UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

Corso di Laurea Magistrale in Ingegneria Informatica

Tesi di Laurea Magistrale

Entity Extraction and Linking for Digital Pathology

Relatore
Prof. Gianmaria Silvello

Riccardo Galiazzo
matricola 1150141

5 Dicembre 2022 (A.A. 2022/2023)

Ringraziamenti

Voglio ringraziare prima di tutto il mio relatore, il Professor Gianmaria Silvello, che con la sua enorme pazienza e disponibilità mi ha permesso di svolgere questa tesi.
Vorrei anche ringraziare i miei amici Fabio Giachelle e Roberto Bodo, che senza il loro aiuto questa tesi non sarebbe stata la stessa.
Inoltre ringrazio il mio psicologo Stefano Cardullo che mi ha dato la spinta mentale per svolgere la parte finale di questa tesi.
Infine ringrazio tutti i miei parenti, i miei amici, i parenti dei miei amici e gli amici dei miei parenti per avermi supportato in questo periodo.

Vorrei dedicare questa tesi a due persone a me molto care, che sono scomparse in questi ultimi mesi di università.
La dedico alla mia cara nonna, Maria Mimo, che mi ha adottato e cresciuto come se fossi un suo figlio.
E soprattutto la dedico al mio caro fratello e purtroppo il mio ex migliore amico, Marco Galiazzo, che mi ha mi ha quasi fatto da figura paterna.

Summary

In recent years, huge amounts of biomedical data have been produced. The rich information content of such data could be exploited for several purposes including diagnostics and supporting the medical decision-making process. Nevertheless, most of this information is stored to date using unstructured formats, as occurs, for instance, for free-text narrative clinical reports and clinical notes saved in Electronic Health Records (EHRs). Hence, these documents are human-readable but not machine-readable. Despite some Laboratory Information Systems (LISs) support structured data and synoptic reports, the adoption of structured and machine-readable formats is still limited. This poses hindrances to the full exploitation of computational approaches for data analysis, pattern recognition, and any other secondary use in general. To mitigate this, knowledge extraction methods could be used to automatically extract meaningful information from biomedical textual data provided in natural language. In this thesis, we tackle the former issues by investigating the application of different knowledge extraction techniques for free-text clinical reports coming from the digital pathology domain. Firstly, we manually defined curated ground truths containing all the relevant information extracted from a set of clinical reports. Secondly, we implemented several state-of-the-art techniques for knowledge extraction. Then, we evaluated the performance of such knowledge extraction algorithms against the ground truths. From the analyses conducted, it emerges that the effectiveness of knowledge extraction algorithms depends on the variability of the pathology reports examined and on the kind of entities to extract. Hence, most of the algorithmic approaches considered in our analyses obtain different results that varies significantly in terms of precision and recall.

Sommario

In anni recenti, enormi quantità di dati biomedici sono stati generati grazie all'avanzamento della tecnica e delle tecnologie digitali associate. Queste grandi moli di dati contengono informazioni preziose che possono essere sfruttate per diversi scopi, tra cui la diagnostica e il supporto al processo decisionale in ambito medico. Tuttavia, l'archiviazione di queste informazioni è avvenuta, per la maggior parte, adoperando formati non strutturati. Ad esempio, i rapporti clinici prodotti in linguaggio naturale sono archiviati spesso nelle cartelle cliniche elettroniche utilizzando uno schema libero. Di conseguenza, questi documenti sono umanamente comprensibili, ma non sono processabili direttamente in maniera automatica dagli elaboratori. Nonostante alcuni sistemi informativi di laboratorio supportino dati strutturati, l'adozione di formati strutturati automaticamente comprensibili per gli elaboratori è ancora limitata. Ciò ostacola il pieno sfruttamento degli approcci computazionali per l'analisi dei dati, il riconoscimento di schemi ricorrenti e riutilizzo dei dati per altri scopi. Per mitigare questo problema, è possibile impiegare metodi di estrazione automatica volti ad estrapolare informazioni significative a partire da dati testuali espressi in linguaggio naturale. In questa tesi, si affrontano i problemi precedentemente esposti attraverso l'applicazione di molteplici tecniche di estrazione di informazioni con riferimento ai referti clinici provenienti dal dominio della patologia digitale. In primo luogo, sono stati definiti manualmente i dati empirici ritenuti veri per il campione di referti clinici di interesse. In secondo luogo, sono state implementate diverse tecniche note in letteratura per l'estrazione automatica di informazioni. Infine, è stata valutata l'efficacia di tali algoritmi rispetto ai dati empirici di riferimento. Dalle analisi condotte, emerge che l'efficacia del processo di estrazione è molto dipendente dalla variabilità dei referti clinici esaminati e dalla tipologia delle entità che si vogliono estrarre. Di conseguenza, si evince che la maggior parte degli approcci algoritmici esaminati possono ottenere risultati molto variabili in termini di misure di valutazione quali precisione e richiamo.

Contents

1 Introduction 1

2 Background 3
- 2.1 History of the digital pathology . . . 3
  - 2.1.1 Histopathology . . . 3
  - 2.1.2 Digital pathology . . . 5
  - 2.1.3 Computational pathology . . . 7
- 2.2 ExaMode . . . 8
- 2.3 Analysed diseases . . . 9
  - 2.3.1 Cervix cancer . . . 11
- 2.4 Semantic web . . . 12
  - 2.4.1 Ontology explanation . . . 14

3 Project development 19
- 3.1 Ontology creation . . . 19
  - 3.1.1 Ontology Construction . . . 19
  - 3.1.2 Ontology description . . . 20
- 3.2 Ground Truth generation . . . 23
  - 3.2.1 Conversion rules . . . 23
  - 3.2.2 Rule Generation . . . 24
  - 3.2.3 Database implementation . . . 28
  - 3.2.4 RDF generation . . . 36
- 3.3 Automatic results generation . . . 39
  - 3.3.1 Entity extraction . . . 39
  - 3.3.2 Entity merge . . . 45
  - 3.3.3 Entity linking . . . 51

4 Results Analysis 55
- 4.1 Evaluation methods . . . 55
- 4.2 Results . . . 57
- 4.3 Analysis of the results . . . 70

5 Conclusions 73

A Full ontology 75

B Full result 79

Bibliography 101

Chapter 1

Introduction

In the context of automatic information extraction from texts and visual contents, the Semantic Web gives the chance to process information according to a formal reasoning based on ontological rules.
In particular, the European ExaMode project [5] proposes a use case regarding the information extraction from medical records diagnosing cancer and celiac diseases, having the most impact on the human quality of life.
This work aims to devise a method to automatically extract information from a collection of textual medical records (in natural language). Such information is formatted under form of machine-readable tags, according to the Resource Description Framework (RDF) format [26].
The workflow comprises the manual synthesis of ontological rules starting from the diagnosis information contained in the records.
Then, an automatic methods to generate such rules are created and evaluated against the manual extracted rules (ground truth).
From the result of the evaluation it is inferred that the rules automatically extracted are heavily dependent from the source of the diagnosis, giving a high variable performance.

The reminder of the thesis is as follows:

- In chapter 2 is present the subject of the study, which is digital pathology, with it evolution trough time.
  It is also shortly presented the use case focuses, which is the diseases study in the ExaMode project or each disease, with an in-depth analysis with the disease present in this thesis, that is cervix cancer.
  Also is presented briefly a bit of background on the techniques and technologies used.

- In chapter 3 is presented the workflow of this thesis, from the definition of the ontology, the building of the ground truth of the statement in a manual

way and the automatic way of extraction of the statement to compare with the manual ones.

- In chapter 4 both manual and automatic methods used to generate the ontological rules are evaluated and their comparison is made.
- In chapter 5 the final conclusions are discussed, together with possible extensions and future work.
- In Appendix A is placed a big figure of the ontology used in ExaMode, that is too big to put in the corpus of this thesis.
- In Appendix B is placed a lot of data of result, that it was too much to put in the corpus of this thesis.

Chapter 2

Background

In this chapter, we introduce the concepts and the knowledge necessary to profitably understand the technical contents of the present thesis.
First of all, we present briefly the historical evolution of digital pathology, its key purposes, as well as the recent advancements in computational pathology.
Moreover, we provide a short description of the ExaMode European project and the related diseases considered in the study. Finally, we report the background concerning semantic Web technologies (e.g., RDF) and ontologies.

2.1 History of the digital pathology

From histopathology, which uses histology techniques, new disciplines such as digital pathology and computational pathology have emerged.
In this section, we summarize histology techniques, the digital pathology contributions and the future developments of computational pathology.

2.1.1 Histopathology

Histopathology (from the union of the Greek words histos = tissue, pathos = suffering, -logia = study of) studies the tissues that are manifesting some disease, by visualizing them through optical microscopes.
The sample of tissue analyzed, also known as *specimen*, is extracted by means of interventions such as biopsies, surgeries or autopsies.
Then, the specimen is analyzed using histology techniques such as *staining*, that is widely used for pathological diagnosis studies. Briefly, a generic histological procedure consists of the following key steps [11]:

Fixation is a practice necessary to preserve the morphology and chemical composition of the tissue analyzed from autolysis and putrefaction. Moreover,

it is used to facilitate processing and advancements concerning the rest of the procedure.

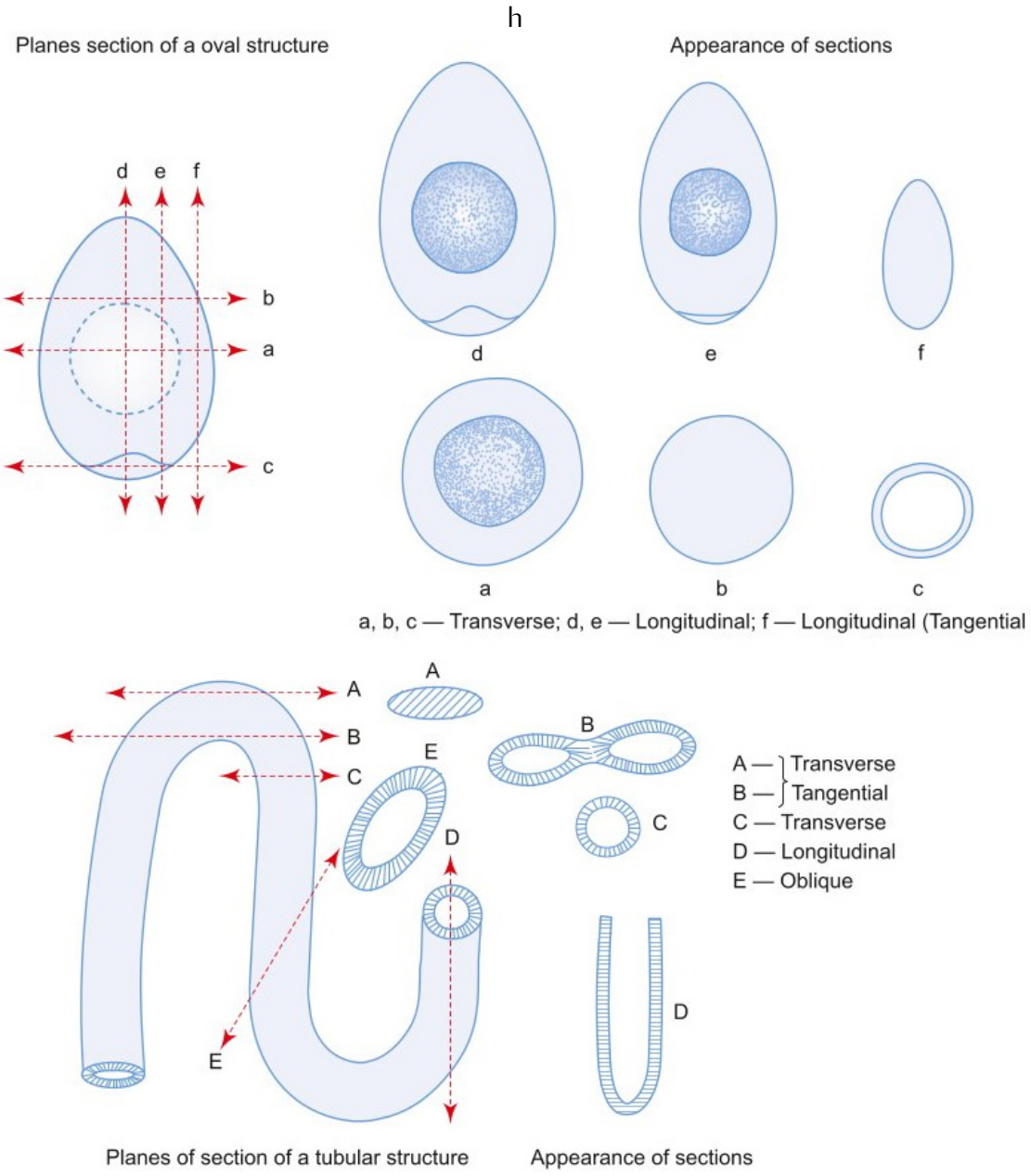

Figure 2.1. Appearance of sections of oval and tubular structures in various planes [11].

Embedding After a dehydration and cleaning, the tissue is infiltrated with embedding medium which gives a rigid consistency to the tissue.

Section Cutting To analyze a tissue specimen through a microscope, the specimen needs to be dissected in $5 - 7\mu m$ thick sections with a rotary microtome, that is, a machine designed to slice tissues.
This section gives a bi-dimensional figure originated from a three dimensional tissue. For an in-depth analysis, the section cutting procedure is repeated several times as shown in Figure 2.1.

Staining After cutting the tissue in *slides*, each slide is stained so that regions of interest are colored to simplify the identification of the different tissue's components.

Microscope Finally, the stained tissue's slide is analyzed through a microscope, providing different levels of magnification depending on the type of the microscope in use.

The slice, obtained with the histology procedure described above, is unique and indivisible, thus multiple analyses cannot be conducted on a single slide at the same time. Moreover, it is worth noting that the histological procedure is time-consuming and require several chemical reagents and materials.

2.1.2 Digital pathology

Nowadays, the advent of electronics, information technologies, and their recent advancements opened new possibilities for diagnostics and the biomedical domain in general. For what concerns histopathology, in recent years it is possible to visualize the scanned glass slides, originated from optical microscopes, directly to computer monitors [35].
In this context, digital pathology emerged as a brand new discipline applying digital techniques for the examination of histopathological images. The principal technique proposed is the Whole Slide Imaging (WSI) analysis, which enables the inspection and analysis of digitized histology slides at different levels of magnification also on specific areas of interest.
Specifically, the WSI workflow is set after the end of the traditional histology workflow and consists of the following steps:

Scanning To digitise a glass slide a scanner is employed.
To achieve high quality digital slide images, the scanning methodologies adopted for capturing images are (i) tile-by-tile, that produces several small patches for the original image (as occurs for instance in a mosaic) or (ii) using a traditional in-line scanning fashion (as occurs for textual documents).

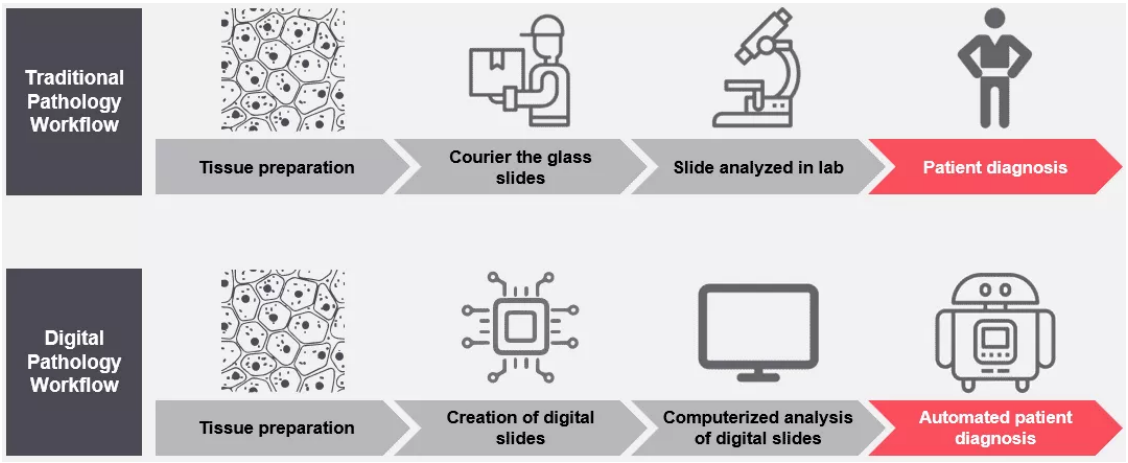

Figure 2.2. Histopathology and digital pathology workflow [8].

By combining several images' patches it is possible to assemble the digital image of the entire slide. In addition, since a tissue section is cut in multiple thin slides, it is possible to analyze the tissue also through the z-axis, which corresponds to the depth of the tissue.
WSI tools allow pathologists and experts to navigate the images in an interactive manner.

Storage To store the images, the strategy used is heavily dependent on the intended use.
The local storage is sufficient if the application require a low number of users and with no need for retention.
Otherwise, if the retention is important, methods to ensure reliable archival are recommended as well as complete backup strategies, like off-site storage, RAID storage, optical/tape storage, or some combination.
If the application require a multitude of users, network-based solutions are recommended, such as a cloud-based or (local) server-based network access with a reliable backup strategy.

Display digitalization practices coupled with WSI allow to enhance the visualization and analysis experience of glass slides images. This kind of analyses is also supported by a large number of tools, like image viewing systems or image management systems.
This tools are able to include digital annotation, rapid navigation/magnification, and computer-assisted viewing and analysis.
Many WSI systems offer software, that can be installed locally on user

computers, to view the image in the Web browsers through put the software on network servers.
Another functionality are the use of algorithms that can detect cells, compute positive staining, perform regional segmentation.
With the use of the image management systems is possible to to organize and access images using image metadata, patient information, or some other characteristic.

So, with a image representation of the slide in a digital domain, is possible to process enables microscopic images of tissue to widespread at multiple users at once and analyze the images with advanced digital tools.
The possible use of the WSI are for the example slide archive, remote consultation, telepathology, in-line scanning, tumor board, education and research.

2.1.3 Computational pathology

The WSI of tissue sections contain a high level of information that includes color, tissue morphology, cell morphology, and complex cell phenotypes, where the latest can be rich of information concerning the pathology of the tissue or the related diseases.
To analyze these complex tissue and cellular phenotypes, a pathologist need to develops specific expertise through years of training and experience, obtained from evaluating case studies and participating in peer review sessions, but the interpretation by pathologists can be altered from inherent cognitive and visual biases.

To mitigate this problem, computer vision algorithms are exploited to detect features thought the correlation of a particular imaging target or disease state. It is possible to employ supervised machine learning techniques, to achieve the automatic discovery of patterns in images to make predictions and derive additional insights. Other strategies take advantage of unsupervised machine learning techniques, which try to identify natural divisions in a data sets without requiring a ground truth of reference in advance.

With a larger adoption of WSI, a higher amount of digital tissue data are becoming available, so that machine learning and artificial intelligence methods can be trained and evaluated over such data, as occurs for deep learning based methods such as convolutional neural networks. The latter methods combine traditional computer vision approaches with modern ML optimizations, so that the algorithm can choose both the intermediate features and the learning applied to those features within a single model.
For the time being, deep learning based techniques have been used for the following tasks: (i) image segmentation; (ii) object classification and recognition;

(iii) and clinical outcomes prediction.
The end goal of artificial intelligence techniques exploiting computer vision, ML, and deep learning approaches for computational pathology, is to support pathologists with useful computer-assisted diagnosis tools, that try to automatize diagnostics and the medical decision-making process.

