Exploration of known and unknown early symptoms of cervical cancer and development of a symptom spectrum

Outline of a data and text mining based approach∗

Claudia Ehrentraut¹, Karin Sundström², and Hercules Dalianis¹

¹Department of Computer and Systems Sciences, (DSV), Stockholm University, Sweden
ehrentraut@dsv.su.se, hercules@dsv.su.se
²Department of Medical Epidemiology and Biostatistics, (MEB), Karolinska Institutet, Stockholm, Sweden
karin.sundstrom@ki.se

March 3, 2015

Abstract

This position paper lays up the structure of some experiments to detect early symptoms of cervical cancer. We are using a large corpora of electronic patient records texts in Swedish from Karolinska University Hospital from the years 2009-2010, where we extracted in total 1,660 patients with the diagnosis code C53. We used a Named Entity Recogniser called Clinical Entity Finder to detect the diagnosis and symptoms expressed in these clinical texts containing in total 2,988,118 words. We found 28,218 symptoms and diagnoses on these 1,660 patients. We present some initial findings, and discuss them and propose a set of experiments to find possible early symptoms or at least a spectrum or finger prints for early symptoms of cervical cancer.

1. Introduction

In the last ten years patient records have become, at least in Sweden, completely digitalized and also centralised in large repositories, so called registers. This is a vast source of knowledge within medical research, however, this resource has not been much exploited.

∗DSV Report Series No. 15-005
The reason is that clinical researchers have little or no knowledge in data and text mining, and also that these repositories due to their sensitive nature are difficult to access in order to perform research.

Lately, these sources have become to a very small extent available to researchers in the U.S. as well as Europe. Meystre et al. [1], wrote a review article back in 2008, about different text mining approaches and tools, mostly for English textual data. The authors mention also among these one approach to detect early findings in breast cancer.

Among others, they mention an approach to detect early symptoms of breast cancer. Dalianis et al. [2] describes clinical text mining including extraction and retrieval specifically for use in Swedish patient records. It is timely to assess to what extent text mining can assist in the evaluation of symptoms encountered in the course of human cancer.

This paper presents the first steps taken within the MINECAN project, which aims at finding unknown early symptoms and developing a symptom spectrum for cervical cancer. Accordingly, the paper

- Outlines the problem of cervical cancer prevention and how text mining is hitherto applied in the cervical cancer domain,
- Illustrates the main aims of the MINECAN project and the approach chosen to reach those
- Presents initial experiments that have been performed and
- Provides an outlook on proposed work for the further course of the project.

2. Project framework

2.1. Cervical cancer

Cervical cancer (ICD-10-code: C.53) is one of the most common cancers worldwide [3], frequently striking young women below age 40, if not screened [4]. While cervical cancer incident and related death rates have dropped in countries where screening procedures have evolved into established prevention techniques [5, 6], the rates in especially less developed countries are still high.

Cervical cancer develops from tissues of the cervix, i.e., the lower third of the womb which is opening up to the vagina [7]. A long-term infection with the Human papillomavirus (HPV) is deemed a necessary but not sufficient factor in the development of cervical cancer [8]. Spreading via sexual contact, HPV causes more than 98% of all cervical cancer cases [9].

2.1.1. Diagnosis

Today, women are offered screening every three to five years in order to detect abnormal changes in the cells, so called dysplasia (ICD-10 code N87) [10], in an early stage where the dysplasia is amenable to treatment. The Pap test is the most commonly used screening
test for cervical cancer and pre-cancerous stages, where cells are collected from the cervix in order to be analyzed under the microscope [6]. About 50% of all cervical cancer cases are detected during screening [5].

Cancer in an intermediate or advanced stage is highly mortal. Early diagnosis is therefore crucial in order to prevent treatable pre-cancer from turning into invasive cancer [6]. Early detection is yet often hindered since not all women wish to participate in cervical screening programs.

