influenza, pneumonia and HIV. A set of 340,142 death certificates was divided into 80% for training and 20% for a test set. All death certificates were coded with ICD-10 diagnosis codes. Both the machine learning-based method SVM (from the Weka toolkit) and a rule-based method based on keyword-matching rules were used. The keywords were selected with assistance from domain experts on words that characterise each disease. Both methods yielded very similar results with an F-score of around 0.96.

10.7.1 Natural Language Generation from SNOMED CT

SNOMED CT is a very complex hierarchical medical terminology that contains many separate pieces of information with the aim of describing a disorder, its cause, symptoms and in which body part the disorder occurs. The concepts are coded in SNOMED CT and are difficult for a human to validate. See for example the IHTSDO SNOMED CT Browser and its description of scarlet fever in Fig. 10.9,

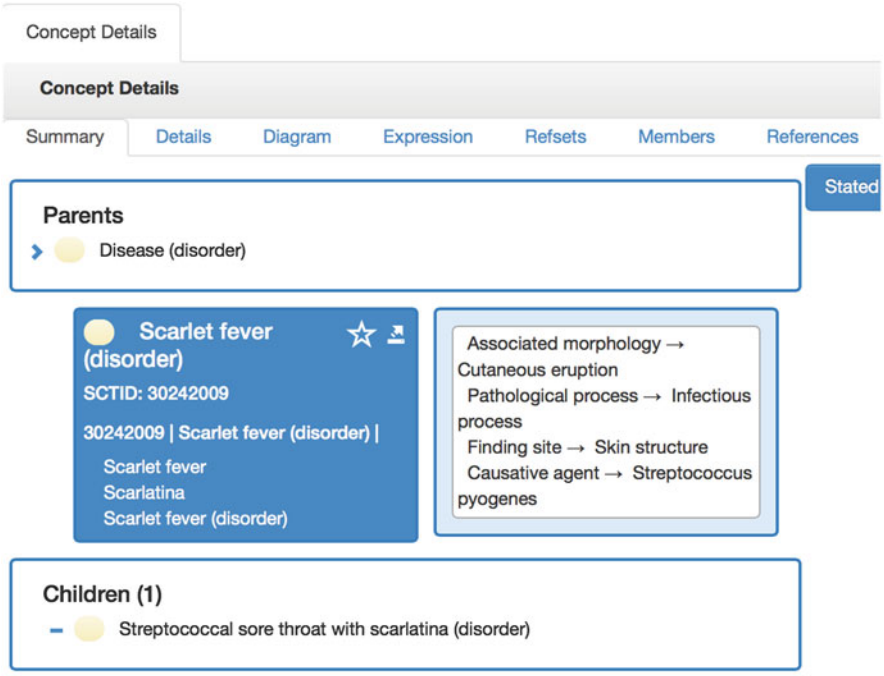

Fig. 10.9 The IHTSDO SNOMED CT Browser and its description of scarlet fever, the browser is described in Sect. 5.2

Input: *Scharlakansfeber*
Output: *Scharlakansfeber är en eruption och hudsjukdom orsakad av streptokocker. Orsaken till sjukdomen är Streptococcus pyogenes. Sjukdomen finns i hud och hudstruktur och hud.*
(Translated with Google translate to English:
Input: *Scarlet fever*
Output: *Scarlet fever is an eruption and skin disease caused by streptococci. The cause of the disease is Streptococcus pyogenes. The disease is found in skin and skin structure and skin.*

Fig. 10.10 Example of natural language output from the SNOgen system, when entering the disorder *scharlakansfeber* (in Eng: scarlet fever) in SNOMED CT format and obtaining the Swedish natural language text output. Below is the corresponding machine translated English text (© 2014 The authors—reprinted with permission from the authors. Published in Kanhov (2014))

therefore, an approach to automatically generate natural language text from the SNOMED CT structure has been carried out using a prototype called SNOGEN.

