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 Ehrentraut[1], Karin Sundström[2], Hercules Dalianis[1]

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

Abstract. This position paper delineates 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 patient records with the ICD-10 diagnosis code C53 for cervical cancer. 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 and/or a spectrum of early symptoms of cervical cancer.

1 Introduction and Motivation

In the last ten years patient records have become, at least in Sweden, completely digitalized and also centralised in large repositories. This is a vast source of knowledge within medical research, however, this resource has not been much exploited. 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. [9], wrote a review article about different text mining approaches and tools, mostly for English textual data. Among others, they mention an approach to detect early symptoms of breast cancer. Dalianis et al. [3] 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.

It has been shown in an interview-based study that young females with cervical cancer frequently delay presentation, and not recognising symptoms as

serious may increase the risk of delay [8]. 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 [14, 1]. So far, symptoms of a disease are mainly 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.

The 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 cervical 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 holistic description of symptoms as recorded in medical text by health care staff. Such a spectrum would include both previously known, and potentially unknown symptoms. The anticipated output is a holistic description of cervical 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 purpose 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,

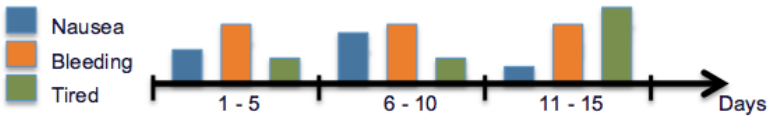

Fig. 1. Proposed visualisation of cervical cancer symptoms spectrum, the height of the bar depicts the number of symptoms. One possibility is also to show number of negated symptoms, or absence of symptoms as negative bars.

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 screening test for cervical cancer, the physician could for example be able to run a program to filter out the patient's 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.

This paper intends to outline the current state of the art within cervical cancer prevention and how text mining is hitherto applied in the cancer domain. Further, this paper presents (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.

2 Background

Cervical cancer (ICD-10 diagnosis code: C.53) is one of the most common cancers worldwide [2], frequently striking young women below age 40, if not screened [15]. A long-term infection with the Human papillomavirus (HPV), which spreads via sexual contact, is deemed a necessary but not sufficient factor in the development of cervical cancer [16].

Today, women are offered screening every three to five years, with the Pap test being most commonly used, in order to detect abnormal changes in the cells in an early stage. 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 [1]. 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. As Lim et al. [7] found, some reasons for the delay may be that the patients (1) do not recognize possible cancer symptoms, especially vaginal discharge, and (2) do not re-attend promptly after first presentation despite persisting symptoms. Delays in diagnosis do also occur on behalf of the provider who may fail to recognize cervical cancer-related symptoms.

According to the state of the art assumption, women with early cervical cancers and pre-cancers usually have no symptoms [14, 1]. Yet, 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 (1) abnormal vaginal bleeding, (2) unusual discharge from the vagina and (3) pain during intercourse [7, 1].

Increasing the awareness of (early) cervical cancer symptoms among women and health care staff might improve diagnostics and chance of survival [6]. 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.

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 aimed at finding a method for retrieving oncology documents relevant to clinical decision within the particular domain of cervix cancer. With a content-based text classification process and similarity analysis at its core, their system obtains its highest accuracy at 92% [11].

The study focusing on HPV aimed 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. Comparing three machine 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% [10].

3 Materials and Methods

The researchers of the MINECAN[1] project and this particular study have access to the Stockholm Electronic Patient Record (SEPR) Corpus that comprises patient records from 2006 to 2014 from Karolinska University Hospital in Stockholm, Sweden, [4]. 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[2] is created from the SEPR Corpus.

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].

3.1 ICD-10 diagnosis codes

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. They are:

[1] MINECAN - Data and text mining of cancer symptoms and comorbidities in electronic patient records in the Nordic languages

[2] This research has been approved by the Regional Ethical Review Board in Stockholm (Etikprövningsnämnden i Stockholm), permission number 2014/1882-31/5

[3] http://hsl.lib.umn.edu/biomed/help/understanding-research-study-designs

- C53.0 (Malignant neoplasm: Endocervix)
- C53.1 (Malignant neoplasm: Exocervix)
- C53.8 (Malignant neoplasm: Overlapping lesion of cervix uteri)
- C53.9 (Malignant neoplasm: Cervix uteri, unspecified)
- D06.0 (Carcinoma in situ: Endocervix)
- D06.1 (Carcinoma in situ: Exocervix)
- D06.7 (Carcinoma in situ: Other parts of cervix)
- D06.9 (Carcinoma in situ: Cervix, unspecified)
- N87.2 (Severe cervical dysplasia, not elsewhere classified).

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(s) where patient is treated
- Daily note (free text) and corresponding date of entrance into hospital system during the years 2009-2010

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.