Women who do not attend screening can be diagnosed via symptomatic presentation. However, diagnoses of cervical cancer may be delayed because of the failure to recognize symptoms as cancer-related. Lim et al. [11] conducted an interview-based study in the UK in which they aim at examining the extend and determinants of these delays. The authors could confirm earlier findings that patients below the age of 25 often delay presentation with cervical cancer and have a more advanced stage. Some reasons for the delay may be that the patients

- do not recognize possible cancer symptoms, especially vaginal discharge
- do not re-attend promptly after first presentation despite persisting symptoms.

Delays in diagnosis do also occur on behalf of the health care provider who may fail to recognize cervical cancer-related symptoms.

2.1.2. Symptoms

According to the state of the art assumption, women with early cervical cancers and pre-cancers usually have no symptoms [10, 6]. But it is possible that there are blood value deviations or other unforeseen symptoms. In most cases, the symptoms do not start until the cancer has reached a more advanced stage. Usual gynecological symptoms at that point are [11, 6]:

- Abnormal vaginal bleeding
- Unusual discharge from the vagina
- Pain during intercourse.

Lim et al. [11] promote earlier presentation of gynecological symptoms but do point out the problem that these symptoms are very common in young women and that promoting early presentations for these may lead to increased visits in primary care as well as unnecessary anxiety in females who do not have cancer.

Increasing the awareness of (early) cervical cancer symptoms among women and health care staff might improve diagnostics and chance of survival [12]. Finding hitherto unknown early symptoms which may appear during a pre-cancerous stage could further help to diagnose cervical cancer at a time when it is still treatable.
2.1.3. Cervical cancer and text mining

Spasic et al. [13] reviewed different approaches for clinical text mining within the cancer domain. Of all studies the authors refer to, only two have focused on cervical cancer and HPV, respectively.

The study focusing on cervical cancer aims at finding a method for retrieving oncolgy documents relevant to clinical decision within the particular domain of cervix cancer. Polpinij et al., [14], propose an ontology-based text analysis to be used where traditional search by using keywords and Boolean logic cannot support clinicians to find and assimilate all the literatures relevant to their research and clinical decision. With a content-based text classification process and similarity analysis at its core, their system obtains its highest accuracy at 92%.

The study focusing on HPV aims at discriminating high-risk HPV types, i.e., those that are related with cervical cancer, from low-risk types, i.e., those that are not related with it. In their paper, Park et al., [15], propose a cost-sensitive learning method to classify the risk types of HPVs using their textual explanation. The authors pre-processed the textual descriptions by performing stop word filtering and TF-IDF ranking.

Comparing three learning algorithms, namely AdaCost, AdaBoost and Naïve Bayes, the authors showed that AdaCost outperforms the other algorithms, yielding an accuracy of circa 93% and F-score of about 87%

2.2. MINECAN project

2.2.1. Project background

It has been shown in an interview-based study that young females with cervical cancer frequently delay presentation for examination by doctor, and not recognising symptoms as serious may increase the risk of delay [16]. Improved identification and awareness of early signs of cervical cancer may reduce both patient and provider delay of investigation and treatment. Thus, it could be highly valuable to establish the cervical cancer symptom spectrum and whether there are additional symptoms that should be added to this. Text mining could as a novel tool aid with a bias-free search of words and biochemical features that may not have been previously suspected/identified by patients or health care.

The hypothesis in this project originates from the assumption that women with early cervical cancers and pre-cancers usually have no symptoms [10, 6]

So far, symptoms of a disease are mainly retrieved/collected by means of capturing descriptions made by the patient spontaneously, or after being questioned by a health care professional. However, few of these are relevant for registration in national health registers. Thus, traditional register-based research cannot access such data.

2.2.2. Project aims

The MINECAN project has two major aims:

1. Determine whether there are unknown early symptoms of cervical cancer, and if so which. This to potentially inform health care and screening processes of symptoms
in women that may be of note. The anticipated output is to find unknown early
symptoms of cervix cancer. In this regard, a list of concrete symptoms is considered
to be the desired finding.