2.2 ExaMode

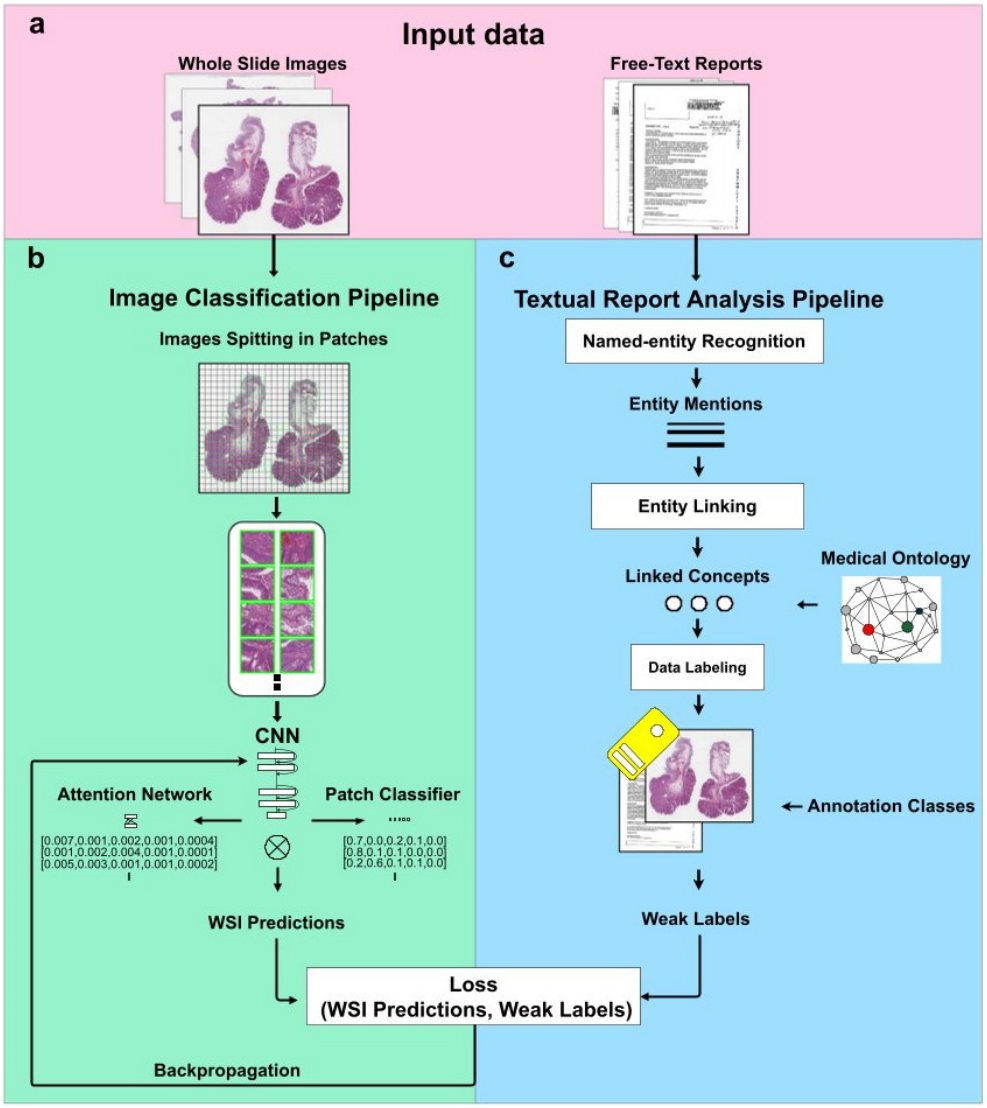

Figure 2.3. ExaMode project workflow [16]

The Extreme-scale Analytics via Multimodal Ontology Discovery & Enhancement (ExaMode) project [5] [16] is co-financed by the European Commission under the Horizon 2020 framework, and its aim is to provide automatic and semi-automatic methods to improve the effectiveness of the diagnoses of illnesses, thus facilitating pathologists' tasks [16].

The focus of ExaMode is on the diagnosis of histopathological tissues with the aim of detecting four relevant disease, which are colon cancer, cervical cancer, lung cancer and celiac disease.

The objectives of the project are centered in: weakly-supervised knowledge discovery for exascale medical data, developing extreme scale analytic tools for heterogeneous exascale multimodal and multimedia data, and to provide to the healthcare industry an extreme-scale analysis and prediction tools.
Nowadays, computer-aided diagnostic tools are based on predictions made by data-driven algorithms. Such algorithms need to be trained on huge collections of annotated data, typically consisting of Whole Slide Images (WSIs) and histological textual reports and analyses.

However, the annotation process is still an expensive and time-consuming task. For this reason, ExaMode is developing automatic methods to extract pathological concepts from the medical records to annotate the WSIs that will be used to train weekly-supervised algorithms supporting the decision-making process in the pathology domain.
In order to achieve automatic annotation, a specific ontology designed within the ExaMode project models technical entities, terminologies, and other aspects concerning the use cases of interest, including diagnosis, anatomical location, the procedure employed to obtain the tissue and the specimen to be analyzed, and the tests performed on the specimen.

2.3 Analysed diseases

In the following section, a brief description of the four diseases covered by the ExaMode ontology, their current scientific relevance, and how the AI systems built using ExaMode can help prevent and help to diagnose them.
In the ExaMode project are being studied the following diseases:

Lung cancer As indicated by the International Agency for Research on Cancer (IARC) [12], the number of lung cancer in both genders and at all ages is estimated to sharply increase by 72% from 2018 to 2040.
Lung cancer is the second most common cancer in both men and women and it covers about 13% of all new cancers diagnoses.
Despite the large number of deaths caused by lung cancer, some people with early-stage lung cancer can be successfully treated, thus is crucial to

perform an early diagnosis.
The most common type of screening for lung cancer are regular chest x-rays and Low Dose Computed Tomography scans (LDTC).
As of today, there are no official recommendations for a screening program for lung cancer in Europe. The American Cancer Society (ACS)[21] instead recommends yearly lung cancer screening with LDCT for people with high lung cancer risk [17].
There are two common types of lung cancer: non-small cell lung cancer (NSCLC) in the 85% of diagnoses, and small cell lung cancer (SCLC) for the remaining 15%. Moreover, NSCLC is divided into: squamous cell (epidermoid) carcinoma (25-30%), adenocarcinoma (40%), and large cell carcinoma (10-15%) [2].

Colon cancer The number of colon cancer in both genders and at all ages is estimated to sharply increase by 75% from 2018 to 2040, so the ACS [21] recommends regular screening for colon cancer for people over 45 years.
At this stage the screening process does not include histopathological examination, but stool-based molecular tests or visual exams can be done to diagnose the disease.
Nevertheless, the number of cases that need further investigation or confirmation of initial findings by histopathological analysis will raise due to the increasing of reported cases.
Cancer detection in biopsies is not very difficult for pathologists, but providing the required huge collection of tested samples is very time-consuming and has a substantial impact on the physicians' workload.
In such context, computer-aided colon cancer diagnosis could be an interesting and effective application field.

Celiac disease The celiac disease (CD) is an immune-mediated disease, with the chronic outcome and genetic predisposition to an intolerance to gluten and its proteins.
In such disease gluten ingestion leads to chronic inflammation, alterations, and damage in the small intestinal mucosa.
CD affects near 1% of people worldwide and its prevalence has significantly increased over the past 20 years [15].
The increasing trend of new reported cases is partly due to better diagnostics and screening of individuals at high risk for the disorder, but there could still be more undiagnosed cases of CD than undiagnosed ones [6].
CD testing is usually only recommended for people at a higher risk of developing this disease, in particular those with a family history of reported cases.
Intestinal biopsies are always necessary if the found CD-specific antibodies are low or negative, and if there are no signs/ symptoms of malabsorption.

A further biopsy may be necessary if there is no clinical improvement after shifting to a strict gluten-free diet.

2.3.1 Cervix cancer

Cervical cancer or Uterine Cervix Carcinoma is the fourth most diffused cancer in women, and the eight most commonly occurring cancer type. The estimated number of cervical cancer cases is predicted to increase by 27% until 2040, among at all ages [12].
Nearly all reported cases of cervical cancer are associated with Human Papilloma Virus (HPV) [1] [34].

Cervical cancer derives its name by the interested anatomical parts, in this case the uterine cervix. It develops in near 85% of all cases as Squamous Cell Carcinoma while in the remaining 15% as Cervix Adenocarcinoma. The disease survival rate is near 70% and women most affected are at age between 55 and 65 years [13].
According to the National Cancer Institute (NCI) [14], modifications on cervix lesions can the classified in to two classes Squamous Intraepithelial Lesion (SIL) or Cervical Intraepithelial Neoplasia (CIN) and according to their severity in low-grade or high-grade.

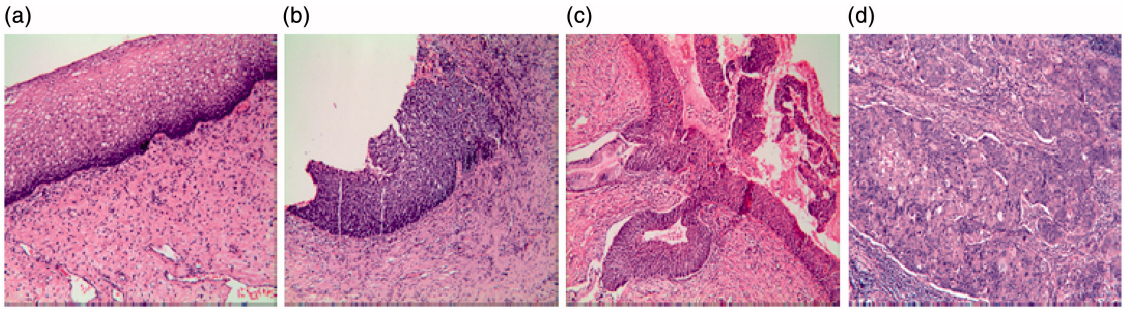

Figure 2.4. The CIN grading image samples. (a)Normal, (b) CINII, (c) CINIII, (d) Cervical Cancer. [27]

The developed malignant cancer instead can be classified into four stages:

- first stage: cancer cells are located only inside the uterine cervix;
- second stage: the cancer is extended in the upper part of the vagina.
- third stage: the cancer is extended also in the lower part of the vagina or the pelvic wall, bleeding and local pain can occur;

- fourth stage: the tumor affects also the rectum and the bladder, or other far parts of the body.

Among the risk factor, the main is represented by the infection of HPV, which represents a necessary but not sufficient cause to the development of the tumor. Sexually transmitted diseases account for an additional risk factor.
To prevent the cervix carcinoma, HPV vaccination and screening can be done in the pre-tumor phase.
Two types of screening test are available: the PAP test and the HPV test.
The Papanicolaou (Pap) test consists in a sampling of cervix cells and their visual inspection under an optical microscope to seek for abnormalities.
The HPV test instead consists in a comparison with the HPV's DNA by using the same sampling technique of the Pap test.
The disease can be treated by surgery, chemotherapy, and radiotherapy.
Surgery is suitable in the first stages, when the tumor has not extended into other parts and consists in the mechanical removal of cancer tissues.
The chemotherapy instead uses antitumoral drugs to block the diffusing of the tumor into other body parts by killing the cancer cells. Anyway, healthy cells in the proximity could be damaged.
Radiotherapy, often accompanied by chemotherapy, uses a source of ionizing radiation to kill the cancer cells. Ionizing sources can be external (rays) or internal (needle electrodes).

So, the screening (by analyses or imaging) is currently a recommended measure that should be considered in order to lower the risk of dying from cancer, whether it is lung, colon, or cervical cancer.
To meet such need, the histopathological diagnosis will become more often a final examination undertaken to unequivocally discriminant between cancerous and non-cancerous lesions found during the screening.

2.4 Semantic web

The Semantic Web is an information network made of different types of data, all identified by a Unique Resource Identifier (URI) [25].
In Computer Science, the Semantic Web is regarded as an extension of the World Wide Web through standards set by the World Wide Web Consortium (W3C). The main goal of such extension is to make contents available on Internet readable from a calculator [32].
This enables to create of a unifying framework that allows data to be shared across different sources, and to be processed automatically by discovering possible new relationships among pieces of data [25].
The collection of Semantic Web technologies (RDF, OWL, SKOS, SPARQL, etc.)

provides an environment where application can query that data, draw inferences using vocabularies, etc [32]. Common applications of Semantic Web include [24]:

- *Data integration*, whereby data in various locations and various formats can be integrated in one seamless application.
- *Resource discovery and classification* to provide better domain specific search engine capabilities.
- *Cataloging* for describing the content and relationships available at a particular Web site, page, or digital library;
- *Content rating*.
- To describe collections of pages that represent a single logical "document".
- To manage the intellectual property rights of Web pages (example, the Creative Commons).

The Semantic Web, on the implementation side, is referred to the formats and technologies that enable it. These technologies are specified as W3C standards, according to the architecture of Figure 2.5.

In the lowest level of the stack, the unique identifier and the resource format describe the resource with a common set of codes, like the characters of an alphabet.
The syntax level, based on the Extensible Markup Language (XML) aggregates such codes to form structured data, like words. This level is similar to the grammatical analysis.
The Data Interchange, Taxonomies, Ontologies and Rules levels and sub-levels enable to combine the structured data in statements and rules.
This process is comparable to the logical analysis of text.
The unifying Logic is responsible for merging queries, rules and the underlying ontology in a unified inference unit.
The proof level instead provides a validation of the previous logical inferences by a demonstration step.
Finally, the trust level supports the semantic correctness of conclusions by verifying that the premises come from trusted sources and the inferences are based

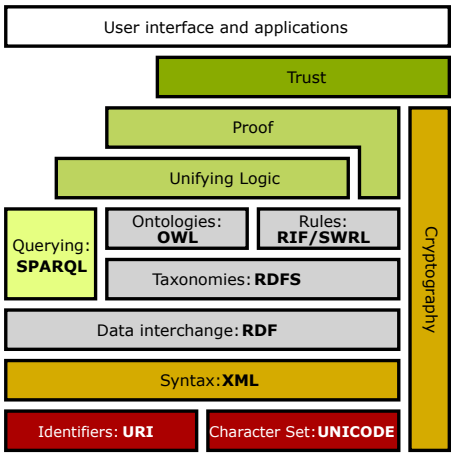

Figure 2.5. Semantic web stack [32]

on a consistent formal inference process.
For example, by reading a newspaper article, a semantic web system is able to quantify the trustness of the read information by evaluating the trustness of its references and by analyzing the presence of logical fallacies in the text.

2.4.1 Ontology explanation

On a philosophical basis, ontology [30] studies concepts such as existence, being, becoming, and reality. It includes the questions of how entities are grouped into basic categories and which of these entities exist on the most fundamental level.
In computer science and information science, an ontology [31] enables to correlate the different parts of discourse, such as names, categories, properties, and relations between the concepts.
It defines a set of concepts and categories that represent the subject and its attributes.
An ontology is a set of ontological relationship, called statement, each composed by a triple subject-predicate-object, e.g.:

$$
\begin{aligned}
Wood &\xrightarrow{\text{is an}} Object \\
Door &\xrightarrow{\text{is an}} Object \\
Stationery &\xrightarrow{\text{is an}} Object \\
Pen &\xrightarrow{\text{is a}} Stationery \\
Pencil &\xrightarrow{\text{is a}} Stationery \\
Pen &\xrightarrow{\text{is made with}} Wood \\
Door &\xrightarrow{\text{is made with}} Wood
\end{aligned}
$$

Subjects, predicates and objects are called ontological entities.
The subject is the ontological entity describe by the discourse while the object is the descriptor, attribute or properties of the object. The predicate instead relates the subject with the object.

A valuable representation of the ontology can be done by using graphs.
For all triples, each subject and object is represented by a node, while for each predicate a directed edge, starting from the subject node to the object node is placed. An exemplification is given in Figure 2.6.

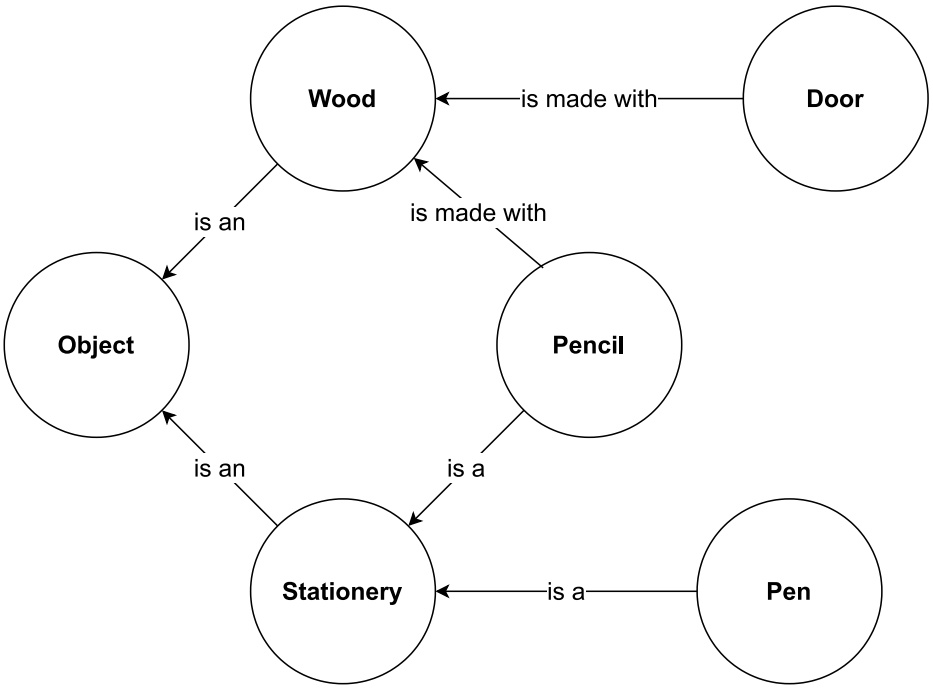

Figure 2.6. Example of ontology

If the object of the triple A is the subject of the triple B, a cascade of relationship can be built as visualized as well.
Moreover, if more triples shares the same subject, a parallel of relationship is built.
In such a way, complex ontology can be represented.

Class ontologies

When a set of specific ontological entities shares a common structure an ontology class can be defined in order to refer to a generic instance of that class. For example, *my computer* is an instance of the ontological class *computers*. Ontological classes can be also defined for object and predicates as well.
Like in object-oriented programming, the use of classes enables to define a structured data template to instantiate as occurrence.
A class ontology is thus an ontology that uses ontological classes as subjects and objects.
This implies that on a class ontology graph representation each node stands for an ontological class while the edges specify the relationship between such ontological classes.
An exemplification is given in Appendix A.

RDF implementation

The Resource Description Framework (RDF) is a framework that aims to represent information about resources, *e.g.* documents, physical objects or concepts. Practically, is a logical way to structure data along with the ontology paradigm. RDF is used when web resources need to be processed by machines, other than displayed to users.
RDF processing is done by a collection of parsers and coder tools.
A rich set of libraries is available for the most common programming languages, thus promoting the popularity of the framework.
The adoption of RDF enables the interoperability of applications by exchanging information in a standardized way.
The RDF is employed for different purposes:

- to embed into Web pages with additional machine-readable information, thus enabling enhanced indexing on search engines;
- to link third-party datasets among each other;
- to interlink API feeds;
- to describe web resources as ontological entities.

RDF uses the concept of triples <*subject*> <*predicate*> <*object*> to format data, such triples in RDF are called *properties*. The RDF uses the logical constructs of ontologies, thus inherits its representation properties in graph form.
In each triple, three types of data can be present: *URI*, literals and blank nodes. URIs, as described above, identify a web resource, and can appear in all three positions of a triple. URIs are used to identify resources such as documents, people, physical objects, and abstract concepts.
Literals are strings that can optionally be associated with a language tag, and may only appear in the *object* position of a RDF property.
Blank nodes are instead uses to make statements without a generic entity (without URI), and can appear only in the *subject* or *object* position of the RDF property.

The RDF is supported by the XML syntax, as reported in the Listing 2.1.
Each RDF serial file contains the XML version and all the data incorporated into the tag:

```
<rdf:RDF>    </rdf:RDF>
```

which represents the root tag containing all the document body.
Inside the tag, the *xmlns:shortName="domainURI#"* rows define the entity namespaces.