The same example describing scarlet fever in SNOMED CT format is entered into the SNOGEN natural language generator. The generated natural language text describes a certain part of the disorder in the context of the cause and where in the body it occurs. The natural language discourse makes it easier to understand and validate the description of scarlet fever expressed in SNOMED CT format, see Fig. 10.10 for an example of *natural language generation* text of Scarlet fever from SNOMED CT (Kanhov et al. 2012; Kanhov 2014).

10.8 Search Cohort Selection and Similar Patient Cases

10.8.1 Comorbidities

Comorbidity means a disorder co-occurs with other disorders as well as a disorder that may cause other disorders. If a patient obtains one disorder which other disorder or disorders are expected to occur sequentially? One application for detecting this is the visualisation tool Comorbidity-View[14] available on the Internet. Comorbidity-View is applied on part of the HEALTH BANK data, more exactly on all patient records from the Karolinska University Hospital encompassing the years 2006–2008, in total 605,587 patients; and visualises them in the form of ICD-10 codes and patients in a comorbidity network, see Fig. 10.11.

The Comorbidity-View tool is a useful tool for quickly exploring the available patient record data, in terms of patients, gender, ages and ICD-10 codes; however, one drawback is that it does not have temporal information on which disorder

[14]Comorbidity-View, https://www2.dsv.su.se/comorbidityview-demo/. Accessed 2018-01-11.