3.2 Statistics of study group

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

In total, 1,660 patients are contained in the study group. Of these patients 1,587 patients have obtained only one ICD-10 diagnosis code, i.e., a C53, D06 or N87 code. 72 patients have had two diagnosis codes in their records, 42 patients had C53 and D06 diagnosis codes in their records while for 29 patients, D06 and N87 co-occurred in the records. No patients had a C53 and N87 co-occurring in the record. For one patient, all three diagnosis codes occurred in the record. Of the 1,587 patients who only had one diagnosis, 603 had a C53 diagnosis code, 955 a D06 code and 29 a N87 code.

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[4] on a small subset of the data.

[4] Only using C53 codes and no D06 and N87 codes is motivated by the fact that we want to start testing

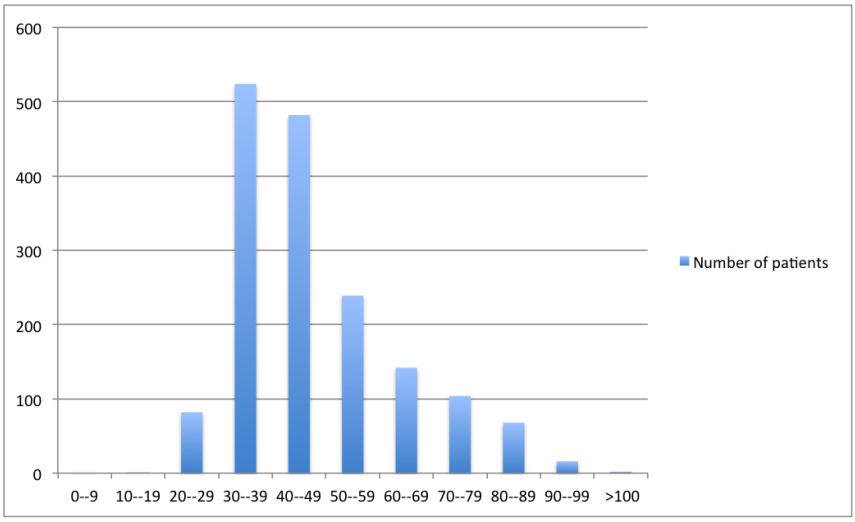

Fig. 2. Age statistics, IDC-10 diagnosis codes C53, D06 and N87.

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 done 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
- 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, implements the idea/task of Named Entity Recognition (NER), i.e., recognizing expressions denoting entities such as diseases, drugs, or people's names in free text documents [9]. This task 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. [12] have

[5] http://nlp.stanford.edu/software/corenlp.shtml. 2014-09-08.
[6] https://opennlp.apache.org/index.html. 2014-09-08.

Table 1. Frequency list for cervical cancer records assigned ICD-10 codes C53.0, C53.1, C53.8 and C53.9

Term frequency	Term	Engl. translation
4252	smärta	pain
3338	illamående	nausea
1895	blödning	bleeding
1735	opåverkad	unaffected
1714	besvär	trouble
1650	mår bra	feel well
1528	ont	ill
1498	smärtor	pains
1495	feber	fever
1472	trött	tired

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*.

After running CEF, the detected cervical cancer symptoms are sorted, counted and saved into a frequency list that is depicted in Table 1.

4 Results

Table 1 depicts the 10 most frequent symptoms in patient records that contain one of the four ICD-10 codes C53.0, C53.1, C53.8 and C53.9. The entire frequency comprises 28,218 symptoms.

We applied the Clinical Entity Finder, CEF, trained on annotated data from one domain to a different domain. To provide an estimate om 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 per record.

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 those notes made before diagnosis, 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 Discussion

During our research work we encountered some challenges. We are not yet at the stage of identifying any early unknown symptoms of cervical cancer but are able to succesfully confirm other known symptoms such as *bleeding* that is a possible symptom of cervical cancer.

Some of the symptoms we identified were actually negated symptoms as *not bleeding*, findings that our system could not identify as negated findings/symptoms, since we did not use any negation detection system. Some of the symptoms which are enumerated in Table 1. are therefore negated.

Findings that are in singular or plural form as *bleeding* or *bleedings* could be reduced to one base form using a lemmatizer. The same approach can be carried out for determined and non-determined form of nouns. Determined nouns in Swedish uses a inflection *en* to change to determined form; *blödning+en =>* *blödning***en**, instead of a modifier as in English, *the bleeding*. Reducing these identical findings would make the analyse easier and increase precision.

Another obstacle was temporality, the patient record stretches over several months or years and we need a method to identify when something occurred. Certainly we have time stamps on each note, but within each note the physician sometimes refer to earlier findings and relate to them.

Regarding identifying terms we saw that there are many non-standard words and abbreviations and compounds of abbreviations and words, as for example, *cervixca.*, that CEF could could not identify as named entities. This could easiest be solved by adding in-domain annotated data.