They contains the shared part of the URI belonging to the same data structure. Along the file, an RDF description, enclosed by tag (that is defined for each subject):

`<rdf:Description> </rdf:Description>`

Inside such description, the *about* attribute contains the subject URI.
In this description, it also contained a list of predicates, each defined by a custom tag of the form

`<shortName:predicateName> </shortName:predicateName>`

In such syntax, a literal object is directly embedded between the tag brackets. If the object is an entity itself,defined by a URI, its URI is embedded into an attribute (*rdf:resource*) of the predicate tag, instead a simple literal value.
An exemplification is below reported.

```
<cd:artist rdf:resource="http://www.recshop.fake/cd/dylan" />
```

An example of an XML-based RDF data is reported below.

Listing 2.1. RDF example taken from [26]

```
<?xml version="1.0"?>

<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns\#"
xmlns:cd="http://www.recshop.fake/cd\#">

<rdf:Description
rdf:about="http://www.recshop.fake/cd/Empire Burlesque">
  <cd:artist>Bob Dylan</cd:artist>
  <cd:country>USA</cd:country>
  <cd:company>Columbia</cd:company>
  <cd:price>10.90</cd:price>
  <cd:year>1985</cd:year>
</rdf:Description>

<rdf:Description
rdf:about="http://www.recshop.fake/cd/Hide your heart">
  <cd:artist>Bonnie Tyler</cd:artist>
  <cd:country>UK</cd:country>
  <cd:company>CBS Records</cd:company>
  <cd:price>9.90</cd:price>
  <cd:year>1988</cd:year>
</rdf:Description>
.
</rdf:RDF>
```

Chapter 3

Project development

In this chapter will present the workflow of this project.
This workflow is composed from the phases of the building the ontology, create the ground truth and development of the automatic method of entity extraction and linking of several records from the diagnosis.

3.1 Ontology creation

The ontology previously discussed section 2.4 aims to describe the medical records with machine RDF code. This process facilitates the information extraction step by using knowledge graphs.
The graph is structured in such by following the subject-verb-object paradigm.
So a medical record is traduced with a subject node, while the verbs represents the graph edges.
Finally, the object ontology entities are shown as terminal nodes.

3.1.1 Ontology Construction

From the diagnosis database in Italian natural language *NDS*, which contains 50 diagnoses, for each entry a manual research is carried on to locate the new entities that not belong to the considered ontology.
This analysis is done by extracting the relevant keywords from the diagnosis field of the medical records.
Such entities are then cross searched into the *UMLS* database and if found, a further cross search is done in the ontology database *ODB* to retrieve the corresponding ontological class.
If the first search in *UMLS* has negative outcome, the located entities are translated into English and searched again into the *UMLS* database and the *ODB* one.

If the searched entities are not found on the second research step, they are manually evaluated by the ExaMode team.
During the entity search in the *ODB* database, when two or more entities share the same superclass, which is relevant for the project, the latter is added to the ontology to obtain more precise results while keeping semantic correctness and computational efficiency.

3.1.2 Ontology description

The fully implemented ontology is schematized in Appendix A.
It is composed of five main areas, according to the following partition:

- the general ExaMode classes (light green), they are shared among all the use-cases;
- Cervix cancer classes (light red);
- Colon cancer classes (light purple);
- Lung cancer classes (light pink);
- Celiac disease classes (light blue).

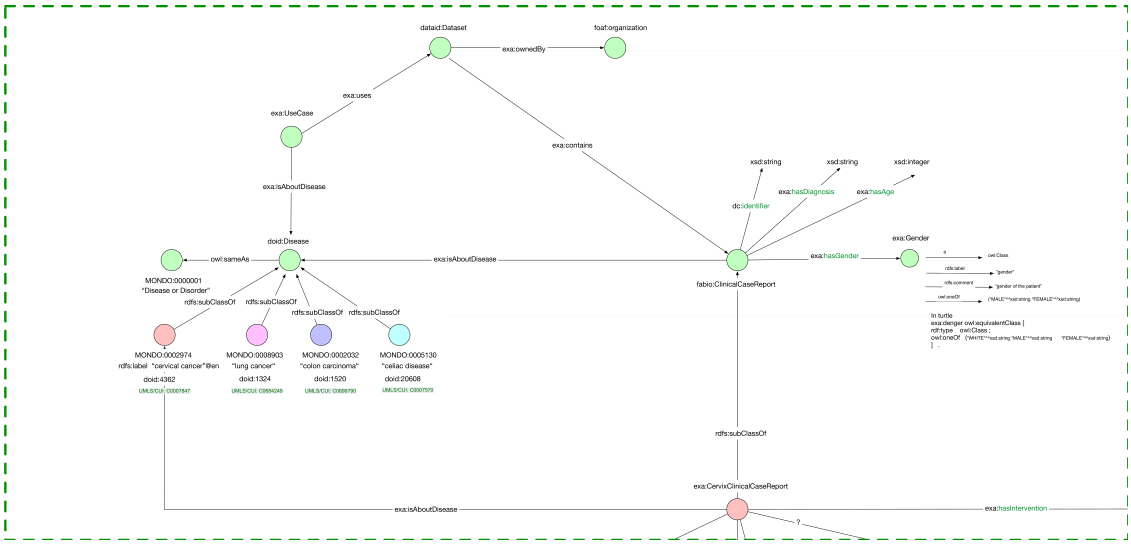

Figure 3.1. ExaMode use cases

With the exception of project classes, each group is related to a specific disease.

Another class grouping can be done with different criteria, according to the functionalities of the information extracted.
The *ExaMode use cases* box, reported in Figure 3.1, collects all the classes related to the patient identity and the medical record traceability.
In particular, it is also present the diagnosis output in terms of disease macro-area (e.g. cervix cancer).

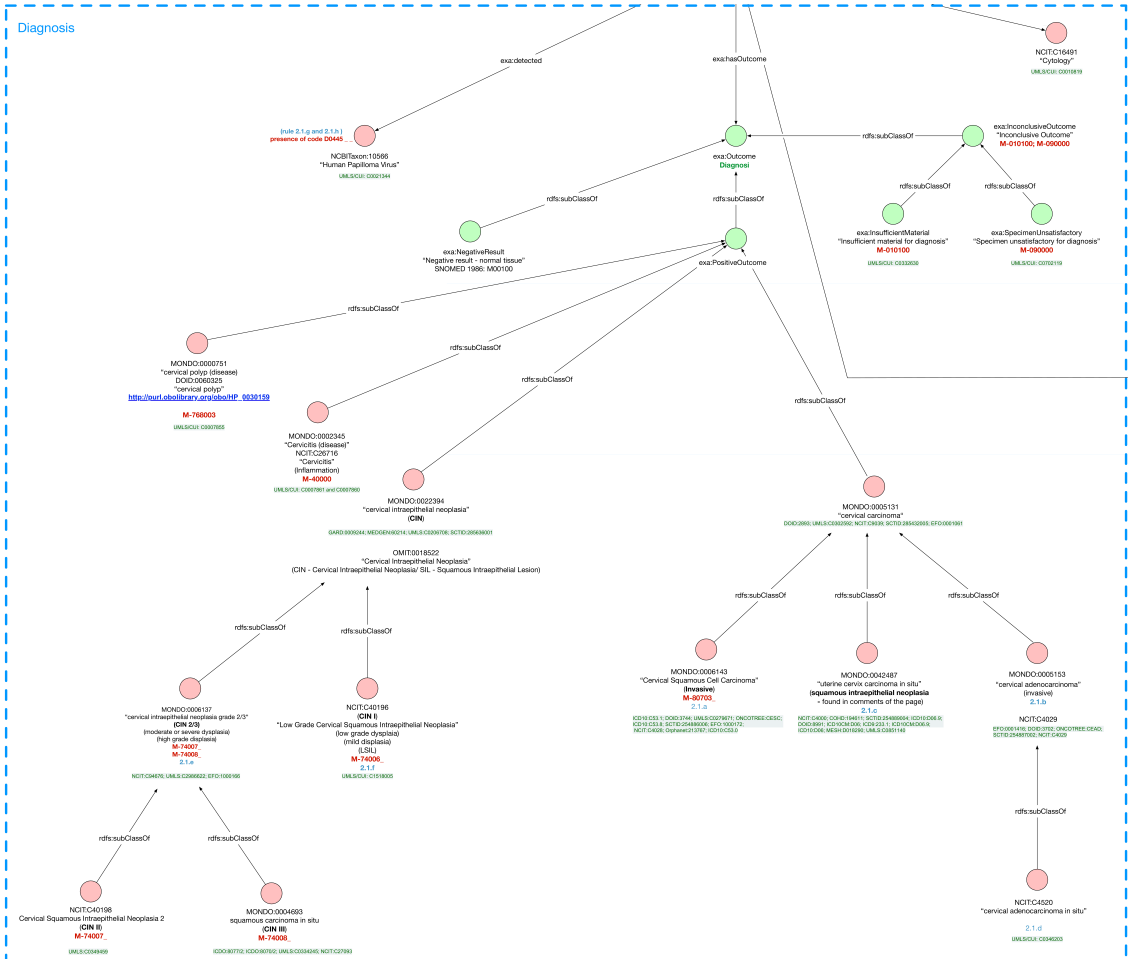

Figure 3.2. ExaMode diagnoses

The *diagnosis* box, depicted in Figure 3.2, reports the details of the diagnosis process.
Notably, in the Cervix cancer case, it reports the diagnosis outcome as positive, uncertain, or negative. In case of positive result, the cancer type is reported.
The main causes are also investigated and reported, like the detection of the *papilloma virus*.

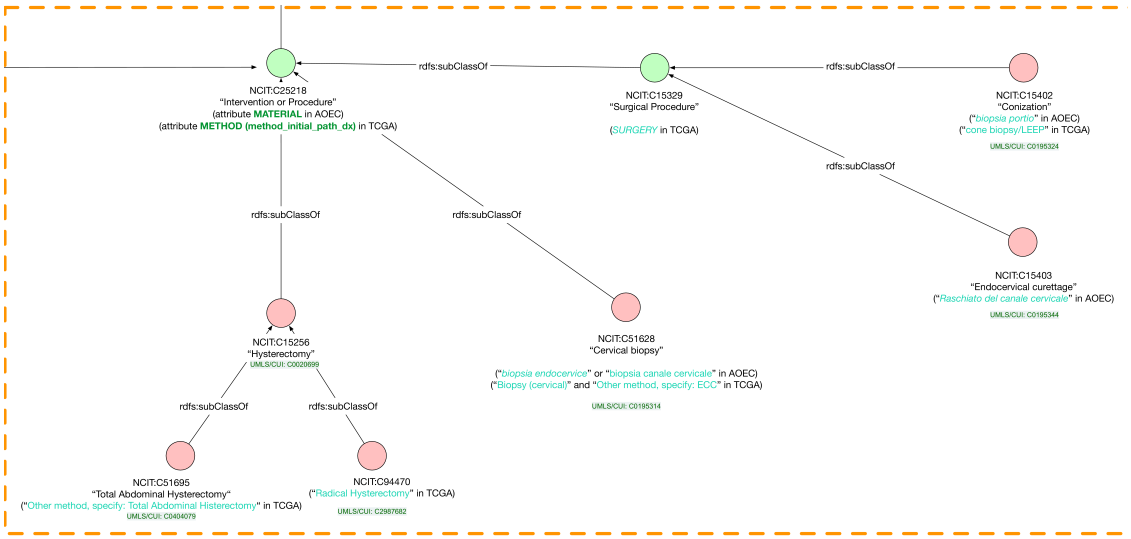

Figure 3.3. ExaMode procedures

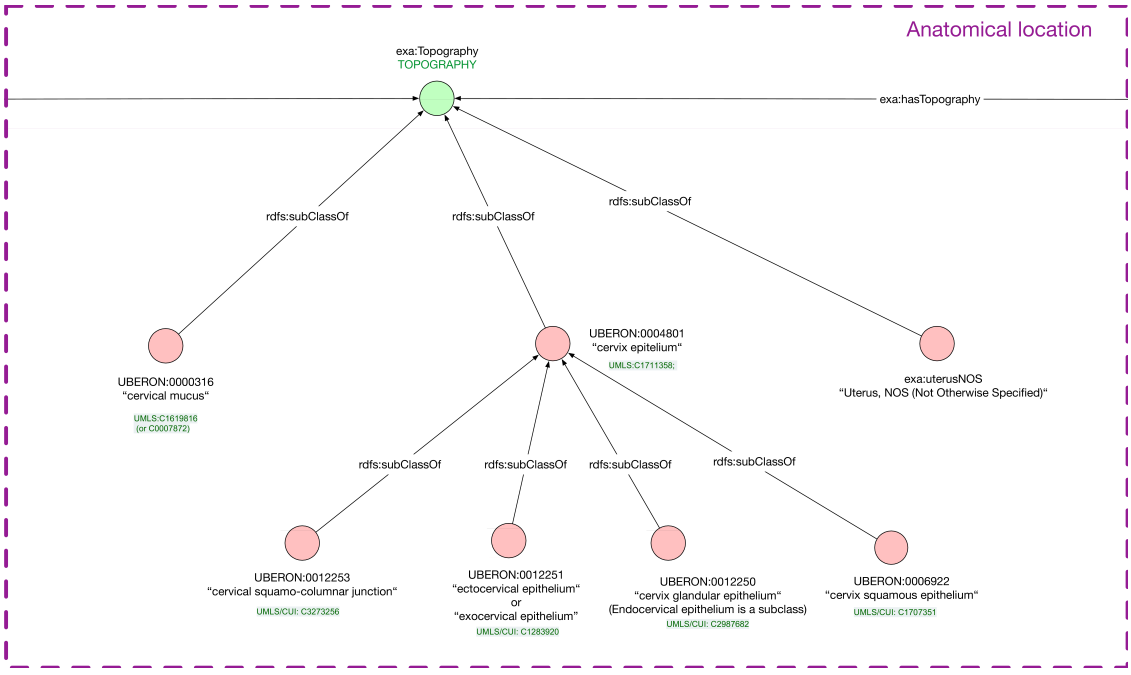

Figure 3.4. ExaMode anatomic location

The *procedure box*, shown in Figure 3.3, describes the extraction procedure of the medical sample used to perform the diagnosis.

In particular, in the actual example, three operation have been conducted: a hysterectomy, a cervical biopsy and a surgical procedure composed of endocervical curettage and conization.

The *anatomical location* box, reported on Figure 3.4, details the information concerning the medical sample extracted.
In the reported example, information about the cervix epithelium and mucus are extracted.
Some terminal nodes have additional codes that relates the entities with external ontologies.
For example, *MONDO:0002974* is related to Mondo Disease Ontology http://www.ontobee.org/ontology/MONDO?iri=http://purl.obolibrary.org/obo/MONDO_0002974.
It also present the *SNOMED* code (in red) that relates the entities to internal hospital record system.
The *UMLS/CU* Unified Medical Language System (UMLS) code, in green, instead relates them to the database managed by National Institute of Health (NIH).
Notably, UMLS is a set of files and software that brings together many health and biomedical vocabularies and standards to enable interoperability between computer systems.

3.2 Ground Truth generation

The present work aims to build a training set to validate the automated *Entity Extraction* and *Entity Linking* methods described in the following chapters.
In particular, a set of rules, based on the ontology previously described in the previous section, is manually created to define the *Ground Truth*.
The process used, called *Rule-Based Graph Extraction* (RBGE), uses the ontology and the previously defined *NDS* database to generate the a series of ontological rules.
Inside the RGBE process, the rule generation is the core part.
In order to standardize the procedure and efficiently manage the input information, a *SQL* database has been implemented.
The output of the RGBE process is a set of ontological rules that complies with the Resource Description Framework (RDF) format.

3.2.1 Conversion rules

In the RBGE process, the rule generation step employs a manual procedure to devise the ontological rules that compose the ground truth.
These rules are assumed true and compose the target set for the Entity Extraction methods.

An ontological rule, as discussed in subsection 2.4.1 is a triple Subject-Verb-Object.
In this case, the subject is represented by the Entity of the medical record, described by its identification number.
The predicates are descriptors used inside the ontology to relate one class entity to another, i.e. *exa:hasDiagnosis* or *exa:hasTopography*. The object complements are instead extracted from the fields of the *NDS* database and comprises:

- the ID of the medical record;
- the integral diagnosis text;
- the patient's age;
- the patient's gender;
- the interested anatomical part contained in the diagnosis;
- the intervention procedure *SNOMED* code;
- the topography *SNOMED* code;
- the diagnosis *SNOMED* code;
- the morphological *SNOMED* code;
- the biological sample description.

They are translated in the ontology as literal (labels) or entities.

3.2.2 Rule Generation

Four procedures are used to generate the rules: the first analyses the *SNOMED* codes in the *NDS* database and directly translates it in ontological object complements.
This type of generation has been done by using automated scripting.
The second procedure automatically extract the relevant keywords contained on the biological sample description and codifies them into the respective ontological entity.
The third procedure automatically codifies the auxiliary information contained in the medical record, *e.g.* the patient's age, and in the *ExaMode* classes, *e.g.* *exa:CervixClinicalCaseReport*.
Finally, the fourth procedure requires the data scientist to the read the diagnosis text and manually extract the additional ontological entities not previously found with the previous techniques.

The following tables report the *translation rules* extracted from the above procedures.

The Table 3.1 reports the rule to translate the intervention procedure in *SNOMED* code into an ontological entity with predicate *exa:hasIntervention*. In particular the field *Value(From)* contains the procedure *SNOMED* code, the *Meaning(ML)* field contains the human language description, the *Value(To)* field contains the ontological entity ID, while the *UMLS/CUI* field contains the *UMLS/CUI* associated to the entity.

Table 3.1. Procedure rules

Value(From)	Meaning(NL)	Value(To)	UMLS/CUI
P-40	"Needle Biopsy"	NCIT:C15190	C0005560

The Table 3.2 details the rule to translate the topography *SNOMED* code into an entity with predicate *exa:hasTopography*. The fields are the same of the above table.

Table 3.2. Topography rules

Value(From)	Meaning(NL)	Value(To)	UMLS/CUI
T-83000	"Uterus, Not Otherwise Specified"	exa:uterusNOS	C0042149

The Table 3.3 contains the rule to translate the diagnosis and diseases *SNOMED* code into an entity with predicate *exa:detected*.

Table 3.3. Diagnosis rules

Value(From)	Meaning(NL)	Value(To)	UMLS/CUI
D-044500	"Human Papilloma Virus"	NCBITaxon:10566	C0021344

The Table 3.4 instead enables to translate the morphology *SNOMED* code through predicate *exa:hasOutcome*.

Table 3.4. Morphology rules

Value(From)	Meaning(NL)	Value(To)	UMLS/CUI
M-010100	"Insufficient material for diagnosis"	exa:InsufficientMaterial	C0332630
M-090000	"Specimen unsatisfactory for diagnosis"	exa:SpecimenUnsatisfactory	C0702119
M-400000	"Cervicitis (Inflammation)"	MONDO:0002345	C0007860
M-740060	"CIN I"	NCIT:C40196	C1518005
M-740070	"CIN II"	NCIT:C40198	C0349459
M-740086	"CIN III"	MONDO:0004693	C0334245
M-768003	"Cervical polyp"	MONDO:0000751	C0007855
M-807030	"Cervical Squamous Cell Carcinoma"	MONDO:0006143	C0279671

The Table 3.5 reports the translation rules to convert the keywords extracted from the biological sample description into ontological entities with the predicate *exa:hasIntervention*.