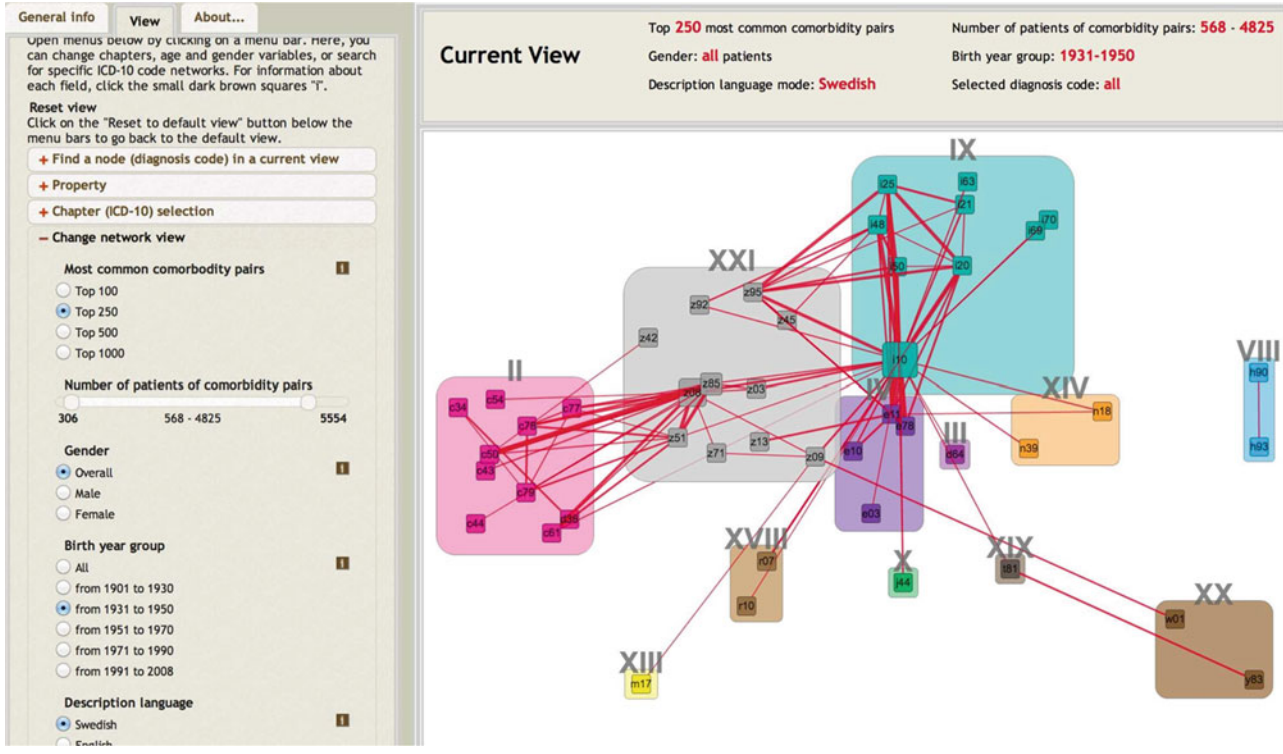

Fig. 10.11 Screenshot of Comorbidity-View, which is a visualisation tool for comorbidity networks. It contains all the disorders in ICD-10 code form that patient records have been assigned. The data contains 605,587 patients over the years 2006–2008 from the Karolinska University Hospital. The thicker the line in Comorbidity-View the more patients who had both ICD-10 codes. The boxes correspond to chapters of ICD-10 codes (Tanushi et al. 2011)

occurred first, second etc. In Comorbidity-View all disorder pairs are presented in the same temporal order.

One similar approach to detect comorbidities using text mining is presented by Roque et al. (2011b), involving Danish clinical corpus containing 5543 psychiatric patients records, their disorders and corresponding manually assigned ICD-10 diagnosis codes.

The methods used for automatically detecting ICD-10 codes were purely rule-based using lists of words from ICD-10 dictionaries, negation detection or detection of relatives of the patient.

31,662 ICD-10 codes were automatically extracted, of which 22,956 were disqualified since they either were negated or they were connected to a relative of the patient. Manual evaluation was carried out on 48 patients and a precision of 87.78% was calculated, the value of the recall is not mentioned.

The most frequent ICD-10 chapter using the manually assigned ICD-10 codes was chapter V *Mental and behavioral disorders* After text mining for other disorders and automatic assignment of ICD-10 diagnosis codes, chapter X *Diseases of the respiratory system* and chapter XIX *Injury, poisoning and certain other consequences of external causes* were more common. The comorbidity network grew significantly after the text mining and this shows that psychiatric patients often have other disorders that are not recorded in their patient record.

10.8.2 Information Retrieval from Electronic Patient Records

A number of shared tasks within clinical text mining have been carried out, for example the TREC 2013 Medical Record Track as well as the ShARe/CLEF eHealth Evaluation Lab from 2013 up to 2016. All shared tasks used electronic patient records that had been de-identified and pseudonymised. All participating teams in the shared task needed to sign confidentiality agreements due to the possible sensitivity of the patient records.

In TREC 2013, the tasks were similar to the previous Text REtrieval Conferences (TREC) using patient records as document collections. The task was constructed in the domain of *cohort* studies, to find a group of patients sharing similar properties. This was simulated by retrieving a particular topic from 50 selected topics contained in 17,264 electronic patient records in English. Each record corresponds to a patient. The records were annotated with ICD-9 diagnosis codes,[15] which also could be used to retrieve topics. Voorhees and Hersh (2012) also reported the challenge of classifying negated findings and negated disorders, which is specific to this type of text.

ShARe/CLEF eHealth Evaluation 2013 contained three different tasks using clinical text written in English (Suominen et al. 2013):

- The first task was to identify and normalise disorders and map them to SNOMED CT.

[15]ICD-9 diagnosis codes are used in the US, and are an earlier revision of ICD-10.

- The second task was to expand abbreviations and acronyms.
- The third task was a traditional Q&A task for patients, but using clinical reports that the patient may ask questions about.

10.8.3 Search Engine Solr

There is one useful and well-known search engine for indexing and retrieving documents, and (of course) electronic patient records, called Solr,[16] previously Lucene. Solr has, for example, been used in the study by Korkontzelos et al. (2012) regarding clinical trials but also in the shared task within precision medicine described in Roberts et al. (2016). The official name of the shared task in precision medicine was *TREC 2014 Clinical Decision Support Track*. Precision medicine means customised treatment to each individual patient. The task to use patient cases, described in free text, as queries to find the best treatment for the patient among 733,138 scientific articles in Pubmed Central (PMC). Pubmed is an online digital database of freely available full-text biomedical literature.