6 Conclusions and Further 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 following tasks:

- Defining and extracting the control group

- Testing and advancing the following methods to identify symptoms captured in the patient records:
  - NER and frequency counting approach
  - Named Entity Recognition and Random Indexing
  - Clustering
- Using and adapting existing text mining tools for the domain and incorporating them into the preceding methods:
  - Lemmatization
  - Negation and certainty detection
  - Temporality
  - Mapping symptoms to ICD-10 codes
- Analyzing and assembling the results as well as designing a visual representation for the developed symptoms spectrum.

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.

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.

Our methodology can also been seen as a part of the HEALTH BANK workbench proposed in [5], that will offer processed aggregated and unaggregated clinical data for research to be used in a wider context.

Acknowledgements

The authors would like to thank the Nordic Information for Action eScience 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 us use the Clinical Entity Finder and for Aron Henriksson for assisting us in executing Clinical Entity Finder.

References

1. American Cancer Society, A.: Cervical Cancer Prevention and Early Detection (2014), http://www.cancer.org/acs/groups/cid/documents/webcontent/003094-pdf.pdf
2. Cancer Research UK, U.: Worldwide cancer incidence statistics, http://www.cancerresearchuk.org/cancer-info/cancerstats/world/incidence/Common, visited: November 13th 2014
3. Dalianis, H.: Clinical text retrieval - an overview of basic building blocks and applications. In: Paltoglou, G., Loizides, F., Hansen, P. (eds.) Professional Search in the Modern World, vol. 8830, pp. 147–165. Springer Verlag, Lecture Notes in Computer Science (2014)
4. Dalianis, H., Hassel, M., Henriksson, A., Skeppstedt, M.: Stockholm EPR Corpus: A clinical database used to improve health care. In: Swedish Language Technology Conference. pp. 17–18 (2012)
5. Dalianis, H., Henriksson, A., Kvist, M., Velupillai, S., Weegar, R.: HEALTH BANK – A Workbench for Data Science Applications in Healthcare. In: Proceedings of CAiSE'15 – Industry Track. Springer Verlag, Lecture Notes in Computer Science (2015)
6. Swedish Council on Health Technology Assessment, SBU, S.: Tidig upptäckt av symtomgivande cancer - En systematisk litteraturöversikt, (In Swedish), (January 2014), http://www.sbu.se/upload/Publikationer/Content0/1/Tidig_upptackt_symtomgivande_cancer_smf.pdf
7. Lim, A.W., Ramirez, A.J., Hamilton, W., Sasieni, P., Patnick, J., Forbes, L.J.: Delays in diagnosis of young females with symptomatic cervical cancer in england: an interview-based study. British Journal of General Practice pp. e602–e610 (October 2014)
8. Lim, A.W., Forbes, L.J., Rosenthal, A.N., Raju, K.S., Ramirez, A.J.: Measuring the nature and duration of symptoms of cervical cancer in young women: developing an interview-based approach. BMC women's health 13(1), 45 (2013)
9. Meystre, S., Savova, G., Kipper-Schuler, K., Hurdle, J.: Extracting information from textual documents in the electronic health record: a review of recent research. IMIA Yearbook of Medical Informatics 47, 128–144 (2008)
10. Park, S.B., Hwang, S., Zhang, B.T.: Mining the risk types of human papillomavirus (HPV) by AdaCost. In: Mařík, V., Retschitzegger, W., Štěpánková, O. (eds.) Database and Expert Systems Applications. Springer (2003)
11. Polpinij, J., Miller, A.: Ontology-based text analysis approach to retrieve oncology documents from PubMed relevant to cervical cancer in clinical trials. In: ICDM Workshop on Advances in Data Mining. IBaI Publishing, Leipzig (2010)
12. Skeppstedt, M., Kvist, M., Nilsson, G.H., Dalianis, H.: Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study. Journal of Biomedical Informatics 49, 148–158 (June 2014)
13. Spasić, I., Livsey, J., Keane, J.A., Nenadić, G.: Text mining of cancer-related information: Review of current status and future directions. International journal of medical informatics 83(9), 605–623 (2014)
14. Storck, S.: Cervical dysplasia. Online (2014), http://www.nlm.nih.gov/medlineplus/ency/article/001491.htm, medlinePlus
15. Sundström, K.: Human Papillomavirus Test and Vaccination - Impact on Cervical Cancer Screening and Prevention. Ph.D. thesis, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden (2013)

16. Walboomers, J.M.M., Jacobs, M.V., Manos, M.M., Bosch, F.X., Kummer, J.A., Shah, K.V., Snijders, P.J.F., Peto, J., Meijer, C.J.L.M., Muñoz, N.: Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. The Journal of Pathology 189(1), 12–19 (1999)