Table 3.5. Material rules

Value(From)	Meaning(NL)	Value(To)	UMLS/CUI
"BIOPSIA ENDOCERVICE"	"Cervical biopsy"	NCIT:C51628	C0195314
"BIOPSIA ENDOCERVICALE"	"Cervical biopsy"	NCIT:C51628	C0195314
"BIOPSIA CANALE CERVICALE"	"Cervical biopsy"	NCIT:C51628	C0195314
"BIOPSIA ORE 1"	"Cervical biopsy"	NCIT:C51628	C0195314
"BIOPSIA PORTIO"	"Conization"	NCIT:C15402	C0195324
"BIOPSIA PORTIO1-3"	"Conization"	NCIT:C15402	C0195324
"BIOPSIOA PORTIO ORE 12"	"Conization"	NCIT:C15402	C0195324
"RASCHIATO DEL CANALE CERVICALE"	"Endocervical curettage"	NCIT:C15403	C0195344
"RASCHIATO CANALE CERVICALE"	"Endocervical curettage"	NCIT:C15403	C0195344

The Table 3.6 contains the conversion rules that deals with auxiliary information extracted. The *Property* field contains a set of shared predicates; the *Class* field contains the type ontological entity, while the *From* field contains the *NDS* database reference.

Table 3.6. Other Properties

Property	Class	From
exa:cervix/??/????	Subject/Instance	ID_identificativo (Derivative)
rdf:type	exa:CervixClinicalCaseReport	
exa:isAboutDisease	MONDO:0002974 C0007847	Name sheet excel file
exa:identifier	Literal	ID_identificativo
exa:hasDiagnosis	Literal	Diagnosi
exa:hasAge	Literal	Età
exa:hasGender	Literal	Sesso

Finally, Table 3.7 details the conversion rules to manage the interested anatomical part contained in the diagnosis, resulting from the manual procedure above described.
The table follows the same format as Table 3.1, while the rules share the same predicate *exa:hasLocation*.

Table 3.7. Anatomical location rules

Value(From)	Meaning(NL)	Value(To)	UMLS/CUI
"Epitelio squamocellulare"	"cervix squamous epithelium"	UBERON:0006922	C1707351
"Epitelio glanulare/endocel."	"cervix glandular epithelium"	UBERON:0012250	C1707350
"Epitelio (eso/ecto)cel."	"ectocervical epithelium"	UBERON:0012251	C1283920
"Giunzione squamo-colonnare"	"cervical squamo-columnar junction"	UBERON:0012253	C3273256
"Muco o mucosa"	"cervical mucus"	UBERON:0000316	C1619816

In this chapter we will take the following diagnosis example, extract from the *NDS* and we will analysing it in the future step:

ID identificativo	Diagnosi	Procedura	Topografia	Diagnosi	Età	Sesso	Materiali
19/1795	Mucosa esocervicale ...	P-40	T-83000	M-740060 D-044500	36	F	BIOPSIA PORTIO

So from the previous example, this will be our set of rules:

Subject	Predicate	Object
exa:cervix/19/1795	exa:hasLocation	NCIT:C15190
exa:cervix/19/1795	exa:hasTopography	exa:uterusNOS
exa:cervix/19/1795	exa:detected	NCBITaxon:10566
exa:cervix/19/1795	exa:hasOutcome	NCIT:C40196
exa:cervix/19/1795	exa:hasIntervention	NCIT:C15402
exa:cervix/19/1795	rdf:type	exa:CervixClinicalCaseReport
exa:cervix/19/1795	exa:isAboutDisease	MONDO:0002974
exa:cervix/19/1795	exa:identifier	"19/1795"
exa:cervix/19/1795	exa:hasDiagnosis	"Mucosa esocervicale con alterazioni ..."
exa:cervix/19/1795	exa:hasAge	"36"
exa:cervix/19/1795	exa:hasGender	"Female"
exa:cervix/19/1795	exa:hasLocation	UBERON:0000316
exa:cervix/19/1795	exa:hasLocation	UBERON:0012250

3.2.3 Database implementation

In order to efficiently manage the ontological rules previously extracted, a relational database based on *PostgreSQL* is devised.

The database Entity Relationship (ER) schematic is depicted in Figure 3.5. Each ontology entity is defined through its *Concept* which has a proper short-name *Concept_SN*, URL *URL_Name* and the *IsInstance* boolean flag which specify if the entity is associated to a medical record.
The *Namespace* instead contains the URL domain related to each ontology entity. The *ConceptMap* collects auxiliary information such as the *UMLS* and *SNOMED* codes, and the ontological entity language description *Name* belonging to the relative *Concept*.
Each ontological rule can be built by aggregating two or more concept according to the following methods:

- *Statement*: a triple Subject-Predicate-Object;
- *StatementLiteral*: a triple Subject-Predicate-Object in which the Object part is a literal.

The Figure 3.6 reports the logical scheme of the database.

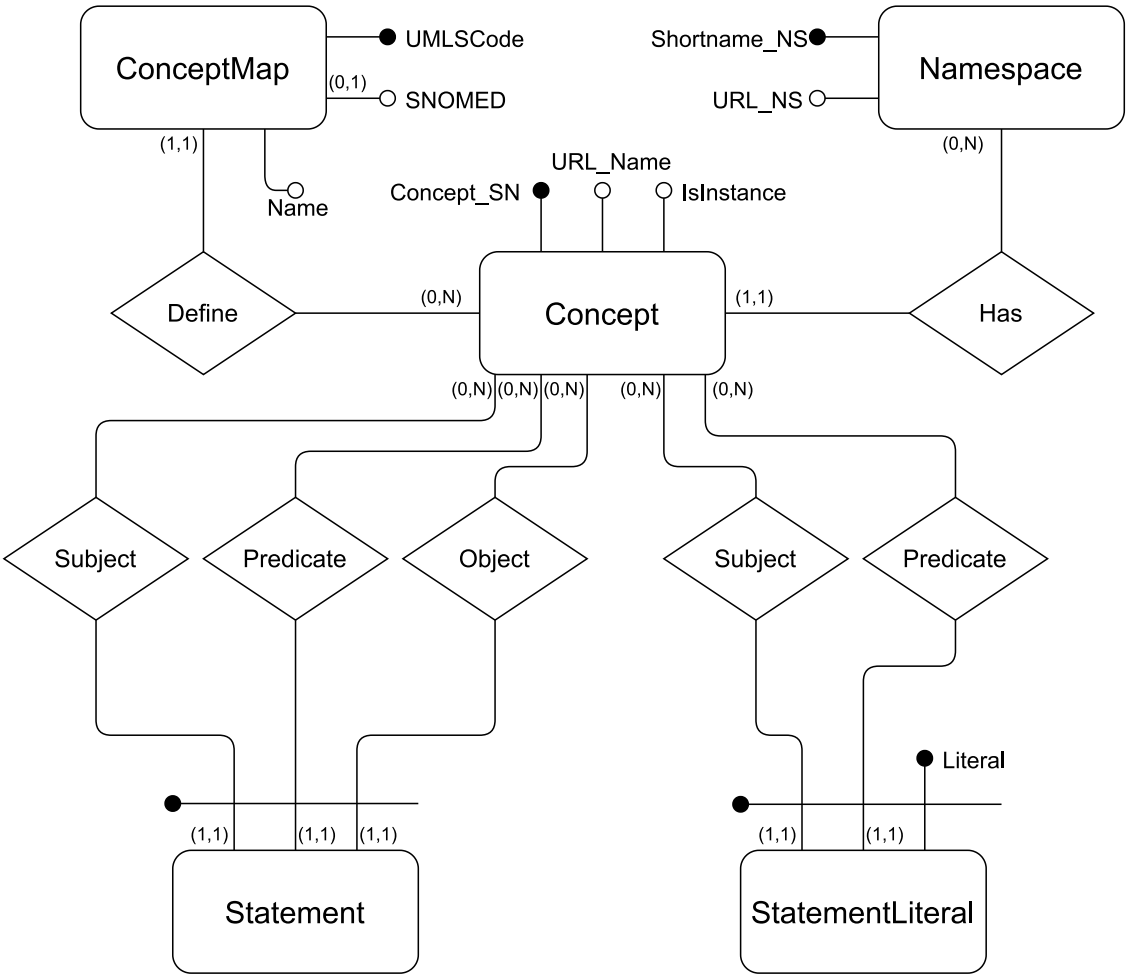

Figure 3.5. Database ER schema

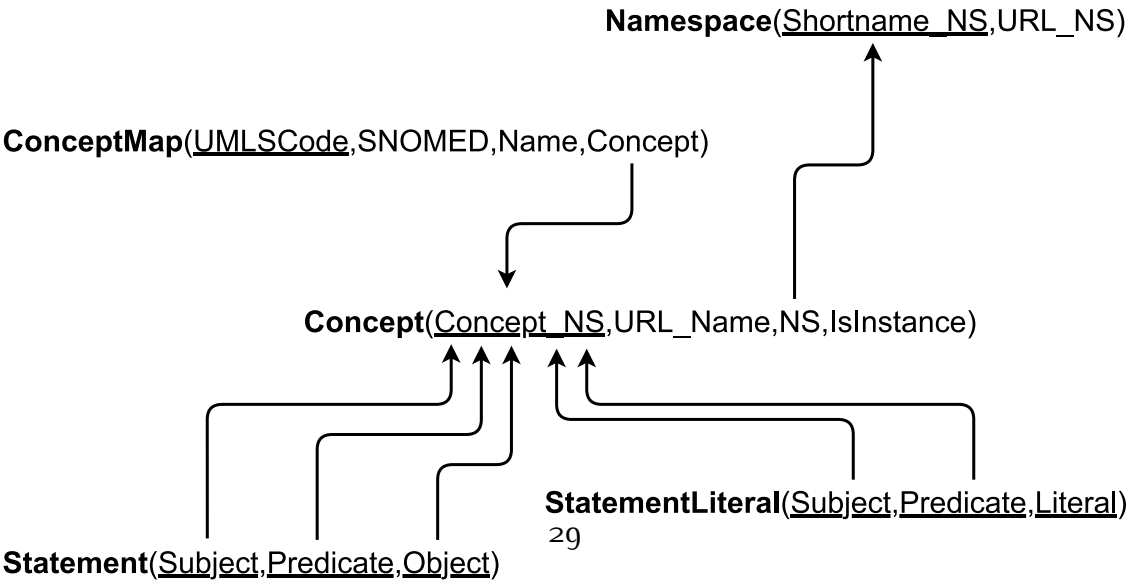

Figure 3.6. Database logical schema

In this schematic, the *ER* schematic is translated into a series of tables, table attributes, and relations; *e.g.* the *Concept* table, in its field *NS*, contains the values of the *Shortname_NS* of the *Namespace* table. In the schematic, the underlined attributes are used as table keys.
The SQL translations of the tables and theirs primary and external key definitions follow.

- The *Namespace* table:

```
CREATE TABLE cervix."Namespace" (
    "Shortname_NS" text NOT NULL,
    "URL_NS" text NOT NULL
);

ALTER TABLE ONLY cervix."Namespace"
    ADD CONSTRAINT "Namespace_pkey" PRIMARY KEY ("Shortname_NS");
```

- the *Concept* table:

```
CREATE TABLE cervix."Concept" (
    "Concept_SN" text NOT NULL,
    "URL_Name" text NOT NULL,
    "NS" text NOT NULL,
    "IsInstance" boolean DEFAULT true
);

ALTER TABLE ONLY cervix."Concept"
    ADD CONSTRAINT "Concept_pkey" PRIMARY KEY ("Concept_SN");

ALTER TABLE ONLY cervix."Concept"
    ADD CONSTRAINT "Concept_NS_fkey" FOREIGN KEY ("NS") REFERENCES
        cervix."Namespace"("Shortname_NS") ON UPDATE CASCADE ON
        DELETE CASCADE;
```

- the *ConceptMap* table:

```
CREATE TABLE cervix."ConceptMap" (
    "UMLSCode" text NOT NULL,
    "SNOMED" text,
    "Name" text NOT NULL,
    "Concept" text NOT NULL
);

ALTER TABLE ONLY cervix."ConceptMap"
    ADD CONSTRAINT "ConceptMap_pkey" PRIMARY KEY ("UMLSCode");
```

```
ALTER TABLE ONLY cervix."ConceptMap"
    ADD CONSTRAINT "ConceptMap_Concept_fkey" FOREIGN KEY
        ("Concept") REFERENCES cervix."Concept"("Concept_SN") ON
        UPDATE CASCADE ON DELETE CASCADE;
```

- the *Statement* table:

```
CREATE TABLE cervix."Statement" (
    "Subject" text NOT NULL,
    "Predicate" text NOT NULL,
    "Object" text NOT NULL
);

ALTER TABLE ONLY cervix."Statement"
    ADD CONSTRAINT "Statement_pkey" PRIMARY KEY ("Subject",
        "Predicate", "Object");

ALTER TABLE ONLY cervix."Statement"
    ADD CONSTRAINT "Statement_Subject_fkey" FOREIGN KEY
        ("Subject") REFERENCES cervix."Concept"("Concept_SN") ON
        UPDATE CASCADE ON DELETE CASCADE;

ALTER TABLE ONLY cervix."Statement"
    ADD CONSTRAINT "Statement_Predicate_fkey" FOREIGN KEY
        ("Predicate") REFERENCES cervix."Concept"("Concept_SN") ON
        UPDATE CASCADE ON DELETE CASCADE;

ALTER TABLE ONLY cervix."Statement"
    ADD CONSTRAINT "Statement_Object_fkey" FOREIGN KEY ("Object")
        REFERENCES cervix."Concept"("Concept_SN") ON UPDATE CASCADE
        ON DELETE CASCADE;
```

- the *StatementLiteral* table:

```
CREATE TABLE cervix."StatementLiteral" (
    "Subject" text NOT NULL,
    "Predicate" text NOT NULL,
    "Literal" text NOT NULL
);

ALTER TABLE ONLY cervix."StatementLiteral"
    ADD CONSTRAINT "StatementLiteral_pkey" PRIMARY KEY ("Subject",
        "Predicate", "Literal");
```

```
ALTER TABLE ONLY cervix."StatementLiteral"
    ADD CONSTRAINT "StatementLiteral_Subject_fkey" FOREIGN KEY
        ("Subject") REFERENCES cervix."Concept"("Concept_SN") ON
        UPDATE CASCADE ON DELETE CASCADE;

ALTER TABLE ONLY cervix."StatementLiteral"
    ADD CONSTRAINT "StatementLiteral_Predicate_fkey" FOREIGN KEY
        ("Predicate") REFERENCES cervix."Concept"("Concept_SN") ON
        UPDATE CASCADE ON DELETE CASCADE;
```

The *PostgreSQL* database above discussed is populated with data coming from the *NDS* database through the script (script 3.1).
Initially, the concepts pertaining to the ontology described in this chapter are manually inserted into the *PostgreSQL* database, then the script is run.

Listing 3.1. Script Python from Excell to SQL

```
import psycopg2
import xlrd

# Function find text in a map
def findmap(mymap, text):
    for key in mymap:
        if key in text:
            return mymap.get(key, "none")
    return ""

# Connection to db
conn = psycopg2.connect(dbname="EXAMODE", host="localhost",
    user="postgres", password="*******")
cur = conn.cursor()

# Db query to get map data
SNOMED_queryP = "SELECT \"SNOMED\", \"Concept\" FROM
    cervix.\"ConceptMap\" WHERE \"SNOMED\" LIKE 'P%';"
SNOMED_queryT = "SELECT \"SNOMED\", \"Concept\" FROM
    cervix.\"ConceptMap\" WHERE \"SNOMED\" LIKE 'T%';"
SNOMED_queryD = "SELECT \"SNOMED\", \"Concept\" FROM
    cervix.\"ConceptMap\" WHERE \"SNOMED\" LIKE 'D%';"
SNOMED_queryM = "SELECT \"SNOMED\", \"Concept\" FROM
    cervix.\"ConceptMap\" WHERE \"SNOMED\" LIKE 'M%';"

# Db query to insert data
```

```
InsertID = "INSERT INTO cervix.\"Concept\" (\"Concept_SN\",
    \"URL_Name\", \"NS\") VALUES (%s, %s, %s) ON CONFLICT DO NOTHING;
InsertState = "INSERT INTO cervix.\"Statement\" (\"Subject\",
    \"Predicate\", \"Object\") VALUES (%s, %s, %s) ON CONFLICT DO
    NOTHING;"
InsertStateLit = "INSERT INTO cervix.\"StatementLiteral\" (\"Subject\",
    \"Predicate\", \"Literal\") VALUES (%s, %s, %s) ON CONFLICT DO
    NOTHING;"

# Snomed maps definition
cur.execute(SNOMED_queryP)
SNOMEDmapP = {}
rows = cur.fetchall()
for row in rows:
    SNOMEDmapP[row[0]] = row[1]

cur.execute(SNOMED_queryT)
SNOMEDmapT = {}
rows = cur.fetchall()
for row in rows:
    SNOMEDmapT[row[0]] = row[1]

cur.execute(SNOMED_queryD)
SNOMEDmapD = {}
rows = cur.fetchall()
for row in rows:
    SNOMEDmapD[row[0]] = row[1]

cur.execute(SNOMED_queryM)
SNOMEDmapM = {}
rows = cur.fetchall()
for row in rows:
    SNOMEDmapM[row[0]] = row[1]

# Map traslation
InterMap = {
    "BIOPSIA ENDOCERVICE": "NCIT:C51628",
    "BIOPSIA ENDOCERVICALE": "NCIT:C51628",
    "BIOPSIA CANALE CERVICALE": "NCIT:C51628",
    "BIOPSIA ORE 1": "NCIT:C51628",
    "BIOPSIA PORTIO": "NCIT:C15402",
    "BIOPSIA PORTIO1-3": "NCIT:C15402",
    "BIOPSIOA PORTIO ORE 12": "NCIT:C15402",
    "RASCHIATO DEL CANALE CERVICALE": "NCIT:C15403",
```

```
    "RASCHIATO CANALE CERVICALE": "NCIT:C15403"
}

# Map of gender
GenderMap = {"M": "MALE", "F": "FEMALE"}

# List of predicate
Predicates = ["rdf:type", "exa:isAboutDisease", "exa:identifier",
    "exa:hasDiagnosis", "exa:hasAge", "exa:hasGender",
             "exa:hasTopography", "exa:detected", "exa:hasOutcome",
                "exa:hasIntervention", "exa:hasLocation"]

# Open the excell file
filename = # Path-file + File-name
sheet = xlrd.open_workbook(filename,
    encoding_override="utf_16_le").sheet_by_name("Cervix")

# Insert in the db the triple
for row in range(2, sheet.nrows):
    id = sheet.cell(row, 0).value
    idconcept = "exa:cervix/" + id
    idurl = "cervix/" + id
    if idurl.endswith("*"):
        idurl = idurl[:(len(idurl) - 1)]
    data = (idconcept, idurl, "exa")
    cur.execute(InsertID, data)
    data = (idconcept, Predicates[0], "exa:CervixClinicalCaseReport")
    cur.execute(InsertState, data)
    data = (idconcept, Predicates[1], "MONDO:0002974")
    cur.execute(InsertState, data)
    data = (idconcept, Predicates[2], id)
    cur.execute(InsertStateLit, data)
    data = (idconcept, Predicates[3], sheet.cell(row, 1).value)
    cur.execute(InsertStateLit, data)
    data = (idconcept, Predicates[4], round(sheet.cell(row, 5).value))
    cur.execute(InsertStateLit, data)
    myvalue = findmap(GenderMap, sheet.cell(row, 6).value)
    if myvalue != "":
        data = (idconcept, Predicates[5], myvalue)
        cur.execute(InsertStateLit, data)
    myvalue = findmap(SNOMEDmapP, sheet.cell(row, 2).value)
    if myvalue != "":
        data = (idconcept, Predicates[9], myvalue)
        cur.execute(InsertState, data)
```

```
    myvalue = findmap(SNOMEDmapT, sheet.cell(row, 3).value)
    if myvalue != "":
        data = (idconcept, Predicates[6], myvalue)
        cur.execute(InsertState, data)
    myvalue = findmap(SNOMEDmapD, sheet.cell(row, 4).value)
    if myvalue != "":
        data = (idconcept, Predicates[7], myvalue)
        cur.execute(InsertState, data)
    myvalue = findmap(SNOMEDmapM, sheet.cell(row, 4).value)
    if myvalue != "":
        data = (idconcept, Predicates[8], myvalue)
        cur.execute(InsertState, data)
    myvalue = findmap(InterMap, sheet.cell(row, 7).value)
    if myvalue != "":
        data = (idconcept, Predicates[9], myvalue)
        cur.execute(InsertState, data)

# Close db connection
conn.commit()
cur.close()
conn.close()
```

After this, some statement are inserted manually in the database, having an human operator reading the full diagnosis and add more statement not added with the previously script.