10.8.4 Supporting the Clinician in an Emergency Department with the Radiology Report

One problem with radiology reports in emergency departments is the time delay between the report from the radiologist and the clinical treatment of the patient. The first radiology report might miss limb fractures and the patient is sent home, after review of the X-rays the radiologist might discover limb fractures and the patient is then re-admitted for treatment. In an study carried by Koopman et al. (2015b) the authors used machine learning techniques to solve this problem.

2378 freetext radiology reports on limb structures from three large Australian public hospitals were studied. The reports are very short, an average only 47 words. Radiologists carried out a review of these free-text radiology reports and classified them as normal or abnormal, i.e. with some fracture on the limb. The machine learning system based on SVM in the Weka toolkit was trained on these classified radiology reports and obtained an F-score of 0.92. This gives the opportunity for the radiologist to only review 11% of the original 2378 radiology reports and to considerably diminish the time delay.

In a nice demonstrator called RadSearch, Bevan Koopman demonstrates the ability to search radiology reports and simultaneously obtain cohorts of patients with a certain medical condition, see Fig. 10.12.

[16] Solr, http://lucene.apache.org/solr/. Accessed 2018-01-11.

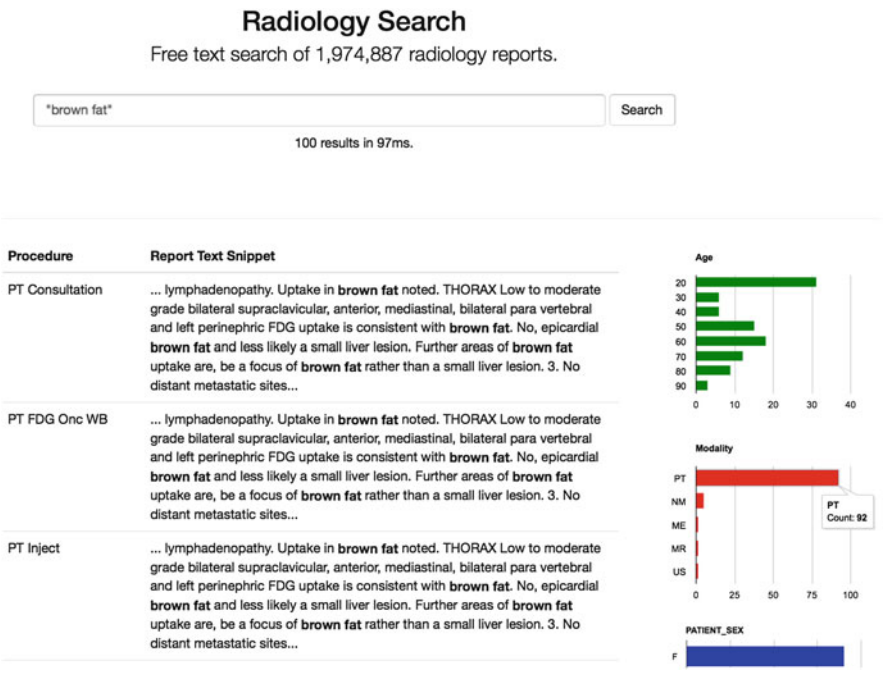

Fig. 10.12 Screenshot of Radsearch, a tool for extracting cohorts of patients with a certain medical condition in radiology. In this case the search terms are *brown fat* (The screenshot is a courtesy of Bevan Koopman)

10.8.5 Incident Reporting

Incident reporting has been mentioned previously, for example to detect healthcare associated infections, see Sect. 10.1, and adverse drug event detection, see Sect. 10.2, but there are some other types of incident reporting, for example to detect adverse events such as pressure sores, patient falls, device failures, nutrition problems, surgical complications as well as healthcare associated infections.

One way of detecting adverse events is to use the *Global Trigger Tool (GTT)* the method which is a manual method to review patient records for trigger words. According to the method there are different groups of adverse events, and within each group specific adverse events that one should review the records for. Doupi et al. (2015) describe one approach at the Karolinska University Hospital in Stockholm, Sweden, where the method was implemented as a semi-automatic method to find candidates for manual review by clinical personnel. The method is called *Modifierad Automatiserad (Modified Automatised) GTT, (MAG)*. MAG searches for triggers both in the structured information and in the free text of the patient record. The results are then reviewed and summarised by the tool. The system was developed in SAS Institute software.