2. Develop and characterize a symptom spectrum for cervical cancer through a holis-
tic description of symptoms as recorded in medical text by health care staff. Such
a spectrum would include both previously known, and potentially unknown symp-
toms. The anticipated output is an evidence-based, holistic description of cervix
cancer symptoms, i.e., likelihood of occurrence, time of occurrence and frequency
of occurrence of diverse symptoms. Ideally, the symptom description will be an
interactive visualization, as for instance depicted in Figure 1. This serves the pur-
pose of generating a better understanding of possible cervical cancer symptoms due
to their potentially ambiguous nature.

The purpose of both aims is to obtain a more concise understanding of symptoms that
occur in cervical cancer patients compared to non-cancer patients, based on evidence that
is gained through a statistical analysis of a large amount of medical data. We intend to
approach these aims by applying and enhancing state of the art text mining tools.

The overall goal is to use our findings as a complement in screening programs for
cervical cancer. In addition to taking a Pap test\(^1\), the physician would be able to run
a program to filter out the patients symptoms, if captured in the medical record and
compare them to a list of possible early cervical cancer symptoms. Ideally, this approach
should be generic in order to be applicable to other cancers.

2.2.3. Flow of screening patients at Karolinska University Hospital

The gynecological clinics at the Karolinska University Hospital perform cervical smears
either as a follow-up from organized cervical screening, or when a woman presents directly
at the hospital with gynecological symptoms.

\(^1\)http://en.wikipedia.org/wiki/Pap_test (Papanicolaou test)

Figure 1: Symptoms spectrum
In terms of the screening program, women are invited every 3-5 years depending on age group. If a screening smear is abnormal, i.e. has cytological changes in the cervical cells, the woman is referred to a gynecological procedure called colposcopy, which is inspection of the cervix and typically accompanied by a biopsy if an abnormality on the cervix is noted. Colposcopy can be done through out-patient clinics contracted by the screening authority, and sometimes these clinics can also perform the follow-up operation termed conization which is used to treat pre-cancerous lesions, depending on in which catchment area the woman lives. In some instances, women are also diagnosed and treated for precancerous changes directly at gynecological clinic at the Karolinska University Hospital, if this is the area where she lives.

If the investigation of the abnormal smear ultimately reveals invasive cervical cancer, women are referred by the outpatient clinics to a Karolinska Hospital-organized multi-disciplinary conference where cytologists, gynecologists, oncologists and others convene to discuss each individual case. There, it is decided whether the woman should start treatment at the gynecological section (i.e. operation) or oncological section (i.e. radiation and/or chemotherapy) first. Continued treatment and flow between clinics is dependent on each woman’s cancer stage, prognosis, treatment outcome and other health issues etc.

3. Data

The researchers of the MINECAN project and this particular study have access to the Stockholm Electronic Patient Record (SEPR) Corpus that comprises patient records from 2009 to 2010 from in-patient clinics at Karolinska University Hospital in Stockholm, Sweden. All patient records are anonymized and kept on an encrypted server at the Department for Computer and Systems Sciences, Stockholm University in Kista, Sweden. The corpus contains records from all units at Karolinska University Hospital except for records from the psychiatric and venereal disease unit. For the MINECAN project, a subcorpus is created from the SEPR Corpus as described in Section 4.2.

4. First experiments

In order to approach the main goal of finding unknown early symptoms and creating a symptom spectrum for cervical cancer, the initial work comprised the construction of part of the subcorpus and initial experiments performed on that corpus.

The approach used for this project resembles a retrospective case-control study. That means past medical records are used to identify exposure and outcome factors, e.g., potential exposures/symptoms for the outcome cervical cancer. The study comprises a group of interest (study group) and a comparison (or control) group.³

³This research has been approved by the Regional Ethical Review Board in Stockholm (Etikprövningsnämnden i Stockholm), permission number 2014/1882-31/5
³http://hsl.lib.umn.edu/biomed/help/understanding-research-study-designs
4.1. Defining the study group

The study group data consists of records that belong to patients diagnosed with cervical cancer. These patients are identified as having cervical cancer if an appropriate ICD-10 diagnosis code is found in their records. All cervical cancer related ICD-10 codes were specified by the project’s medical expert and are listed in Table 1.