In the end, in the database, for the diagnosis cervix/19/1795 are present the following tables:

- From the table *Statement*:

Subject	Predicate	Object
exa:cervix/19/1795	exa:detected	NCBITaxon:10566
exa:cervix/19/1795	exa:hasIntervention	NCIT:C15190
exa:cervix/19/1795	exa:hasIntervention	NCIT:C15402
exa:cervix/19/1795	exa:hasLocation	UBERON:0000316
exa:cervix/19/1795	exa:hasLocation	UBERON:0012250
exa:cervix/19/1795	exa:hasOutcome	NCIT:C40196
exa:cervix/19/1795	exa:hasTopography	exa:uterusNOS
exa:cervix/19/1795	exa:isAboutDisease	MONDO:0002974
exa:cervix/19/1795	rdf:type	exa:CervixClinicalCaseReport

- From the table *StatementLiteral*:

Subject	Predicate	Literal
exa:cervix/19/1795	exa:hasAge	36
exa:cervix/19/1795	exa:hasDiagnosis	Mucosa esocervicale con alterazioni...
exa:cervix/19/1795	exa:hasGender	FEMALE
exa:cervix/19/1795	exa:identifier	19/1795

3.2.4 RDF generation

In order to produce results compatible with the RDF format, the *PostgreSQL* database is serialized accordingly.
The serialization process is automated by using the following script:

```
import psycopg2
import rdflib
from rdflib import Literal, Namespace

# Connection to db
conn = psycopg2.connect(dbname="EXAMODE", host="localhost",
    user="postgres", password="********")
cur = conn.cursor()

# Query for URI and Statements
QueryURI = "SELECT \"Concept_SN\", \"URL_Name\", \"URL_NS\", \"NS\"" \
          " FROM cervix.\"Concept\" INNER JOIN cervix.\"Namespace\" ON
              \"NS\" = \"Shortname_NS\";"
QueryNS = "SELECT * FROM cervix.\"Namespace\";"
QueryStatement = "SELECT * FROM cervix.\"Statement\";"
QueryLiteral = "SELECT * FROM cervix.\"StatementLiteral\";"

# Map for element to URI
URIMap = {}
cur.execute(QueryURI)
rows = cur.fetchall()
for row in rows:
    URIMap[row[0]] = [row[1], row[3]]

# Create Graph rdf
myrdf = rdflib.Graph()

# Add Namespaces
mapNS = {}
cur.execute(QueryNS)
rows = cur.fetchall()
for row in rows:
```

```
    ns = Namespace(row[1])
    mapNS[row[0]] = ns
    myrdf.namespace_manager.bind(row[0], row[1], override=True)

# Add triple with literal
cur.execute(QueryLiteral)
rows = cur.fetchall()
for row in rows:
    mySubject = mapNS.get(URIMap.get(row[0], "none")[1],
        "none")[URIMap.get(row[0], "none")[0]]
    myPredicate = mapNS.get(URIMap.get(row[1], "none")[1],
        "none")[URIMap.get(row[1], "none")[0]]
    myLiteral = Literal(row[2])
    myrdf.add((mySubject, myPredicate, myLiteral))

# Add triple without literal
cur.execute(QueryStatement)
rows = cur.fetchall()
for row in rows:
    mySubject = mapNS.get(URIMap.get(row[0], "none")[1],
        "none")[URIMap.get(row[0], "none")[0]]
    myPredicate = mapNS.get(URIMap.get(row[1], "none")[1],
        "none")[URIMap.get(row[1], "none")[0]]
    myObject = mapNS.get(URIMap.get(row[2], "none")[1],
        "none")[URIMap.get(row[2], "none")[0]]
    myrdf.add((mySubject, myPredicate, myObject))

# Write rdf file
myrdf.serialize(destination='Cervix_GT_Serial_V2.rdf',
    format='pretty-xml')
myrdf.serialize(destination='Cervix_GT_Triple_V2.nt', format='nt')
myrdf.serialize(destination='Cervix_GT_Trig_V2.trig', format='trig')
myrdf.serialize(destination='Cervix_GT_Turtle_V2.ttl', format='turtle')

# Close db connection
cur.close()
conn.close()
```

An exemplification of the ontological entity description, related to the medical record cervix/19/1795 is depicted in Figure 3.7.

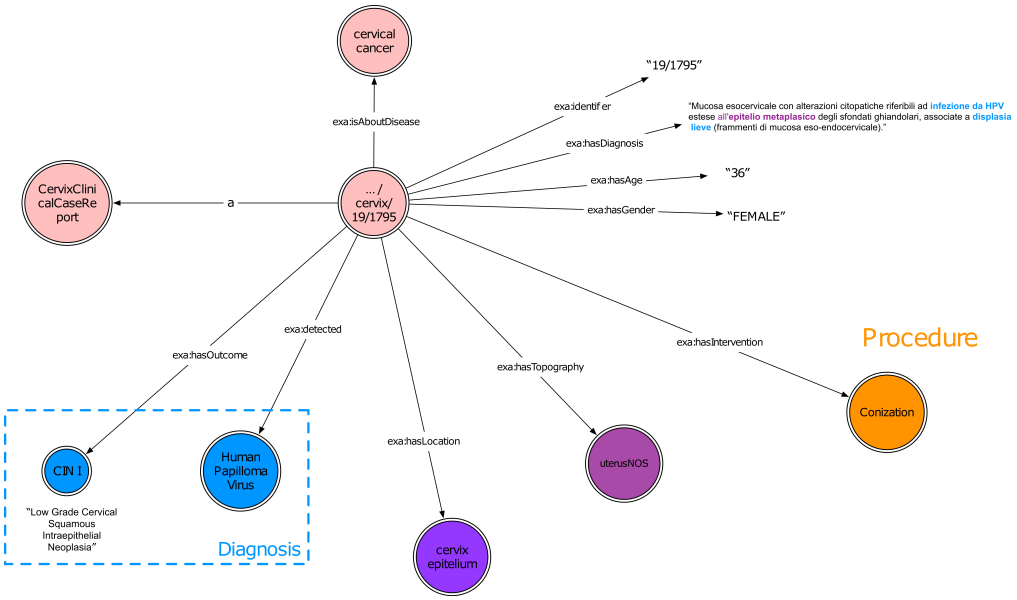

Figure 3.7. Ontology cervix/19/1795

The associated output file, in the RDF format, is instead reported in Listing 3.2.4.
The first file defines the ontological subject, namely the case report associated with the medical record.
The following lines reports a series of predicate-object pairs of the type *exa:hasTopography* (predicate), *https://www.examode.eu/ontology/uterusNOS* (object), both codified by the ontology of subsection 2.4.1.
In lines where the object is a literal, the predicate tags surround the object values; *e.g.* <*exa:hasAge*> <*/exa:hasAge*> are the predicate tags while *36* is the literal object.
The ending line contains the closing tags associated with the medical record main entity.

```
<?xml version="1.0" encoding="utf-8"?>
<rdf:RDF
  xmlns:exa="https://www.examode.eu/ontology/"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
>
...
  <exa:CervixClinicalCaseReport
      rdf:about="https://www.examode.eu/ontology/cervix/19/1795">
    <exa:hasTopography
        rdf:resource="https://www.examode.eu/ontology/uterusNOS"/>
```

```
    <exa:hasIntervention
        rdf:resource="http://purl.obolibrary.org/obo/NCIT_C15402"/>
    <exa:hasAge>36</exa:hasAge>
    <exa:identifier>19/1795</exa:identifier>
    <exa:isAboutDisease
        rdf:resource="http://purl.obolibrary.org/obo/MONDO_0002974"/>
    <exa:hasOutcome
        rdf:resource="http://purl.obolibrary.org/obo/NCIT_C40196"/>
    <exa:hasIntervention
        rdf:resource="http://purl.obolibrary.org/obo/NCIT_C15190"/>
    <exa:hasLocation
        rdf:resource="http://purl.obolibrary.org/obo/UBERON_0012250"/>
    <exa:hasDiagnosis>Mucosa esocervicale con alterazioni citopatiche
        riferibili ad infezione da HPV estese allo epitelio metaplasico
        degli sfondati ghiandolari, associate a displasia lieve
        (frammenti di mucosa eso-endocervicale).</exa:hasDiagnosis>
    <exa:hasLocation
        rdf:resource="http://purl.obolibrary.org/obo/UBERON_0000316"/>
    <exa:hasGender>FEMALE</exa:hasGender>
    <exa:detected
        rdf:resource="http://purl.obolibrary.org/obo/NCBITaxon_10566"/>
  </exa:CervixClinicalCaseReport>
```

3.3 Automatic results generation

After manually having generated the ground truth, it would be an improvement in term of time to be able to generate automatically the statement.
To do so we need to use several tools of entity extraction to, like it says, extract entities from the diagnosis, after merge the entities from the different tools and to the end linking this entities to the ontology to create automatically the statement.
In this section, this procedure used in this thesis will be explained.

3.3.1 Entity extraction

Among Natural Language Programming (NLP), several tasks are accomplished. *Entity Extraction* (EE), or Named Entity Recognition (NER) is the process that aims to automatically mine and label entities from text, by matching its syntactical meaning.
In the present work EE is used to extract medical entities from the diagnosis texts embedded in records.
For example, in "The patient presents Human Papilloma Virus" the EE process

should correctly identify "Human Papilloma Virus" as medical entity. The entities are then matched with the reference ontology.
To extract the entities, EE algorithms receives input the text and a knowledge base in form of dictionary.
Then, text is split into its minimal chunks and arranged in a search tree. In such a way, single chunks can be easily extracted from leaves to compose simple queries, but complex queries can be created as well by combining leaves with parent nodes.
For example, from the expression "Human Papilloma Virus" the search query set Q={"Human","Papilloma","Virus","Human Papilloma","Papilloma Virus","Human Virus" ,"Human Papilloma Virus"} can be created.
Next, each query in Q is searched the knowledge base; such bases, depending on available technology, can be dictionaries, search engine lists of results, and machine learning model predictions. In the present analysis, results obtained

from different EE tools have been combined to improve the output accuracy and the number of entities extracted.
The following tools are thus employed.

- MetaMap[3]: MetaMap is a highly configurable program developed since 1994 by Dr. Alan (Lan) Aronson at the National Library of Medicine (NLM). The program enable the linking of biomedical text into UMLS entities.
  MetaMap uses an extraction mechanism based on symbolic, natural language processing (NLP) and computational-linguistic techniques.
  MetaMap is regarded as one of the main Medical Text Indexer (MTI) reference tool. It can be used for both semiautomatic and fully automatic EE [3].
  To use MetaMap on your project, first you need an account in NIH [19] and log in, after that you need to download the library to use Metamap.
  Then use the following function in Python to analyze the diagnosis:

```
from pymetamap import MetaMap # Import MetaMap library

# Load Metamap object
mm = MetaMap.get_instance(path +
    "Metamap/public_mm/bin/metamap18", version=mm_version_year)

# Extract the entities
concepts, error = mm.extract_concepts(text, [1, 1])
    entities = []
    for concept in concepts:
        preferred_name = concept[3]
        entities.append(preferred_name.lower())
```

- SciSpacy[18]: is a specialized NLP library to process biomedical texts. It is built on the spaCy library and it concludes an extended support for part of speech (POS) tagging, dependency parsing, named entity recognition (NER) and sentence segmentation.
  To use SciSpacy you need to get the SciSpacy library with your favorite SciSpacy model available at https://allenai.github.io/scispacy.
  Then use the following function in Python to analyze the diagnosis:

```
# Import SciSpacy libraries
import scispacy
import spacy

# Load the SciSpacy object
nlp = spacy.load(python_path +
    "/site-packages/en_core_sci_lg/en_core_sci_lg-0.2.4")

# Extract the entities
    doc = nlp(str(text))
    concepts = doc.ents
    entities = []
    for concept in concepts:
        entities.append(str(concept).lower())
```

- TagMe[7]: is a powerful tool that, on-the-fly and with high precision/recall, annotates short text chunks with hyperlinks to Wikipedia contents.
  Tagme uses Wikipedia anchor texts as key and the correlated pages as possible concept descriptions.
  Tests resulted in a confirmation of the good classification performance and the claimed computation efficiency.
  First you need to get an identification token at https://sobigdata.d4science.org/web/tagme/tagme-help, with a previously login at the site linked.
  Then use the following function in Python to analyze the diagnosis:

```
import tagme # Import Tagme Library

# Call the function (with italian language)
tagme.GCUBE_TOKEN = # Insert here the TagMe Token
concepts = tagme.annotate(str(text), lang="it")

# Extract the entities
entities = []
for concept in concepts.get_annotations(0.01):
    entities.append(concept.entity_title.lower())
```

All the previous tools accept only the English language (except for Tagme that accept the Italian language too), a translation from the diagnosis in Italian language to the English language is needed.
For the translation is been used DeepL[4] in a manually way (with copy and paste), because the use of the API of DeepL are beyond payment.
To test the Italian entity extraction algorithm of TagMe, after the extraction, the records have been translated with the API of Google Translate[9] using the 3.2 code in Python, with the attention to parse the result because you obtain the following result:

```
translated(src=it, dest=en, text=dysplasia,
pronunciation=dysplasia, extra\_data=\\"\{'translat...\\")
```

The Google Translate snippet code is as follows:

Listing 3.2. Google Translate snippet script

```
from googletrans import Translator #Google Translator library

# Translation with Google Translate function
def googtran(text_ita):
    # Use the library
    translator = Translator()
    text_eng_ugly =str( translator.translate(text_ita, dest="en",
        src="it"))
    # Clean the result
    start = text_eng_ugly.find(", text=") + len(", text=")
    end = text_eng_ugly.find(", pronunciation=")
    text_eng = text_eng_ugly[start:end]
    return text_eng
```

An other way to translate the records, it is using the Italian Wikipedia[33] entries and search the relative English Wikipedia entries. In Python is used the following code:

```
import wikipediaapi # Wiki library

# Translation with wikipedia function
def wikitran(text_ita):
    wiki_wiki = wikipediaapi.Wikipedia("it",
        extract_format=wikipediaapi.ExtractFormat.WIKI)
    wiki_ita = wiki_wiki.page(text_ita)
    try:
        wiki_eng = wiki_ita.langlinks["en"]
        return str(wiki_eng.title)
    except:
        return ""
```

Eventually, to save the result data, is created a JSON file for each export method with the result of the diagnosis cervix/19/1795:

- This is the list of record using the DeepL translated English diagnosis on MetaMap:


```
"diagnosis": "exa:cervix/19/1795",
"concepts": [
   "human papilloma virus infection",
   "mucous membrane",
   "crushing injury",
   "associated with",
   "mild dysplasia",
   "related personal status",
   "relationships",
   "glandular epithelium",
   "fragment of (qualifier value)",
   "fragments",
   "specimen source codes - mucosa",
   "crushing procedure",
   "extended",
   "extension",
   "extent",
   "metaplastic",
   "endocervical mucosa",
   "alteration"
]
```


- This is the list of record using the DeepL translated English diagnosis on SciSpacy:


```
"diagnosis": "exa:cervix/19/1795",
"concepts": [
   "exocervical mucosa",
   "cytopathic",
   "alterations",
   "hpv infection",
   "metaplastic epithelium",
   "glandular crushed",
   "associated with",
   "mild dysplasia",
   "fragments",
   "exo-endocervical mucosa"
```

```
]
```

- This is the list of record using the DeepL translated English diagnosis on TagMe:

```
"diagnosis": "exa:cervix/19/1795",
"concepts": [
   "mucous membrane",
   "cytopathic effect",
   "human papillomavirus",
   "infection",
   "metaplasia",
   "epithelium",
   "gland",
   "dysplasia",
   "mucous membrane"
]
```

- This is the list of record using the Italian diagnosis on Tagme and the record translate using Google Translate:

```
"diagnosis": "exa:cervix/19/1795",
"concepts": [
   "mucosa",
   "alteration (music)",
   "infection",
   "human papilloma virus",
   "epithelial tissue",
   "gland",
   "dysplasia",
   "fragments of ...",
   "fragments of ...",
   "mucosa"
]
```

- This is the list of record using the Italian diagnosis on Tagme and the record translate using Wikipedia method:

```
"diagnosis": "exa:cervix/19/1795",
"concepts": [
   "mucous membrane",
   "accidental (music)",
```

```
    "infection",
    "human papillomavirus infection",
    "epithelium",
    "gland",
    "dysplasia",
    "",
    "",
    "mucous membrane"
]
```

The MetaMap list and the SciSpacy list have a lot of medical term, although a bit different from each other. Instead the three lists derivate from TagMe are very similar with each other, but with the DeepL one more similar to SciSpacy than the other two, which are very similar.

3.3.2 Entity merge

After obtain the lists of records from the different tools, to set an usable list of records for the entity linking step, is needed to merge the lists of records.
To do so we need compare the records with each other to define the similarity of the strings that th records are been defined.
The value the similarity are been used the following string distance:

- The Cosine distance used in this project is the similarity between phrases, not the words itself. So the two vectors represent a numeric value of the vectors of words.
  The value of the cosine distance is the follow (higher is better)[28]:

$$\text{cosine similarity} = S_C(A, B) := \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\mathbf{A}\mathbf{B}} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}, \quad (3.1)$$

In python is being implemented in this way:

```
# Import useful libraries
import math
import re
from collections import Counter

# Definition of the cosine similarity measure
```

```
def get_cosine(text1, text2):
    WORD = re.compile(r"\w+")
    vec1 = Counter(WORD.findall(text1))
    vec2 = Counter(WORD.findall(text2))
    intersection = set(vec1.keys()) & set(vec2.keys())
    numerator = sum([vec1[x] * vec2[x] for x in intersection])
    sum1 = sum([vec1[x] ** 2 for x in list(vec1.keys())])
    sum2 = sum([vec2[x] ** 2 for x in list(vec2.keys())])
    denominator = math.sqrt(sum1) * math.sqrt(sum2)
    if not denominator:
        return 0.0
    else:
        return float(numerator) / denominator

# Call of the function
value = get_cosine(String1, String2)
```

- The Hamming distance [22] is the measure of the number of characters that differ between two strings.
  Typically Hamming distance is undefined when strings are of different length, but this implementation considers extra characters as differing (lower is better).
  For example, the Hamming distance between abc and abcd is equal to 1.
  In Python, using the [23] library, is being implemented in this way:

```
import jellyfish #Import library
# Function call
value = jellyfish.hamming_distance(string1, string2)
```

- The Levenshtein distance [22] represents the number of insertions, deletions, and substitutions required to change one word to another (lower is better).
  For example, the Levenshtein distance between berne and born is equal to 2 and representing the transformation of the first e to o and the deletion of the second e.
  In Python, using the [23] library, is being implemented in this way:

```
import jellyfish #Import library
# Function call
value = jellyfish.levenshtein_distance(string1, string2)
```

- The Damerau Levenshtein distance [22] a modification of Levenshtein distance, Damerau-Levenshtein distance counts transpositions (such as ifsh

for fish) as a single edit (lower is better).
Where the Levenshtein distance between fish and ifsh is equal to 2 as it would require a deletion and an insertion, though the Damerau Levenshtein distance fish and ifsh is equal to 1 as this counts as a transposition. In Python, using the [23] library, is being implemented in this way:

```
import jellyfish #Import library
# Function call
value = jellyfish.damerau_levenshtein_distance(string1, string2)
```

- The Jaro distance [29] between the string s_1 and the string s_2 is defined as follow:

$$\begin{cases} 0 & \text{if } m = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right) & \text{otherwise} \end{cases} \tag{3.2}$$

where m is the number of matching character between s_1 and s_2 that are the same and not farther than $\left\lfloor \frac{\max(|s_1|,|s_2|)}{2} \right\rfloor - 1$ characters apart, and t the number of transposition between the two strings.
In Python, using the [23] library, is being implemented in this way:

```
import jellyfish #Import library
# Function call
value = jellyfish.jaro_distance(string1, string2)
```

For each distance measure, the records list is being combined from the several extraction method into one list of record.
To evaluated the several distance, for each diagnoses, for each distance and for each value of distance (now called threshold), with the average value between each initial lists, is being generated a list of records to put in output in a JSON file.