For an academic article see Gerdes and Hardahl (2012) for an approach using SAS® Text Miner and SAS® Enterprise Content Categorization to find pressure ulcers (pressure injuries) in Danish patient records. The approach obtained a negative predictive value, i.e. a value for not finding indications of pressure ulcers was 97%, which is very good, but the positive predictive value for finding indications of pressure ulcers was only 56%, which is comparatively low.

In Australia an approach is used to classify incident reports into types of incidents. Around 10% of all admissions to acute care hospitals result in an adverse event and an incident report. Wang et al. (2017) performed the following study: from 137,522 submitted incident reports 6000 were randomly selected and in turn manually annotated by three experts. Part of this set was divided into a balanced subset of the following 11 incident types *falls, medications, pressure injury, aggression, documentation,*[17] *blood product, patient identification, infection, clinical handover, deteriorating patient* and *others*. This subset comprised in total 260 reports in each class in total 2860 reports, using the Support Vector Machine (SVM) algorithm gave an F-score of 0.783. Another imbalanced dataset of 5950 reports was also annotated and gave an F-score of 0.739. The inter-rater reliability for determining incident types was 0.93 Cohen's κ calculated on a small separate training set for the tree annotators. See Sect. 10.2.4 about adverse event detection systems.

10.8.6 Hypothesis Generation

One more domain is the generation and testing of new hypotheses. One example of this is presented in Dalianis et al. (2009), where the document clustering system Infomat[18] based on the vector space model was used. 4000 electronic patient records in Swedish from geriatric clinics were extracted from the Stockholm EPR corpus. The free text fields *bedömning* (assessment) and the structured entry gender were used for the clustering experiments. 62% of the patients in the corpus were women and 38% were men. One observation was that more documents describing female patients contained the words (translated from Swedish) *crutch, pelvis, femur, walkers, support, lift* and *broken bone*, than documents describing male patients, one new hypothesis is therefore that more women than men suffer from bone brittleness. There were also more men than women who had problems with memory and dementia. The hypothesis that more women than men suffers from bone brittleness was also supported after some preliminary literature studies. Figure 10.13 shows a screenshot of Infomat applied on electronic patient records in geriatrics. The "bone brittleness" words in Swedish can been seen in the middle of the figure in the boxes and these are also gathered in the diagonal line of the Infomat tool.

[17] The type "documentation", include various errors in documentation, that lead to an incident.

[18] Informat, http://www.csc.kth.se/tcs/projects/infomat/infomat/. Accessed 2018-01-11.