<table>
<thead>
<tr>
<th>Type of cervix cancer</th>
<th>ICD-10 code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Malignant neoplasm: Endocervix</td>
<td>C53.0</td>
</tr>
<tr>
<td>Malignant neoplasm: Exocervix</td>
<td>C53.1</td>
</tr>
<tr>
<td>Malignant neoplasm: Overlapping lesion of cervix uteri</td>
<td>C53.8</td>
</tr>
<tr>
<td>Malignant neoplasm: Cervix uteri, unspecified</td>
<td>C53.9</td>
</tr>
<tr>
<td>Carcinoma in situ: Endocervix</td>
<td>D06.0</td>
</tr>
<tr>
<td>Carcinoma in situ: Exocervix</td>
<td>D06.1</td>
</tr>
<tr>
<td>Carcinoma in situ: Other parts of cervix</td>
<td>D06.7</td>
</tr>
<tr>
<td>Carcinoma in situ: Cervix, unspecified</td>
<td>D06.9</td>
</tr>
<tr>
<td>Severe cervical dysplasia, not elsewhere classified</td>
<td>N87.2</td>
</tr>
</tbody>
</table>

4.2. Construction of the subcorpus and basic statistics

The SEPR Corpus is stored in a database. Ultimately, the subset that is created from this corpus for the cervical cancer project will comprise records that belong to the study as well as as control group. As part of the first experiments, only data for the study group has been extracted. Defining and extracting data for the control group will be done at a later point in time.

For the study group, the following information is extracted from the database using MySQL queries:

- Gender and age of patient
- Date of patients’ admission to and discharge from hospital
- Clinic where patient is treated
- Daily note (free text) and corresponding date of entrance into hospital system

Once the data is extracted, all information about the patients is saved into a text file with one file per patient, containing patient number, age and gender information as well as all the patients’ daily notes sorted by date. These files are then used for further processing and analysis.

Statistics for the study group were obtained according to the following parameters: age, clinic, time of diagnosis, length of treatment.

In total, 1660 patients are contained in the study group. Of these patients 1587 patients have obtained only one ICD-10 diagnosis code, i.e., a C53, D06 or N87 code. 72 patients
have had two diagnosis code in their records, 42 patients had C53 and D06 diagnosis code in their records while for 29 patients, D06 and N87 co-occurred in the records. 0 patients had a C53 and N87 co-occurring in the record. For one patient, all three diagnosis codes occurred in the record. Of the 1587 patients who only had one diagnosis, 603 had a C53 diagnosis code, 955 a D06 code and 29 a N87 code.

4.3. Named Entity Recognition and frequency counting

The following section describes an initial approach of generating a frequency list of symptoms captured in records of patients that were assigned a C53 code on a small subset of the data.

4.3.1. Method

This method aims at identifying symptom words in patient records, extract them from the records and sort them according to their frequencies.

Ultimately, this step will be carried out for the records of the study group and the control group, resulting in two frequency lists, a cervical cancer list and a control list. The two frequency lists will be compared to one another to see

- if and how the symptoms differ between cases and controls

\footnote{Only using C53 codes and no D06 and N87 codes is motivated by the fact that we want to start testing...}
• if well-known cervical cancer symptoms are identified most frequently in the cervical cancer list or

• if there are other symptoms that occur more frequently

• whether our methods can accurately identify a priori known/suspected associations, which should validate whether the methodology is appropriate

As part of these first experiments, an initial cervical cancer frequency list was created in a two step process.

• Identify all symptoms by using the tool Clinical Entity Finder (CEF)

• Extract, sort and count all found symptoms and save them into a frequency list

The Clinical Entity Finder (CEF) by Skeppstedt et al. [17], implements the idea/task of Named Entity Recognition (NER), i.e., recognizing expressions denoting entities such as symptoms, diseases, body parts and drugs in clinical text documents.