- For the diagnosis cervix/19/1795 the global list of records in JSON format is:

```
"method": "all",
"th": -1,
"record": [
   "mucous body substance",
   "finding",
   "show",
   "endocervical",
```

```
    "endocervical (intended site)",
    "endocervix",
    "specimen source codes - mucus",
    "mucus layer",
    "epithelial",
    "small size",
    "limited (extensiveness)",
    "limited lifting ability",
    "limited walking ability",
    "limited component (foundation metadata concept)",
    "alteration",
    "squamous",
    "usual",
    "dysplasia",
    "evidence",
    "dystrophic",
    "typical",
    "squamous mucus",
    "discous epithelials",
    "alterations",
    "canal of the cervix",
    "squamous epithelial cell",
    "mucus",
    "small intestine",
    "family",
    "television program",
    "dystrophic lake",
    "without evidence",
    "epithelium",
    "accidental (music)",
    "dystrophy",
    "epithelial tissue",
    "alteration (music)"
]
```

- For the diagnosis cervix/19/1795 the list of records using the cosine distance at the value of threshold equal to 0.48 circa in JSON format is:

```
"method": "cosine",
"th": 0.4787693700234704,
"record": [
    "specimen source codes - mucus",
```

```
    "mucus layer",
    "small size",
    "dysplasia",
    "evidence",
    "dystrophic",
    "squamous mucus",
    "squamous epithelial cell",
    "mucus",
    "dystrophic lake",
    "without evidence"
]
```

- For the diagnosis cervix/19/1795 the list of records using the Hamming distance at the value of threshold equal to 5.0 in JSON format is:

```
"method": "hamming",
"th": 5.0,
"record": [
    "show",
    "epithelial",
    "small size",
    "alteration",
    "usual",
    "dysplasia",
    "evidence",
    "dystrophic",
    "mucus",
    "family",
    "dystrophic lake",
    "dystrophy"
]
```

- For the diagnosis cervix/19/1795 the list of records using the Levenshtein distance at the value of threshold equal to 5.0 in JSON format is:

```
"method": "levenshtein",
"th": 5.0,
"record": [
    "finding",
    "show",
    "epithelial",
    "small size",
```

```
    "alteration",
    "squamous",
    "usual",
    "dysplasia",
    "evidence",
    "dystrophic",
    "typical",
    "mucus",
    "family",
    "dystrophic lake",
    "dystrophy"
]
```

- For the diagnosis cervix/19/1795 the list of records using the Damerau Levenshtein distance at the value of threshold equal to 5.0 in JSON format is:

```
"method": "damerau_levenshtein",
"th": 5.0,
"record": [
    "finding",
    "show",
    "epithelial",
    "small size",
    "alteration",
    "squamous",
    "usual",
    "dysplasia",
    "evidence",
    "dystrophic",
    "typical",
    "mucus",
    "family",
    "dystrophic lake",
    "dystrophy"
]
```

- For the diagnosis cervix/19/1795 the list of records using the Jaro distance at the value of threshold equal to 0.8 circa in JSON format is:

```
"method": "jaro",
"th": 0.8002886002886003,
```

```
"record": [
   "mucus layer",
   "epithelial",
   "small size",
   "alteration",
   "dysplasia",
   "dystrophic",
   "alterations",
   "mucus",
   "dystrophic lake",
   "dystrophy",
   "epithelial tissue"
]
```

How we can see the Levenshtein distance list and the Damerau Levenshtein distance are equal, with a bit less element from the Hamming distance. Instead the Jaro distance list has the same number of record of the cosine distance list, but the former has a higher threshold respect the latter.

3.3.3 Entity linking

Merging web data with knowledge bases is one of the fundamental step to achieve the vision of Semantic Web (section 2.4). To do such operation, Entity Linking (EL), or named entity disambiguation in the NLP community, is thus developed to link named entity mentions appearing in Web texts with their corresponding entities in knowledge bases.
It enables to perform tasks like knowledge base population, question answering, and information integration. To insert newly extracted knowledge derived from the information extraction system into an existing knowledge base, a system that maps an entity mention associated to the corresponding entity in the knowledge base is needed.
EL needs to deal with challenges due to to name variations and entity ambiguity. Indeed, a named entity could have multiple forms, such as its full name, partial names, aliases, acronyms, and alternate spellings.
An EL system thus needs to to identify the correct mapping entities for entity mentions of various forms and to disambiguate the entity mention in the textual context and identify the mapping entity for each entity mention [20].

EL is used for a variety of applications among the following:

- Information extraction: named entities extracted are usually ambiguous, but a good disambiguation can be done by EL with a known knowledge base.

- Information retrieval: the main tasks are the entity-based search, and query disambiguation.

- Content analysis: EL is used to link users' topics of interest into a knowledge base of categories. In such a way, a topic search can be done.

- Question answering: EL is also used in question answering system to disambiguate entities in question texts and build a suitable query to answer to the question.

- Knowledge base population: to automatically populate and keep up to date existing knowledge bases with newly extracted facts to support Semantic Web and knowledge management techniques.

So in this project, the entity linking has been used for knowledge base population.

For accomplish the entity linking scope is needed to create the triple statements, so we need to find the predicate and the object of each statement.
The subject are already know because we know the shortname of the diagnosis, for example the exa:cervix/19/1795 diagnosis.
To find the remaining two elements, the procedure is the same for each diagnosis, merge method and threshold, described in the previous step.

To find the object we need to find the name in the ontology that is more similar to the name of the record that it will be analyzed. To do so the list of the ontology names are found in the database at the attribute *Name* of the table *ConceptMap*.
The record are being associated from the one of the list that has the minimal Levenshtein distance from the record or the one from the list that has the maximal Jaro distance from the record, from now the selected name from the list is described with the term of *extracted entity*.
Until the end of this subsection, in this project are being used both association, with similar procedure unless recalled.

Now to find the predicate we using the table *ConceptStatement* in the database and the shortname of the object that we extract.
To do so we need to use a recursive function where:

- If the shortname of the object is in the attribute *Subject* and the attribute *Predicate* is equal to *rdfs:subClassOf*, we recall the recursive function with the value of the attribute *Object* has a new parameter.

- If the shortname of the object is in the attribute *Object* and the attribute *Subject* is equal to *exa:CervixClinicalCaseReport*, we return the value of *Predicate* and this will be the return value of the previously called functions.
- Else return *not found*.

This is the function is being implemented in Python in this project:

```
# Path_list: list from ConceptStatement
# Value_name: the name to search the predicate
def find_path(path_list, value_name):
    for row in path_list:
        # Recursive condition
        if value_name == row[0] and row[1] == "rdfs:subClassOf":
            [found, predicate] = find_path(path_list, row[2])
            if found:
                return [True, predicate]
        # Return condition
        elif value_name == row[2] and row[0] ==
            "exa:CervixClinicalCaseReport":
            return [True, row[1]]
    return [False, ""]
```

In the end we have the list of the statement extracted automatically from the text for each diagnosis, method and threshold.
The missing statement are the literal statement and the statement that define our diagnosis as a *exa:CervixClinicalCaseReport*.

Chapter 4

Results Analysis

Arrived at this point, we have build the ontology, defined the ground truth and create the list of statement in an automatic method, the only this that remain to do is to evaluate this work with proper method.
In this chapter is being included the comparison method between the ground truth and the automatic extracted statement list, is also included the measures used, the results with a proper explanation and a comment about the results.

4.1 Evaluation methods

In this section, as said before, there are the method to compare the ground truth, the list of statements extracted in a manual way, and the list of statements extracted in an automatic way.
Like said in subsection 3.3.3, the list of record are defined for each diagnosis, for each distance measured in subsection 3.3.1 and with the relative threshold.
To compare the two list we need to check the follow conditions:

$$\begin{gathered} Subject_{GT} = Subject_{AE} \\ \wedge \\ Predicate_{GT} = Predicate_{AE} \\ \wedge \\ Object_{GT} = Object_{AE} \end{gathered}$$

with X_{GT} the X element of the ground truth and X_{AE} the X element of the automatic extracted entities.
To remember that both list have only unique statements.
In this thesis, the following Python code is being used:

```
# Function to examine the presence of the extracted statement in the
    ground truth statement
# name_in: the object value of the extracted statement
# id_in: the subject value of both the extracted statement and the
    ground truth statement
# predicate_of(name_in): the predicate of the statement name_in as
    object
def search_path(name_in, id_in):
    query_sql = "SELECT \"Subject\", \"Predicate\", \"Object\", \"Name\"
        " \
                "FROM cervix.\"ConceptMap\" JOIN cervix.\"Statement\" " \
                "ON cervix.\"ConceptMap\".\"Concept\" =
                    cervix.\"Statement\".\"Object\" " \
                "WHERE \"Subject\" = %s;"
    cur.execute(query_sql, (id_in, ))
    rows = cur.fetchall()
    found = False
    sub = ""
    pre = ""
    obj = ""
    for row in rows:
        if row[3] == name_in and predicate_of(name_in) == row[1]:
            found = True
            sub = row[0]
            pre = row[1]
            obj = row[2]
            break
    return = [found, sub, pre, obj]
```

Now that we compare the statement is now the time to have some measure. In the beginning we need to count the equal statement and have some initial score, that are defined as follow:

$$Count = \sum_{i=1}^{|GT|} \sum_{j=1}^{|AE|} x_{ij}$$

$$Score_{Lev} = \sum_{i=1}^{|GT|} \sum_{j=1}^{|AE|} x_{ij} \cdot \frac{1}{Dist_{Lev}(GT_i, AE_j)}$$

$$Score_{Jaro} = \sum_{i=1}^{|GT|} \sum_{j=1}^{|AE|} x_{ij} \cdot Dist_{Jaro}(GT_i, AE_j)$$

with $|GT|$ the number of statement presents in the ground truth list, $|AE|$ the number of statement presents in the automatic extracted entities list, $x_{ij} = 1$ if the i-th element of GT is equal to the j-th element of AE or $x_{ij} = 0$ otherwise, $Dist_{Lev}(GT_i, AE_j)$ the value of the Levenshtein distance between the i-th element of GT and the j-th element of AE and $Dist_{Jaro}(GT_i, AE_j)$ the value of the Jaro distance between the i-th element of GT and the j-th element of AE.
To x_{ij} and count is not considered the literal statements and the (rdf:type) from the ground truth, because impossible to extract from only the diagnosis text or irrelevant to the automatic extraction.
To give every measure the concept higher is better, is being chosen the inverse number of the Levenshtein distance value, because the Levenshtein distance value ha a properties of lower is better.

The next measures used are the precision, recall and F_1 score [10] but to define them we need to describe their component first, as follow:

True Positive (TP) : are the statements that are being selected in both the ground truth list and the automatic extracted entities list.

True Negative (TN) : are the statements that are not being selected in both the ground truth list and the automatic extracted entities list.

False Positive (FP) : are the statements that are being selected only in the automatic extracted entities list.

False Negative (FN) : are the statements that are being selected only in the ground truth list.

The precision, recall and F_1 score, in this project, are defined in this way:

$$Precision = \frac{|TP|}{|TP| + |FP|} = \frac{Count}{|AE|}$$
$$Recall = \frac{|TP|}{|TP| + |FN|} = \frac{Count}{|GT|}$$
$$F_1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

4.2 Results

After the background, the process of developing the thesis and explaining the measurements, we arrived at the fateful results.
How we see in section 4.1, we have several measures applied in this section.

At the end of this project, we finally can generate the RDFs from the automatic extraction; to do so we use for each diagnosis, from the possible candidate of lists extracted, the list with the best score (see section 4.1) is chosen.
In the following example, is presented a few RDFs code represented as follow:

- The RDF of the Ground truth for the diagnosis *exa:cervix/19/1795* is the following:

Listing 4.1. Ground truth RDF

```
<exa:CervixClinicalCaseReport
    rdf:about="https://www.examode.eu/ontology/cervix/19/1795">
   <exa:hasTopography
      rdf:resource="https://www.examode.eu/ontology/uterusNOS"/>
   <exa:hasIntervention
      rdf:resource="http://purl.obolibrary.org/obo/NCIT_C15402"/>
   <exa:hasAge>36</exa:hasAge>
   <exa:identifier>19/1795</exa:identifier>
   <exa:isAboutDisease
      rdf:resource="http://purl.obolibrary.org/obo/MONDO_0002974"/>
   <exa:hasOutcome
      rdf:resource="http://purl.obolibrary.org/obo/NCIT_C40196"/>
   <exa:hasIntervention
      rdf:resource="http://purl.obolibrary.org/obo/NCIT_C15190"/>
   <exa:hasLocation rdf:resource=
   "http://purl.obolibrary.org/obo/UBERON_0012250"/>
   <exa:hasDiagnosis>Mucosa esocervicale con alterazioni
      citopatiche riferibili ad infezione da HPV estese
      all'epitelio metaplasico degli sfondati ghiandolari,
      associate a displasia lieve (frammenti di mucosa
      eso-endocervicale).</exa:hasDiagnosis>
   <exa:hasLocation rdf:resource=
   "http://purl.obolibrary.org/obo/UBERON_0000316"/>
   <exa:hasGender>FEMALE</exa:hasGender>
   <exa:detected rdf:resource=
   "http://purl.obolibrary.org/obo/NCBITaxon_10566"/>
</exa:CervixClinicalCaseReport>
```

- The RDF of the automatic extracted entities, with the Levenshtein distance used in the linking phase for the diagnosis *exa:cervix/19/1795* is the following:

Listing 4.2. Levenshtein distance RDF

```
<exa:CervixClinicalCaseReport
    rdf:about="https://www.examode.eu/ontology/cervix/19/1795">
```

```
   <exa:identifier>19/1795</exa:identifier>
   <exa:hasDiagnosis>Mucosa esocervicale con alterazioni
       citopatiche riferibili ad infezione da HPV estese
       all'epitelio metaplasico degli sfondati ghiandolari,
       associate a displasia lieve (frammenti di mucosa
       eso-endocervicale).</exa:hasDiagnosis>
   <exa:hasAge>36</exa:hasAge>
   <exa:hasGender>FEMALE</exa:hasGender>
   <exa:detected rdf:resource=
   "http://purl.obolibrary.org/obo/NCBITaxon_10566"/>
   <exa:hasIntervention
       rdf:resource="http://purl.obolibrary.org/obo/NCIT_C15402"/>
   <exa:hasOutcome
       rdf:resource="http://purl.obolibrary.org/obo/NCIT_C40196"/>
</exa:CervixClinicalCaseReport>
```

- The RDF of the automatic extracted entities, with the Jaro distance used in the linking phase for the diagnosis *exa:cervix/19/1795* is the following:

Listing 4.3. Jaro distance RDF

```
<exa:CervixClinicalCaseReport
    rdf:about="https://www.examode.eu/ontology/cervix/19/1795">
   <exa:identifier>19/1795</exa:identifier>
   <exa:hasDiagnosis>Mucosa esocervicale con alterazioni
       citopatiche riferibili ad infezione da HPV estese
       all'epitelio metaplasico degli sfondati ghiandolari,
       associate a displasia lieve (frammenti di mucosa
       eso-endocervicale).</exa:hasDiagnosis>
   <exa:hasAge>36</exa:hasAge>
   <exa:hasGender>FEMALE</exa:hasGender>
   <exa:detected rdf:resource=
   "http://purl.obolibrary.org/obo/NCBITaxon_10566"/>
   <exa:hasIntervention
       rdf:resource="http://purl.obolibrary.org/obo/NCIT_C15402"/>
   <exa:hasOutcome
       rdf:resource="http://purl.obolibrary.org/obo/NCIT_C40196"/>
   <exa:isAboutDisease
       rdf:resource="http://purl.obolibrary.org/obo/MONDO_0002974"/>
</exa:CervixClinicalCaseReport>
```

After seeing the resulting some RDFs achieved through the automatic extraction methods, some numerical results are now being shown.
First of all we analyze, for only the diagnosis *exa:cervix/19/1795*, how the trends

of the value of threshold influence the measures.
So, in the next two figures, are represented how the measures of precision (red line), recall (blue line) and F_1 (black line), are affected from the value of threshold.
Remember that the measures of the cosine distance and the Jaro distance are a number within the range [0,1] (with higher the value, best is the result), meanwhile the Hamming distance, the Levenshtein distance and the Damerau Levenshtein distance are a number ≥ 0 (with lower the value, best the result).
There are two figure, because in the linking phase are being used two distance, in the Figure 4.1 represented the use of the Levenshtein distance in the linking phase, meanwhile in the Figure 4.2 represented the use of the Jaro distance in the linking phase.

After noticing trough the plot how the threshold value impact the measure, the measure values are being show.
At the beginning, on the next two tables (in Table 4.1 and in Table 4.2), are shown, for each diagnosis, how all measures (count, score, precision, recall and F_1) values when all the automatic extracted entities are put together, without having no one measure and threshold impacted in the merge phase.
In Table 4.1 are presented the values of the measures when the Levenshtein distance is used in the linking phase with all the record extracted in the extract phase, meanwhile in Table 4.2 are presented the values of the measures when the Jaro distance is used in the linking phase with all the record.

For the rest of the following table are present the following values:

- For each table is used a different distance measure in the merge phase, while all use the Levenshtein distance for the linking phase.
- For each row is presented a different diagnosis.
- For each column is presented the average value and the maximal value for the relative measure (precision, recall, F_1).
- The average value and the maximal value are obtained from all the measure values obtain from the different threshold values.

For a typographic reason, in this section are included only the short version of the results, meanwhile in the Appendix B are included the results fully.