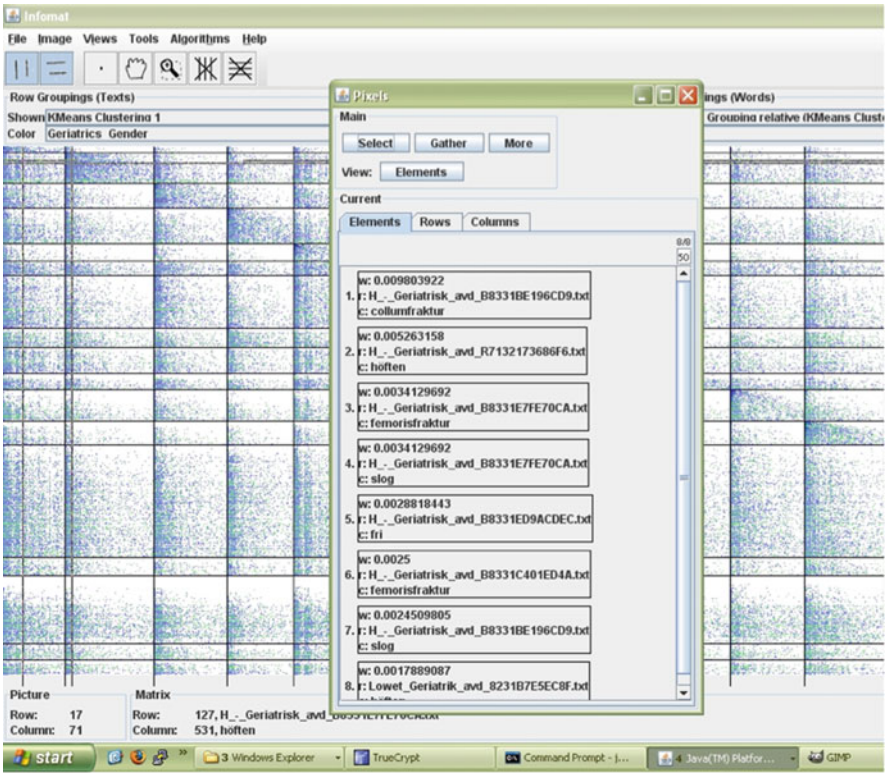

Fig. 10.13 Screenshot of hypothesis generation in geriatrics using the Infomat clustering tool. The "bone brittleness" words in Swedish such as *collumfractur, femorisfracture, höft (hip) etc.* indicating female patients can be observed in the boxes

10.8.7 Practical Use of SNOMED CT

Regarding SNOMED CT, there are two studies describing the use of the terminology (Lee et al. 2013, 2014). The first study (Lee et al. 2013) is an interview survey and the other study (Lee et al. 2014) is a literature review.

In the first study (Lee et al. 2013), called "A survey of SNOMED CT implementations", the authors contacted over 50 users of SNOMED CT in February 2012, this resulted in 14 interviews for 13 different implementation in over eight countries. The interviewees were professionals ranging from physicians, academics, clinical terminologists and software developers to vendors.

Some of the success factors for the use of SNOMED CT were: simplicity, involvement by clinicians and ease of demonstrating value and training.

The most common success factor was keeping the user interface simple for clinicians, hiding the complexity of SNOMED CT.

One implementation was in a historical electronic patient record system, where over 10 million patient records were analysed using SNOMED CT resulting in over 20,000 unique descriptions.

One other implementation was for assisting clinicians in finding new useful body site terms to be used for measuring blood pressure.

The conclusion of the study was that a lot effort and resources have been spent on developing SNOMED CT but there is still much work required to bring SNOMED CT into practical use.

In the second study (Lee et al. 2014), the authors investigated over 488 articles on SNOMED CT published between 2001 and 2012. Most articles comprised academic work mostly on a theoretical level, but there was also work on how to harmonise SNOMED CT to other terminologies and standards.

10.8.8 ICD-10 and SNOMED CT Code Mapping

There have been many attempts, mostly manual, to map from either ICD-10 to SNOMED CT or sometimes from SNOMED CT to ICD-10. The reason for the mapping is to achieve interoperability between the terminologies, or at least some degree of interoperability. The interoperability will make it possible to use SNOMED CT in systems that use ICD-10 coding, and also the other way around. Of course SNOMED CT is much more expressful and powerful than ICD-10, but also more difficult to use.

One early method was carried out by Wang et al. (2008) where their system mapped ICPC-2 PLUS to SNOMED CT with a precision of 96.46% and overall recall of 44.89%.

The mapping was performed both as string mapping including substring matching after removing stop words etc, and expanding abbreviation matching, but also mapping using WordNet synonym lexicon matching (in English), UMLS mapping, and finally post-coordination mapping, where several terms in a concept were extracted and matched partially into a SNOMED CT procedure.

ICPC-2 PLUS (also known as the BEACH coding system) is a coding system developed for primary care in Australia. It contains 7410 concepts. Wang et al. (2008) used only 5971 of these concepts to map to SNOMED CT that contains over 300,000 concepts.