NER is a traditional task of recognizing Personal names, Organisations, Locations and Time points, but are now also used as a description of how to identify for example also clinical entities such as symptoms and diseases. NER can be performed automatically and over the past years multiple NER algorithms have been implemented. NER modules for English are for instance available via the Stanford CoreNLP\(^5\) package or Apache OpenNLP\(^6\).

Skeppstedt et al. [17] have implemented the Clinical Entity Finder that can automatically recognize entities in narrative text of Swedish health records. The tool is based on CRF++, an implementation of the conditional random fields algorithm, and is initially implemented to detect the terms within the entity categories Disorder, Finding, Pharmaceutical Drug and Body Structure. The training material comprises patient records from a Swedish internal and medicine emergency unit that were manually annotated. 1,988 disorders, 3,681 findings, 735 body structures and 1,542 drugs instances, in total 7,946 annotated instances distributed over 70,852 tokens.

After running the Clinical Entity Finder, the detected cervical cancer symptoms are sorted, counted and saved into a frequency list that is depicted in Table 2. The table depicts the 10 most frequent symptoms in patient records that contain one of the four ICD-10 diagnosis codes C53.0, C53.1, C53.8 and C53.9. The entire frequency comprises 28,218 symptoms and diagnoses in total 2,988,118 tokens of text. The 100 most frequent ones can be found in Appendix A.


Table 2: Frequency list for C53.0, C53.1, C53.8 and C53.9

<table>
<thead>
<tr>
<th>Term frequency</th>
<th>Term</th>
<th>Engl. translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>4252</td>
<td>smärta</td>
<td>pain</td>
</tr>
<tr>
<td>3338</td>
<td>illamående</td>
<td>nausea</td>
</tr>
<tr>
<td>1895</td>
<td>blödning</td>
<td>bleeding</td>
</tr>
<tr>
<td>1735</td>
<td>opåverkad</td>
<td>unaffected</td>
</tr>
<tr>
<td>1714</td>
<td>besvår</td>
<td>trouble</td>
</tr>
<tr>
<td>1650</td>
<td>mår bra</td>
<td>feel well</td>
</tr>
<tr>
<td>1528</td>
<td>ont</td>
<td>ill</td>
</tr>
<tr>
<td>1498</td>
<td>smärtor</td>
<td>pains</td>
</tr>
<tr>
<td>1495</td>
<td>feber</td>
<td>fever</td>
</tr>
<tr>
<td>1472</td>
<td>trött</td>
<td>tired</td>
</tr>
</tbody>
</table>

4.3.2. Qualitative analysis of two pre-annotated patient records

We applied the Clinical Entity Finder (CEF) to data from a different domain. To provide an estimate on how well CEF works within the new domain, one member of the group conducted a qualitative analysis of two pre-annotated patient records, by manually reading through them and checking whether the symptoms were annotated correctly.

It was found that CEF, which is trained on data from the internal and medicine emergency domain, failed to detect some cervical cancer related terms. While cervix, cervix cancer, as well as the abbreviations cervixca., skivepitelca. were missed, cancer and cervixcancer were correctly detected. In the two files, 9 respectively 14 symptoms (findings + disorders) were negated, indicating the absence of those particular symptoms. To sum up, CEF yielded promising results, missing only 3 to 6 percent of the symptoms in the two records. In total, there were 458 respectively 373 annotations per record, containing 10,169 and 6,089 tokens each. Our results on data from the radiology and oncology domain were in line with the ones obtained for the internal and medicine emergency domain, i.e., where they obtained an F-score of 81% and 69% for finding and disorder respectively, as described in Skeppstedt el al. [17].

4.3.3. Initial findings

We identified several restrictions and drawbacks that have to be handled in order to obtain a representative frequency list of cervical cancer symptoms.