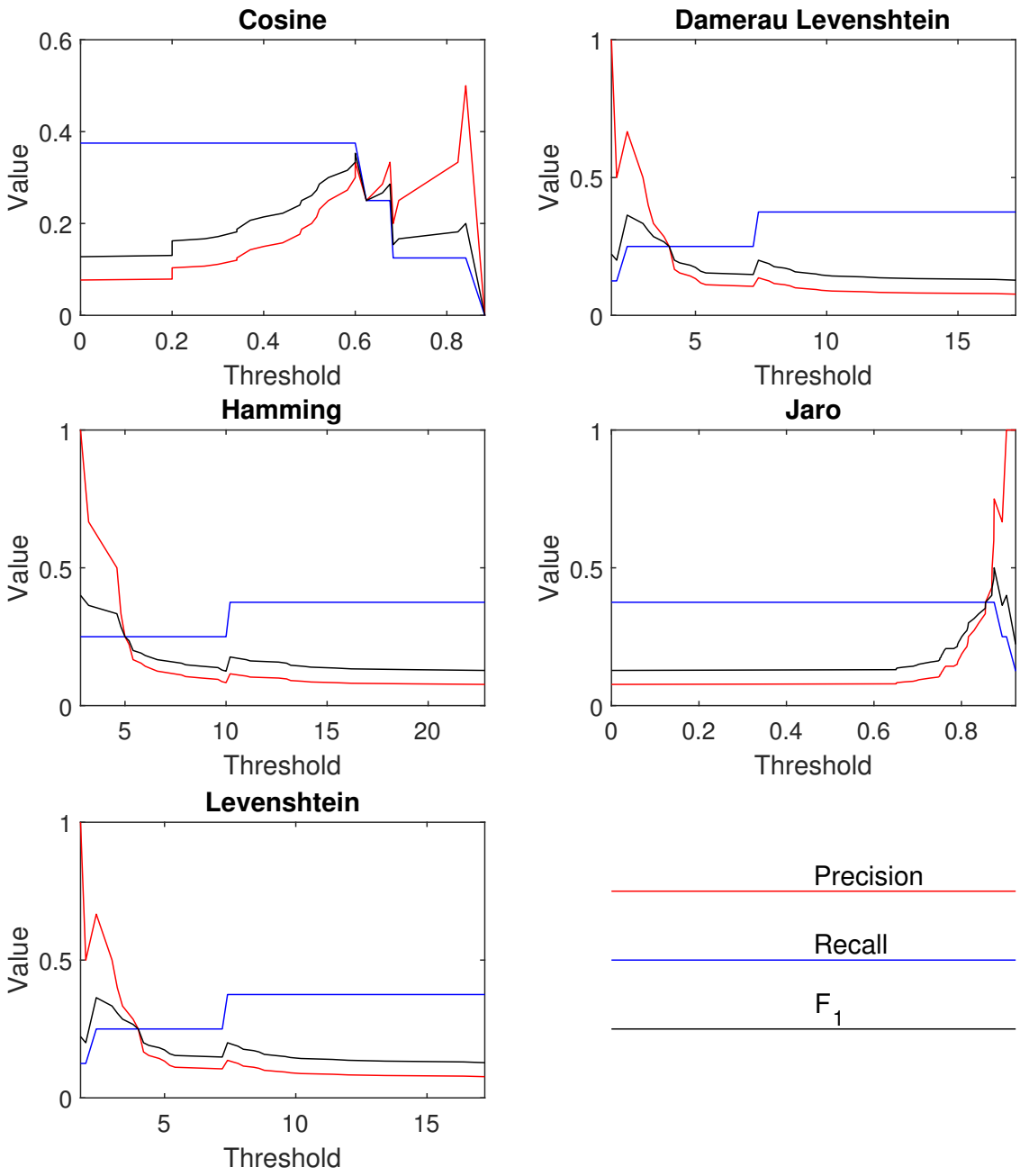

Figure 4.1. Plots of trends of various measures depend on threshold, using the Levenshtein distance in the linking phase.

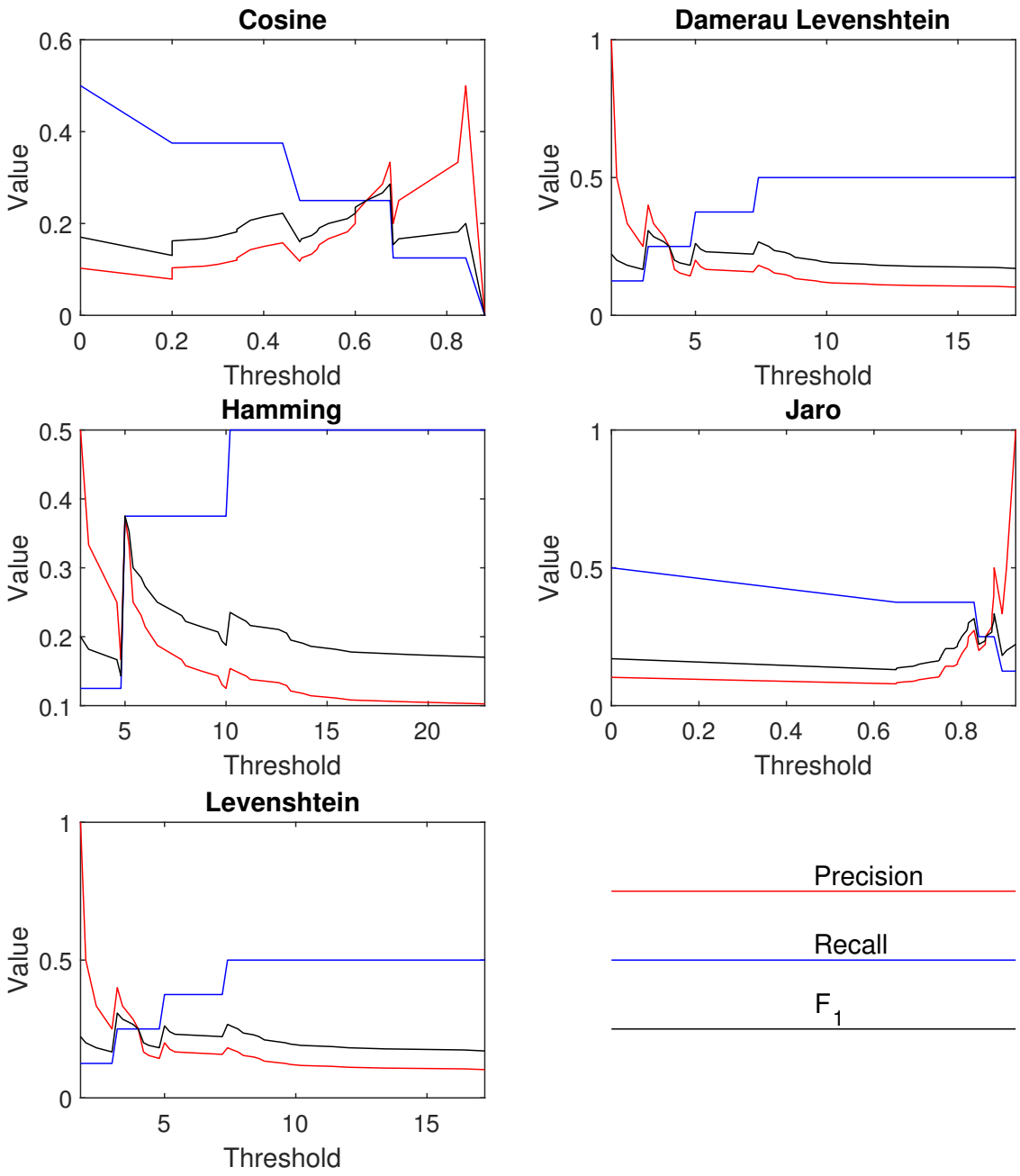

Figure 4.2. Plots of trends of various measures depend on threshold, using the Jaro distance in the linking phase

Id diagnosis	Count	Score	Precision	Recall	F_1
exa:cervix/19/1015	3	0.12564	0.081081	0.375	0.13333
exa:cervix/19/1016	3	0.089277	0.11111	0.375	0.17143
exa:cervix/19/1017	5	0.13333	0.1087	0.5	0.17857
exa:cervix/19/1803	4	0.13362	0.16	0.57143	0.25
exa:cervix/19/1795	3	0.092981	0.075	0.375	0.125
exa:cervix/19/1794	3	0.11453	0.085714	0.42857	0.14286
exa:cervix/19/1798	4	0.10951	0.097561	0.5	0.16327
exa:cervix/19/1792	3	0.08183	0.071429	0.5	0.125
exa:cervix/19/1012	3	0.11146	0.076923	0.42857	0.13043
exa:cervix/19/1802	5	0.0969	0.15625	0.71429	0.25641
exa:cervix/19/1011*	4	0.1086	0.083333	0.44444	0.14035
exa:cervix/19/985	5	0.082082	0.064935	0.625	0.11765
exa:cervix/19/1010_1A	3	0.11449	0.096774	0.42857	0.15789
exa:cervix/19/1010_2B	2	0.049517	0.038462	0.25	0.066667
exa:cervix/19/1000*	3	0.10317	0.058824	0.42857	0.10345
exa:cervix/19/1812*	1	0.083333	0.029412	0.16667	0.05
exa:cervix/19/1025*	3	0.1803	0.081081	0.375	0.13333
exa:cervix/19/280	4	0.10446	0.10526	0.5	0.17391
exa:cervix/19/975	2	0.076316	0.045455	0.33333	0.08
exa:cervix/19/1789	2	0.065476	0.11765	0.4	0.18182
exa:cervix/19/1793	3	0.079345	0.073171	0.42857	0.125
exa:cervix/19/1018	4	0.12946	0.095238	0.44444	0.15686
exa:cervix/19/966	3	0.10434	0.057692	0.42857	0.10169
exa:cervix/19/282*	3	0.10897	0.09375	0.42857	0.15385
exa:cervix/19/990	3	0.093074	0.076923	0.42857	0.13043
exa:cervix/19/1800	5	0.1169	0.17241	0.71429	0.27778
exa:cervix/19/969	5	0.10632	0.061728	0.55556	0.11111
exa:cervix/19/993	2	0.064777	0.064516	0.33333	0.10811
exa:cervix/19/995	2	0.13846	0.03125	0.28571	0.056338
exa:cervix/19/1797*	4	0.11762	0.066667	0.5	0.11765
exa:cervix/19/281	1	0.076923	0.03125	0.16667	0.052632
exa:cervix/19/997	5	0.11117	0.070423	0.625	0.12658
exa:cervix/19/1013	4	0.10417	0.097561	0.44444	0.16
exa:cervix/19/1807	2	0.063462	0.042553	0.25	0.072727
exa:cervix/19/1004	1	0.1	0.022222	0.16667	0.039216
exa:cervix/19/1006	3	0.12261	0.066667	0.42857	0.11538
exa:cervix/19/1024	4	0.097261	0.13793	0.5	0.21622
exa:cervix/19/1023	1	0.2	0.047619	0.16667	0.074074
exa:cervix/19/1811	1	0.083333	0.026316	0.2	0.046512
exa:cervix/19/1790	2	0.077295	0.058824	0.28571	0.097561
exa:cervix/19/1022*	4	0.11696	0.125	0.57143	0.20513
exa:cervix/19/279	3	0.1312	0.0625	0.5	0.11111
exa:cervix/19/1002	5	0.10632	0.071429	0.55556	0.12658
exa:cervix/19/1008*	4	0.13846	0.066667	0.5	0.11765
exa:cervix/19/1799	3	0.086488	0.083333	0.42857	0.13953
exa:cervix/19/1808	2	0.063462	0.044444	0.28571	0.076923
exa:cervix/19/996	3	0.076599	0.048387	0.42857	0.086957
exa:cervix/19/1020	4	0.15673	0.088889	0.5	0.15094
exa:cervix/19/1809	3	0.081197	0.13043	0.5	0.2069
exa:cervix/19/1014	4	0.10632	0.14815	0.57143	0.23529

Table 4.1. Table with all the entities extracted with Levenshtein distance used in linking phase.

Id diagnosis	Count	Score	Precision	Recall	F_1
exa:cervix/19/1015	2	0.62653	0.054054	0.25	0.088889
exa:cervix/19/1016	1	0.7232	0.037037	0.125	0.057143
exa:cervix/19/1017	4	0.6865	0.086957	0.4	0.14286
exa:cervix/19/1803	3	0.72445	0.12	0.42857	0.1875
exa:cervix/19/1795	4	0.52296	0.1	0.5	0.16667
exa:cervix/19/1794	3	0.49111	0.085714	0.42857	0.14286
exa:cervix/19/1798	4	0.502	0.097561	0.5	0.16327
exa:cervix/19/1792	4	0.64125	0.095238	0.66667	0.16667
exa:cervix/19/1012	3	0.73259	0.076923	0.42857	0.13043
exa:cervix/19/1802	4	0.69302	0.125	0.57143	0.20513
exa:cervix/19/1011*	4	0.66428	0.083333	0.44444	0.14035
exa:cervix/19/985	4	0.69745	0.051948	0.5	0.094118
exa:cervix/19/1010_1A	2	0.69755	0.064516	0.28571	0.10526
exa:cervix/19/1010_2B	3	0.62383	0.057692	0.375	0.1
exa:cervix/19/1000*	3	0.40143	0.058824	0.42857	0.10345
exa:cervix/19/1812*	2	0.62702	0.058824	0.33333	0.1
exa:cervix/19/1025*	4	0.71889	0.10811	0.5	0.17778
exa:cervix/19/280	4	0.69984	0.10526	0.5	0.17391
exa:cervix/19/975	4	0.61836	0.090909	0.66667	0.16
exa:cervix/19/1789	0	0	0	0	0
exa:cervix/19/1793	3	0.49111	0.073171	0.42857	0.125
exa:cervix/19/1018	2	0.63735	0.047619	0.22222	0.078431
exa:cervix/19/966	3	0.73648	0.057692	0.42857	0.10169
exa:cervix/19/282*	4	0.65923	0.125	0.57143	0.20513
exa:cervix/19/990	2	0.62683	0.051282	0.28571	0.086957
exa:cervix/19/1800	4	0.69984	0.13793	0.57143	0.22222
exa:cervix/19/969	5	0.53313	0.061728	0.55556	0.11111
exa:cervix/19/993	4	0.63932	0.12903	0.66667	0.21622
exa:cervix/19/995	3	0.70808	0.046875	0.42857	0.084507
exa:cervix/19/1797*	3	0.70988	0.05	0.375	0.088235
exa:cervix/19/281	3	0.66909	0.09375	0.5	0.15789
exa:cervix/19/997	4	0.49682	0.056338	0.5	0.10127
exa:cervix/19/1013	3	0.64101	0.073171	0.33333	0.12
exa:cervix/19/1807	4	0.61163	0.085106	0.5	0.14545
exa:cervix/19/1004	5	0.59699	0.11111	0.83333	0.19608
exa:cervix/19/1006	4	0.5153	0.088889	0.57143	0.15385
exa:cervix/19/1024	1	0.62077	0.034483	0.125	0.054054
exa:cervix/19/1023	1	0.67926	0.047619	0.16667	0.074074
exa:cervix/19/1811	2	0.62702	0.052632	0.4	0.093023
exa:cervix/19/1790	3	0.51539	0.088235	0.42857	0.14634
exa:cervix/19/1022*	3	0.6842	0.09375	0.42857	0.15385
exa:cervix/19/279	3	0.70988	0.0625	0.5	0.11111
exa:cervix/19/1002	4	0.51904	0.057143	0.44444	0.10127
exa:cervix/19/1008*	4	0.54333	0.066667	0.5	0.11765
exa:cervix/19/1799	2	0.33963	0.055556	0.28571	0.093023
exa:cervix/19/1808	4	0.60864	0.088889	0.57143	0.15385
exa:cervix/19/996	3	0.64075	0.048387	0.42857	0.086957
exa:cervix/19/1020	2	0.65788	0.044444	0.25	0.075472
exa:cervix/19/1809	1	0.58207	0.043478	0.16667	0.068966
exa:cervix/19/1014	2	0.77308	0.074074	0.28571	0.11765

Table 4.2. Table with all the entities extracted with Jaro distance used in linking phase.

Id diagnosis	Prec Avg	Prec Max	Rec Avg	Rec Max	F_1 Avg	F_1 Max
exa:cervix/19/1000*	0.05746	0.14286	0.20476	0.42857	0.08736	0.19048
exa:cervix/19/1002	0.16096	1	0.38889	0.55556	0.18307	0.26087
exa:cervix/19/1004	0.03666	0.09091	0.13043	0.16667	0.05532	0.11765
exa:cervix/19/1006	0.12362	0.25	0.35238	0.42857	0.16946	0.3
exa:cervix/19/1008*	0.15492	1	0.35	0.5	0.17331	0.25
exa:cervix/19/1010_1A	0.12197	0.33333	0.20833	0.28571	0.13928	0.21053
exa:cervix/19/1010_2B	0.07007	0.16667	0.16935	0.25	0.09216	0.14815
exa:cervix/19/1011*	0.23191	1	0.30303	0.44444	0.21137	0.33333
exa:cervix/19/1012	0.10595	0.25	0.25824	0.42857	0.13673	0.2
exa:cervix/19/1013	0.22963	1	0.23977	0.44444	0.181	0.26667
exa:cervix/19/1014	0.14153	0.25	0.26316	0.57143	0.17238	0.28571
exa:cervix/19/1015	0.03803	0.11765	0.10625	0.375	0.05518	0.16
exa:cervix/19/1016	0.1047	0.22222	0.18382	0.375	0.12895	0.23529
exa:cervix/19/1017	0.21639	0.5	0.288	0.5	0.21552	0.27027
exa:cervix/19/1018	0.16952	0.5	0.22778	0.44444	0.16332	0.27273
exa:cervix/19/1020	0.10305	0.15789	0.23958	0.5	0.13753	0.22222
exa:cervix/19/1022*	0.08276	0.21429	0.21978	0.57143	0.11656	0.28571
exa:cervix/19/1023	0.04267	0.11111	0.09524	0.16667	0.05791	0.13333
exa:cervix/19/1024	0.31411	1	0.36364	0.5	0.29216	0.4
exa:cervix/19/1025*	0.27035	1	0.30729	0.375	0.24024	0.33333
exa:cervix/19/1789	0.17976	0.33333	0.27273	0.4	0.20285	0.30769
exa:cervix/19/1790	0.09652	0.2	0.21675	0.28571	0.12777	0.23529
exa:cervix/19/1792	0.08258	0.16667	0.29885	0.5	0.12566	0.22222
exa:cervix/19/1793	0.12266	0.28571	0.31336	0.42857	0.16551	0.28571
exa:cervix/19/1794	0.11794	0.2	0.28571	0.42857	0.15942	0.23529
exa:cervix/19/1795	0.19916	0.5	0.3125	0.375	0.21524	0.35294
exa:cervix/19/1797*	0.12609	0.33333	0.30357	0.5	0.15641	0.25
exa:cervix/19/1798	0.18915	0.5	0.37054	0.5	0.21741	0.28571
exa:cervix/19/1799	0.18903	0.5	0.31818	0.42857	0.20632	0.33333
exa:cervix/19/1800	0.22625	0.5	0.35	0.71429	0.24967	0.36364
exa:cervix/19/1802	0.22622	0.4	0.37662	0.71429	0.2589	0.375
exa:cervix/19/1803	0.22846	0.5	0.32331	0.57143	0.24656	0.36364
exa:cervix/19/1807	0.02443	0.07407	0.09274	0.25	0.0383	0.11429
exa:cervix/19/1808	0.02912	0.07407	0.11735	0.28571	0.04615	0.11765
exa:cervix/19/1809	0.11356	0.2	0.24074	0.5	0.1476	0.28571
exa:cervix/19/1811	0.02132	0.05882	0.104	0.2	0.03513	0.09091
exa:cervix/19/1812*	0.02501	0.06667	0.09091	0.16667	0.03887	0.09524
exa:cervix/19/279	0.21032	1	0.43889	0.5	0.22881	0.36364
exa:cervix/19/280	0.18092	0.5	0.28879	0.5	0.18962	0.33333
exa:cervix/19/281	0.05919	0.16667	0.13492	0.16667	0.07757	0.16667
exa:cervix/19/282*	0.18965	0.5	0.27891	0.42857	0.19661	0.30769
exa:cervix/19/966	0.16255	0.5	0.36607	0.42857	0.19667	0.33333
exa:cervix/19/969	0.1293	1	0.35859	0.55556	0.1507	0.2
exa:cervix/19/975	0.06917	0.16667	0.22778	0.33333	0.10025	0.16667
exa:cervix/19/985	0.07005	0.125	0.30313	0.625	0.10735	0.14925
exa:cervix/19/990	0.12059	0.2	0.29121	0.42857	0.16111	0.27273
exa:cervix/19/993	0.03524	0.11111	0.11111	0.33333	0.05209	0.13333
exa:cervix/19/995	0.05863	0.16667	0.21429	0.28571	0.08632	0.16
exa:cervix/19/996	0.11403	0.33333	0.35135	0.42857	0.156	0.23529
exa:cervix/19/997	0.16224	1	0.42442	0.625	0.18827	0.3

Table 4.3. Short table with cosine distance used in merge phase with Levenshtein distance used in linking phase.