Andersson and Sjöberg (2016) carried out a master's thesis work where they mapped the Swedish version of SNOMED CT to the Swedish version of ICD-10 using lexical similarity. The normalisation of the text descriptions included non-functional word (stop word) removal, stemming and decompounding using a specially built decompounder for Swedish medical words.

Andersson and Sjöberg (2016) constructed different matching algorithms including semantic similarity for SNOMED CT and ICD-10, meaning that if two concepts describe the same thing they are semantically similar. 948 ICD-10 concepts in the test set were evaluated by 10 domain experts obtaining a precision of 68.6% and a recall of 69.9%.

10.8.9 Analysing the Patient's Speech

Dementia is a gradual decline of cognitive abilities, often resulting from neurodegeneration. To detect early signs of dementia one project was carried out analysing the patient's speech. The patient's speech is recognised by a speech recogniser and then converted to text that is analysed with natural language processing methods.

In a study by Fraser et al. (2015) the DementiaBank corpus was used, which contains 167 patients diagnosed with "possible" or "probable" Alzheimer's disease. The DementiaBank corpus contains 240 narrative samples, and 97 controls sample provide another 233 speech samples. A method based on linguistic features was applied both to the pure speech and to the recognised speech in the form of text. The annotated data was used to train machine learning classifiers, such as logistic regression (LR) and support vector machines (SVM), to automatically classify patients with Alzheimer's disease and healthy patients. The approach gave an accuracy of over 81%.

The results from the research of Fraser et al. (2015) are currently being transferred into a Swedish context and adapted to Swedish dementia patients. The Swedish study will extend the speech analysis with eye movement analysis of the patients and other cognitive markers. The planned work is described in Kokkinakis et al. (2017).

10.8.10 MYCIN and Clinical Decision Support

There is a research area called clinical decision systems, these systems support the physician in his or her decision on diagnosis and treatment. The area has it origins in the expert systems and knowledge based systems of the 1980s. There has been discussion on the ethics on such systems, that is if they give the wrong decision to the physician, who is the responsible the machine or the physician?

One well-known clinical decision system is MYCIN, developed in 1970s, for detecting and giving treatment advice on blood diseases such as bacteremia and meningitis. It contained over 600 rules and outperformed the medical staff at Stanford University. One interesting feature of MYCIN was that it had an NLP interface for both questions and answers. The system also gave explanations for its reasoning (Buchanan and Shortliffe 1984).

10.8.11 IBM Watson Health

IBM Watson Health is an approach by IBM to collect a large amount of information about healthcare in the form of scientific journals, clinical trials, guidelines and textbooks as well as other clinical documents. IBM Watson lets a program index

and retrieve, or even interpret, the information and make it available for physicians as answers to their health related questions. IBM Watson Health supports natural language understanding of basic queries that are interpreted and sent to the system. IBM Watson has been used in the areas of cancer kinases and drug repositioning. Cancer kinases, or as the correct term is protein kinases, are enzymes used to treat cancer and drug repositioning is finding a new use for old drugs. The answers are provided as a list of suggestions with a ranking of the closeness or usefulness for the treatment of the patient. The physician will obtain feedback from the system in the form of explanations and the system will also request additional information, as in a dialogue. However, no information has been made available on the performance of the system (Chen et al. 2016; High 2012).

10.9 Summary of Applications of Clinical Text Mining

This chapter has presented a number of applications in clinical text mining as detection and prediction of adverse events such as healthcare associated infections (HAI), and detection of adverse drug events (ADE). Other applications presented were assignment and validation of ICD-10 diagnosis codes for patient records as well as automatic mapping of ICD-10 diagnosis codes to SNOMED CT.

Applications of summarising the patient records were presented, as well as simplification of the patient records for laypeople. An application for generation of patient records text in the neonatal intensive care area was presented.

Search and retrieval techniques for patient records were presented, specifically for cohort studies and comorbidities, along with hypothesis generation using clustering techniques. An information retrieval task, the ShARe/CLEF eHealth Evaluation task for clinical text retrieval was presented, finally classic medical decision (expert) systems such as MYCIN were briefly described, together with IBM Watson Health for assisting the physician in diagnosing the patient.