- Multiple inflectional forms of the same word, such as smärta (Engl.: pain) and smärtor (Engl.: pains), occur in the frequency list. Using lemmatization, they should be reduced to their base form in order to only include the main symptom concept in the frequency list.

- So far the frequency list contains symptoms which are negated and that should be removed from the list. Negation detection will need to be applied in order to filter out these symptoms.
• Since we are interested in early symptoms, mainly daily patient notes that are added to the EHR before the cancer diagnosis are of interest. So far we used all patient notes that exist in the EHR for a patient with a cervical cancer diagnosis. A future task aims at using only these notes when detecting symptoms and generating frequency lists from them.

• Identifying symptoms by applying CEF yielded promising results. Yet CEF should be adapted to the domain by using more domain relevant training data and incorporating negation.

5. Proposed work

This paper described the first steps towards finding unknown early symptoms and building a symptom spectrum for cervical cancer. As the projects progresses we plan to work on the tasks presented in Sections 5.1 to 5.4.

5.1. Defining and extracting the control group

So far, only the study group has been defined and extracted. A crucial task in the future will be to define the control group, i.e., specify according to which parameters, such as age, clinic and calendar year, the control group should be aligned to the study group. One important parameter for the control group is that the subjects have not been diagnosed with cervical cancer or pre-stages of cervical cancer. As a subsequent step, the data of the control group will need to be extracted from the database.

5.2. Primary methods

Primary methods serve the purpose of finding/obtaining any possibly relevant symptoms. The proposed methods are

• Advance NER and frequency counting approach
• Named Entity Recognition and Random Indexing
• Clustering

5.2.1. Advance NER and frequency counting approach

We have applied NER and frequency counting in a first attempt (see Section 4.3.1), in order to develop a list of cervical cancer symptoms that appear in the EHRs. As described in Section 4.3.3, we identified several restrictions and drawbacks that have to be handled in order to obtain a representativ frequency list of cervical cancer symptoms. Taking these initial findings into account, this approach will be advanced in the future.
5.2.2. Named Entity Recognition and Random Indexing

Random Indexing refers to an incremental word space model that was developed by Sahlgren [18]. The basic idea behind it is to accumulate context vectors based on the occurrence of words in contexts. It can be applied to the patient records in order to extract diagnosis-symptom pairs. That means, by inserting diagnosis codes for cervical cancer and its pre-cancers, which are depicted in Table 1, the system can filter out symptoms that occur in the records of a patient with that specific diagnosis code.

Random Indexing can and has been applied to Swedish patient records to calculate co-occurrences of words and ICD-10 codes on a document level [19]. Prior to running the Random Indexing tool, the Clinical Entity Finder will be applied to detect symptoms in the patient records.

5.2.3. Clustering

Yet another idea to find unknown early symptoms is to use clustering. Clustering aims at finding structure in data, thus being exploratory in nature. For decades, clustering has been applied within various fields, with the first clustering algorithm, K-Means, dating back to 1955. By now, a vast number of different clustering algorithms exists and subpart of this project will be to find the most appropriate one for our task. [20]

Infomat\(^7\), a vector space visualization tool aimed at Information Retrieval and text clustering in particular, provides a number of different clustering algorithms and could be used for the task. Using clustering could ideally reveal clusters of symptom words, which differ depending on whether the patient records they appear in belong to the study or control group.

5.3. aDEX

We plan to use a tool called aDEX, constructed within the DADEL-project by Isak Karlsson, that can find comorbidities related to cervix cancer ICD-10 codes, as specified in Table 1. The aDEX-demo can been seen here\(^8\).

5.4. Secondary (preprocessing methods) methods

The secondary methods serve the purpose of increasing the (retrieval results). Methods to be applied are:

- Lemmatization
- Negation and certainty detection
- Temporality
- Mapping symptoms to ICD-10 diagnosis codes

\(^7\)http://www.csc.kth.se/tcs/projects/infomat/infomat/[2014-11-19]
\(^8\)https://drive.google.com/file/d/0B5E5Nh0dI65GevMmazVxcV12Ex/view?usp=sharing
These four indirect methods will be tested in combination with the various direct methods. Further indirect methods may be added as the project proceeds.

5.4.1. Lemmatization

Lemmatization is a method to reduce inflected words to their lemma or base form. For example, the inflected words blödning, blödningar (Eng.: bleeding, bleedings) are reduced to blöda (Eng.: bleed). We plan to use the CST lemmatizer [21], which is adapted to Swedish. Lemmatization has been used extensively in information retrieval for highly inflectional languages and has improved recall, in most cases without decreasing precision, for Swedish and other such languages [22].

Lemmatizing the patient journals before generating the frequency list is important to avoid redundant terms, i.e., various inflected versions of the same word.

5.4.2. Negation and certainty detection

"In clinical reports the presence of a term does not necessarily indicate the presence of the clinical condition represented by that term. In fact, many of the most frequently described findings and diseases in discharge summaries, radiology reports, history and physical exams, and other transcribed reports are denied in the patient" [23]

As pointed out by the citation, negations occur frequently in clinical text. For reasons of documentation and clarity, sometimes also to protect against subsequent legal action, it can be vital for a clinician to note that certain signs, symptoms or risk factors are not present/applicable in the individual patient case. Identifying symptoms that are negated is therefore a crucial task if one wants to obtain a frequency list that represents the records it is based on properly. For this project’s purposes, negated cervical cancer symptoms will be removed from the frequency list. Negation detection is the technique to identify negated entities, such as symptoms, disorders or drugs. NegEx is a negation detection system for clinical text that has been adapted to Swedish [24] and that could be used in this project.

Similar to negation, medical text is determined by uncertain expressions, such as strongly suggests, possible or likely, that reflect the diagnostic alternatives or difficult judgements due to uncertain, incomplete and complex data that practitioners are faced with in clinical medicine [25]. As part of the future work, it has to be identified, whether uncertain expressions do occur in combination with symptoms. If this is the case, it further has to be discussed, how uncertain symptoms should be handled, i.e., should they be removed from, ranked or kept in the symptoms list.

In case uncertainty detection becomes relevant in addition to negation detection, the pyConTextSwe algorithm, which is based on the assertion algorithm for English pyConTextNLP, could be utilized to identify uncertain and negated terms in the patient records [26].
5.4.3. Temporality

In order to decide whether a certain symptom may be an early symptom of cervical cancer, it is crucial to know at what point in time the symptom occurs. Does it refer to a symptom which the patient encountered before seeking help at a health facility, occur in the beginning of a hospitalization or emerge just one day before the cancer is diagnosed.

The electronic patient records in our dataset that are used in this task comprise daily notes written by physicians and nurses. These contain timestamps which specify when a certain note was entered to the record. This timestamp is however not sufficient to indicate the point of time at which the symptom occurred. The textual context is deciding for when the symptom set in. For example if the patient note contains an entrance such as "The patient experienced heavy bleedings yesterday" on October 27th 2010, the correct time of the symptom bleeding would be October 26th 2010.

In this regard, a temporal analysis of the context in which a symptom occurs needs to be performed. Temporal analysis has been the topic in several natural language processing (NLP) challenges and multiple systems have been developed for different languages [27]. Yet little research has been performed on Swedish clinical text. Velupillai [27] has conducted an initial study on automatic identification of temporal expressions, such as igår (Engl.: yesterday), in Swedish clinical text by translating an adapting the HeidelTime system. A promising precision of 92% is obtained, which renders the future development of a temporal reasoning system for Swedish that could be applied for the purpose of this study possible.

5.4.4. Mapping symptoms to ICD-10 diagnosis codes

Another technique that will be tested is to map the extracted symptoms to ICD-10 diagnosis codes. The classification is divided into 22 chapters from I to XXII, whereat the division is based on diseases of body systems, special diseases or external factors. Chapter XVIII is used to code Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified. The idea is to map the symptoms to these ICD-10 symptom codes, in order to group the symptoms to a main concept symptom

6. Possible limitations

One limitation may be that Aim 1, finding previously early and/or unknown symptoms of cervical cancer, cannot be fulfilled. However, this in turn could actually inform health care practice and confirm the current evidence base for cervical cancer as a relatively symptom-less disease, demonstrated by systematically exploiting a novel data source; medical records. Regardless of Aim 1, our Aim 2 should provide valuable information on the symptom spectrum in cervical cancer.

It is well known that manual ICD-10 code assignment is error prone. Estimations carried out by the National Board of Health and Welfare (Socialstyrelsen) in Sweden concluded that the errors in the main diagnoses was 20 percent [28].
6.1. Additional ICD-10 diagnosis codes

Other early codes that might be relevant are for example N93.9 - Abnormal uterine and vaginal bleeding, unspecified B97.7 Papillomavirus as the cause of diseases classified elsewhere. (HPV virus is the main cause of cervical cancer and can therefore been seen as early symptoms)

7. Conclusion

This paper has outlined the current state of the art within cervical cancer prevention and how text mining is hitherto applied in the cancer domain. Further, this paper presented (1) initial experiments that have been performed as well as (2) an outlook on proposed work in order to find unknown early symptoms and develop a symptom spectrum for cervical cancer.

We believe that outlining the scope of the project, including aims, state-of-the-art research, proposed future work and limitations, as well as performing initial experiments was crucial for enabling a stringent work flow in the project.

Acknowledgements

The authors would like to thank the Nordic Center of Excellence in Health-Related e-Sciences (NIASC) and Nordforsk for funding of the project and to Eric Thuning and Per "Pelle" Olofsson both IT experts at DSV for help with the management of the Stockholm EPR Corpus server. We would also like to thank Maria Skeppstedt for letting use use the Clinical Entity Finder and for Aron Henriksson for assisting us in executing Clinical Entity Finder.

References


A.

Cervical cancer frequency list

4252 smärta
3338 illamående
1895 blödning
1735 opåverkad
1714 besvär
1650 mär bra
1528 ont
1498 smärtor
1495 feber
1472 trött
1401 orolig
1252 blödningar
1224 opåverkat
940 god effekt
888 frisk
871 mjuk och oöm
793 smärtan
709 svullnad
680 värk
629 oro
625 rodnad
618 mjuk
589 diarréer
572 biljud
560 diarré
543 sår
515 palp ua
505 svullen
493 buksmärtor
472 smärtlindrad
461 trötthet
454 ledsn
449 afebril
446 recidivfri
422 klåda
419 hosta
416 cirkulatoriskt stabil
412 ömhet
409 sveda
408 kräkningar
402 buksmärta
381 yrsel
378 tamponad
377 tumörsuspekt
369 huvudvärk
354 såret
350 vulva ua
349 palpabla resistenser
334 nedsatt aptit
332 välmående
326 god diures
312 mår väl
304 vagina ua
303 lös avföring
289 frossa
285 fri vätska
271 smärfteri
269 pleuravätska
268 bi- eller bläsljud
266 bläsljud
262 palp resistenser
260 palperas ua
249 god aptit
242 kräkning
240 försämring
234 vaken
225 pigg
222 tryck
222 smårtorna
220 tumör
219 smårtproblematik
218 kräkts
214 mjuk, oöm
212 obehag
211 ödem
211 utslag
208 infektionstecken
200 svullna
196 domningar
196 cirkulatorisk stabil
195 opåverkad i vila
194 stabil
194 andningsbesvär
192 mår bättre
189 piggare
187 illamåendet
185 god blodretur
182 vulva u a
182 cirk stabil
180 progress
179 respiratoriskt stabil
176 smärt päverkad
176 ronki
176 magerlagd
175 corpus resistens
172 svaghet
172 rassel
172 hematuri
172 förstorade lymfkörtlar
170 opåv