Id diagnosis	Prec Avg	Prec Max	Rec Avg	Rec Max	F_1 Avg	F_1 Max
exa:cervix/19/1000*	0.11533	1	0.26939	0.42857	0.11584	0.25
exa:cervix/19/1002	0.14418	1	0.38406	0.55556	0.15834	0.27273
exa:cervix/19/1004	0.0418	0.14286	0.15238	0.16667	0.06246	0.15385
exa:cervix/19/1006	0.1257	0.5	0.33641	0.42857	0.1588	0.33333
exa:cervix/19/1008*	0.12638	1	0.3125	0.5	0.13343	0.22222
exa:cervix/19/1010_1A	0.10801	0.33333	0.23214	0.28571	0.13625	0.21053
exa:cervix/19/1010_2B	0.03428	0.08333	0.15625	0.25	0.05574	0.125
exa:cervix/19/1011*	0.16408	1	0.34758	0.44444	0.18301	0.25
exa:cervix/19/1012	0.10612	0.33333	0.30476	0.42857	0.14452	0.2
exa:cervix/19/1013	0.17589	1	0.3037	0.44444	0.1745	0.21053
exa:cervix/19/1014	0.18797	0.33333	0.40714	0.57143	0.24339	0.33333
exa:cervix/19/1015	0.06119	0.10714	0.19167	0.375	0.09127	0.16667
exa:cervix/19/1016	0.07966	0.14286	0.19079	0.375	0.11125	0.19355
exa:cervix/19/1017	0.18584	1	0.32647	0.5	0.19006	0.23529
exa:cervix/19/1018	0.1047	0.25	0.26882	0.44444	0.1392	0.2069
exa:cervix/19/1020	0.08098	0.13636	0.30303	0.5	0.12588	0.2
exa:cervix/19/1022*	0.17622	0.33333	0.42286	0.57143	0.23593	0.31579
exa:cervix/19/1023	0.07444	0.2	0.13158	0.16667	0.09084	0.18182
exa:cervix/19/1024	0.15757	0.33333	0.30952	0.5	0.18936	0.26667
exa:cervix/19/1025*	0.18012	1	0.29924	0.375	0.18408	0.26087
exa:cervix/19/1789	0.01548	0.125	0.04706	0.4	0.02328	0.19048
exa:cervix/19/1790	0.11096	0.25	0.25824	0.28571	0.14408	0.25
exa:cervix/19/1792	0.0153	0.07692	0.09444	0.5	0.02631	0.13333
exa:cervix/19/1793	0.0902	0.33333	0.24603	0.42857	0.11604	0.2
exa:cervix/19/1794	0.14366	1	0.24138	0.42857	0.13532	0.25
exa:cervix/19/1795	0.18488	1	0.30603	0.375	0.18582	0.4
exa:cervix/19/1797*	0.05896	0.16667	0.24128	0.5	0.08901	0.14286
exa:cervix/19/1798	0.19017	0.66667	0.37903	0.5	0.2158	0.36364
exa:cervix/19/1799	0.13982	1	0.25238	0.42857	0.13573	0.25
exa:cervix/19/1800	0.21898	0.66667	0.40571	0.71429	0.25	0.4
exa:cervix/19/1802	0.21256	0.66667	0.4	0.71429	0.23753	0.4
exa:cervix/19/1803	0.24676	0.66667	0.37662	0.57143	0.26254	0.4
exa:cervix/19/1807	0.01509	0.05263	0.07566	0.25	0.02512	0.08696
exa:cervix/19/1808	0.01346	0.05405	0.07589	0.28571	0.02285	0.09091
exa:cervix/19/1809	0.09224	0.15789	0.23684	0.5	0.13033	0.24
exa:cervix/19/1811	0.02405	0.05556	0.12353	0.2	0.03997	0.08696
exa:cervix/19/1812*	0.02303	0.05882	0.09195	0.16667	0.03656	0.08696
exa:cervix/19/279	0.13055	1	0.32381	0.5	0.13434	0.28571
exa:cervix/19/280	0.11986	0.5	0.25	0.5	0.13483	0.2
exa:cervix/19/281	0.01593	0.04545	0.07051	0.16667	0.02594	0.07143
exa:cervix/19/282*	0.15592	1	0.23377	0.42857	0.14049	0.25
exa:cervix/19/966	0.14323	1	0.31513	0.42857	0.14444	0.28571
exa:cervix/19/969	0.11833	1	0.34722	0.55556	0.12777	0.2
exa:cervix/19/975	0.02095	0.06667	0.125	0.33333	0.03581	0.11111
exa:cervix/19/985	0.0301	0.06667	0.2217	0.625	0.05267	0.12048
exa:cervix/19/990	0.04601	0.09091	0.17857	0.42857	0.07198	0.15
exa:cervix/19/993	0.06293	0.14286	0.17949	0.33333	0.08936	0.15385
exa:cervix/19/995	0.0325	0.09091	0.16717	0.28571	0.05218	0.11111
exa:cervix/19/996	0.03535	0.08333	0.19601	0.42857	0.05817	0.10526
exa:cervix/19/997	0.1309	1	0.41071	0.625	0.15216	0.22222

Table 4.4. Short table with Hamming distance used in merge phase with Levenshtein distance used in linking phase.

Id diagnosis	Prec Avg	Prec Max	Rec Avg	Rec Max	F_1 Avg	F_1 Max
exa:cervix/19/1000*	0.09233	0.5	0.26531	0.42857	0.11146	0.22222
exa:cervix/19/1002	0.13854	1	0.38647	0.55556	0.15427	0.22222
exa:cervix/19/1004	0.04284	0.14286	0.15054	0.16667	0.06274	0.15385
exa:cervix/19/1006	0.12021	0.4	0.33766	0.42857	0.15737	0.33333
exa:cervix/19/1008*	0.12415	1	0.29878	0.5	0.13014	0.22222
exa:cervix/19/1010_1A	0.11658	0.28571	0.2517	0.28571	0.15112	0.28571
exa:cervix/19/1010_2B	0.03751	0.09091	0.16532	0.25	0.06044	0.13333
exa:cervix/19/1011*	0.17216	1	0.34877	0.44444	0.18245	0.27273
exa:cervix/19/1012	0.12002	0.33333	0.31905	0.42857	0.15694	0.21429
exa:cervix/19/1013	0.19932	1	0.30108	0.44444	0.18683	0.30769
exa:cervix/19/1014	0.19437	0.4	0.40714	0.57143	0.25105	0.375
exa:cervix/19/1015	0.05911	0.10345	0.19828	0.375	0.0903	0.16216
exa:cervix/19/1016	0.11384	0.2	0.22727	0.375	0.14557	0.21053
exa:cervix/19/1017	0.20745	1	0.34444	0.5	0.20494	0.28571
exa:cervix/19/1018	0.09977	0.2	0.27556	0.44444	0.13773	0.24
exa:cervix/19/1020	0.07611	0.13636	0.29643	0.5	0.11977	0.2
exa:cervix/19/1022*	0.14913	0.25	0.39429	0.57143	0.20868	0.31579
exa:cervix/19/1023	0.07087	0.16667	0.13158	0.16667	0.08946	0.16667
exa:cervix/19/1024	0.15007	0.33333	0.29348	0.5	0.18053	0.27586
exa:cervix/19/1025*	0.1848	1	0.32083	0.375	0.19061	0.3
exa:cervix/19/1789	0.04951	0.15385	0.1375	0.4	0.07269	0.22222
exa:cervix/19/1790	0.13748	0.5	0.2619	0.28571	0.15468	0.30769
exa:cervix/19/1792	0.01729	0.07692	0.10417	0.5	0.0296	0.13333
exa:cervix/19/1793	0.08958	0.25	0.24286	0.42857	0.11743	0.18182
exa:cervix/19/1794	0.1559	1	0.26667	0.42857	0.14905	0.25
exa:cervix/19/1795	0.20863	1	0.30645	0.375	0.19198	0.36364
exa:cervix/19/1797*	0.07592	0.33333	0.26645	0.5	0.10093	0.18182
exa:cervix/19/1798	0.2038	1	0.36638	0.5	0.20498	0.28571
exa:cervix/19/1799	0.11452	0.5	0.24868	0.42857	0.13007	0.22222
exa:cervix/19/1800	0.21109	0.5	0.375	0.71429	0.23873	0.30769
exa:cervix/19/1802	0.21461	0.66667	0.40571	0.71429	0.24091	0.4
exa:cervix/19/1803	0.22428	0.5	0.35714	0.57143	0.24466	0.33333
exa:cervix/19/1807	0.01319	0.05	0.06818	0.25	0.02209	0.08333
exa:cervix/19/1808	0.01379	0.05263	0.07792	0.28571	0.02342	0.08889
exa:cervix/19/1809	0.0888	0.15385	0.225	0.5	0.1247	0.23077
exa:cervix/19/1811	0.02193	0.05556	0.11724	0.2	0.03672	0.08696
exa:cervix/19/1812*	0.01996	0.05556	0.08333	0.16667	0.03203	0.08333
exa:cervix/19/279	0.11839	1	0.30882	0.5	0.12761	0.28571
exa:cervix/19/280	0.12601	0.5	0.27232	0.5	0.14308	0.2
exa:cervix/19/281	0.01559	0.04545	0.06944	0.16667	0.02541	0.07143
exa:cervix/19/282*	0.17338	1	0.25143	0.42857	0.1569	0.25
exa:cervix/19/966	0.12373	0.5	0.32707	0.42857	0.14775	0.36364
exa:cervix/19/969	0.11366	1	0.36	0.55556	0.12561	0.2
exa:cervix/19/975	0.02677	0.0625	0.14865	0.33333	0.04514	0.10526
exa:cervix/19/985	0.03054	0.06667	0.23295	0.625	0.05377	0.12048
exa:cervix/19/990	0.04881	0.1	0.17734	0.42857	0.07453	0.13953
exa:cervix/19/993	0.06327	0.14286	0.17901	0.33333	0.0895	0.15385
exa:cervix/19/995	0.03543	0.11111	0.18214	0.28571	0.0565	0.125
exa:cervix/19/996	0.06589	0.16667	0.30357	0.42857	0.10088	0.15385
exa:cervix/19/997	0.12904	1	0.41274	0.625	0.15211	0.22222

Table 4.5. Short table with Levenshtein distance used in merge phase with Levenshtein distance used in linking phase.

Id diagnosis	Prec Avg	Prec Max	Rec Avg	Rec Max	F_1 Avg	F_1 Max
exa:cervix/19/1000*	0.09233	0.5	0.26531	0.42857	0.11146	0.22222
exa:cervix/19/1002	0.13815	1	0.38647	0.55556	0.15404	0.22222
exa:cervix/19/1004	0.04284	0.14286	0.15054	0.16667	0.06274	0.15385
exa:cervix/19/1006	0.11944	0.4	0.34034	0.42857	0.15726	0.33333
exa:cervix/19/1008*	0.12412	1	0.29878	0.5	0.13009	0.22222
exa:cervix/19/1010_1A	0.11658	0.28571	0.2517	0.28571	0.15112	0.28571
exa:cervix/19/1010_2B	0.03751	0.09091	0.16532	0.25	0.06044	0.13333
exa:cervix/19/1011*	0.17459	1	0.34641	0.44444	0.18213	0.26087
exa:cervix/19/1012	0.12002	0.33333	0.31905	0.42857	0.15694	0.21429
exa:cervix/19/1013	0.19932	1	0.30108	0.44444	0.18683	0.30769
exa:cervix/19/1014	0.1927	0.4	0.40714	0.57143	0.24994	0.35294
exa:cervix/19/1015	0.05911	0.10345	0.19828	0.375	0.0903	0.16216
exa:cervix/19/1016	0.11384	0.2	0.22727	0.375	0.14557	0.21053
exa:cervix/19/1017	0.20745	1	0.34444	0.5	0.20494	0.28571
exa:cervix/19/1018	0.09977	0.2	0.27556	0.44444	0.13773	0.24
exa:cervix/19/1020	0.07611	0.13636	0.29643	0.5	0.11977	0.2
exa:cervix/19/1022*	0.14913	0.25	0.39429	0.57143	0.20868	0.31579
exa:cervix/19/1023	0.07087	0.16667	0.13158	0.16667	0.08946	0.16667
exa:cervix/19/1024	0.15007	0.33333	0.29348	0.5	0.18053	0.27586
exa:cervix/19/1025*	0.1848	1	0.32083	0.375	0.19061	0.3
exa:cervix/19/1789	0.04951	0.15385	0.1375	0.4	0.07269	0.22222
exa:cervix/19/1790	0.13656	0.5	0.2619	0.28571	0.15407	0.30769
exa:cervix/19/1792	0.01724	0.07692	0.10417	0.5	0.02954	0.13333
exa:cervix/19/1793	0.08958	0.25	0.24286	0.42857	0.11743	0.18182
exa:cervix/19/1794	0.1559	1	0.26667	0.42857	0.14905	0.25
exa:cervix/19/1795	0.20863	1	0.30645	0.375	0.19198	0.36364
exa:cervix/19/1797*	0.07537	0.33333	0.26923	0.5	0.10079	0.18182
exa:cervix/19/1798	0.20653	1	0.3625	0.5	0.20704	0.28571
exa:cervix/19/1799	0.11253	0.5	0.2449	0.42857	0.1284	0.22222
exa:cervix/19/1800	0.20931	0.5	0.38286	0.71429	0.2395	0.30769
exa:cervix/19/1802	0.21387	0.66667	0.41714	0.71429	0.2419	0.4
exa:cervix/19/1803	0.22428	0.5	0.35714	0.57143	0.24466	0.33333
exa:cervix/19/1807	0.01281	0.05	0.06618	0.25	0.02144	0.08333
exa:cervix/19/1808	0.01379	0.05263	0.07792	0.28571	0.02342	0.08889
exa:cervix/19/1809	0.0888	0.15385	0.225	0.5	0.1247	0.23077
exa:cervix/19/1811	0.02193	0.05556	0.11724	0.2	0.03672	0.08696
exa:cervix/19/1812*	0.01996	0.05556	0.08333	0.16667	0.03203	0.08333
exa:cervix/19/279	0.11839	1	0.30882	0.5	0.12761	0.28571
exa:cervix/19/280	0.1285	0.5	0.26339	0.5	0.14269	0.2
exa:cervix/19/281	0.01559	0.04545	0.06944	0.16667	0.02541	0.07143
exa:cervix/19/282*	0.17762	1	0.25	0.42857	0.15867	0.25
exa:cervix/19/966	0.12357	0.5	0.32601	0.42857	0.14823	0.36364
exa:cervix/19/969	0.11437	1	0.36054	0.55556	0.12557	0.2
exa:cervix/19/975	0.02677	0.0625	0.14865	0.33333	0.04514	0.10526
exa:cervix/19/985	0.03125	0.06667	0.23837	0.625	0.05502	0.12048
exa:cervix/19/990	0.04934	0.1	0.18095	0.42857	0.07556	0.13953
exa:cervix/19/993	0.06233	0.14286	0.17949	0.33333	0.0885	0.15385
exa:cervix/19/995	0.03533	0.11111	0.18118	0.28571	0.05637	0.125
exa:cervix/19/996	0.06586	0.16667	0.30357	0.42857	0.10083	0.15385
exa:cervix/19/997	0.12904	1	0.41274	0.625	0.15211	0.22222

Table 4.6. Short table with Levenshtein Damerau distance used in merge phase with Levenshtein distance used in linking phase.

Id diagnosis	Prec Avg	Prec Max	Rec Avg	Rec Max	F_1 Avg	F_1 Max
exa:cervix/19/1000*	0.08448	0.33333	0.26331	0.42857	0.11299	0.2
exa:cervix/19/1002	0.16546	1	0.42857	0.55556	0.19732	0.36364
exa:cervix/19/1004	0.05862	0.25	0.15909	0.16667	0.07739	0.2
exa:cervix/19/1006	0.16039	0.5	0.37143	0.42857	0.19617	0.33333
exa:cervix/19/1008*	0.16039	1	0.37292	0.5	0.1829	0.35294
exa:cervix/19/1010_1A	0.18667	1	0.26267	0.28571	0.17868	0.30769
exa:cervix/19/1010_2B	0.11313	1	0.21324	0.25	0.11395	0.22222
exa:cervix/19/1011*	0.21219	1	0.33796	0.44444	0.20896	0.375
exa:cervix/19/1012	0.12059	0.5	0.27839	0.42857	0.14493	0.22222
exa:cervix/19/1013	0.17548	1	0.27371	0.44444	0.1734	0.25
exa:cervix/19/1014	0.21353	1	0.2963	0.57143	0.20149	0.25
exa:cervix/19/1015	0.06811	0.14286	0.17568	0.375	0.09371	0.14815
exa:cervix/19/1016	0.10761	0.2	0.19907	0.375	0.13294	0.21053
exa:cervix/19/1017	0.1963	1	0.31957	0.5	0.19677	0.28571
exa:cervix/19/1018	0.14833	0.5	0.26455	0.44444	0.16242	0.25
exa:cervix/19/1020	0.1132	0.22222	0.30114	0.5	0.15205	0.23529
exa:cervix/19/1022*	0.08753	0.17647	0.25446	0.57143	0.12795	0.25
exa:cervix/19/1023	0.07211	0.2	0.12698	0.16667	0.08814	0.18182
exa:cervix/19/1024	0.20974	0.5	0.34483	0.5	0.23896	0.33333
exa:cervix/19/1025*	0.2395	1	0.33446	0.375	0.22776	0.36364
exa:cervix/19/1789	0.10109	0.22222	0.22353	0.4	0.13737	0.28571
exa:cervix/19/1790	0.15396	1	0.2395	0.28571	0.14881	0.25
exa:cervix/19/1792	0.03076	0.09091	0.16667	0.5	0.05169	0.15385
exa:cervix/19/1793	0.12858	0.33333	0.33798	0.42857	0.16991	0.26087
exa:cervix/19/1794	0.18039	1	0.31429	0.42857	0.18601	0.28571
exa:cervix/19/1795	0.25175	1	0.3625	0.375	0.24253	0.5
exa:cervix/19/1797*	0.10373	0.33333	0.27292	0.5	0.13044	0.28571
exa:cervix/19/1798	0.23428	1	0.40549	0.5	0.2399	0.4
exa:cervix/19/1799	0.12641	0.5	0.25794	0.42857	0.14717	0.22222
exa:cervix/19/1800	0.23126	0.5	0.40394	0.71429	0.26154	0.375
exa:cervix/19/1802	0.19795	0.42857	0.39286	0.71429	0.24082	0.42857
exa:cervix/19/1803	0.2502	0.5	0.37143	0.57143	0.26478	0.375
exa:cervix/19/1807	0.00805	0.05	0.04255	0.25	0.01353	0.08333
exa:cervix/19/1808	0.01123	0.04878	0.06032	0.28571	0.01889	0.08333
exa:cervix/19/1809	0.14516	0.33333	0.26812	0.5	0.17447	0.25
exa:cervix/19/1811	0.02646	0.06667	0.12632	0.2	0.04329	0.1
exa:cervix/19/1812*	0.0245	0.0625	0.09314	0.16667	0.03845	0.09091
exa:cervix/19/279	0.17712	1	0.41667	0.5	0.19909	0.44444
exa:cervix/19/280	0.16398	0.66667	0.26974	0.5	0.1735	0.36364
exa:cervix/19/281	0.02796	0.07143	0.09896	0.16667	0.04318	0.1
exa:cervix/19/282*	0.23052	1	0.29464	0.42857	0.20945	0.44444
exa:cervix/19/966	0.18277	1	0.40385	0.42857	0.20718	0.42857
exa:cervix/19/969	0.13286	1	0.39781	0.55556	0.16274	0.3
exa:cervix/19/975	0.01441	0.05405	0.08712	0.33333	0.02468	0.09302
exa:cervix/19/985	0.04329	0.07143	0.24178	0.625	0.07121	0.12658
exa:cervix/19/990	0.08788	0.16667	0.24908	0.42857	0.12358	0.21053
exa:cervix/19/993	0.08382	0.33333	0.16667	0.33333	0.09951	0.22222
exa:cervix/19/995	0.07325	0.22222	0.25893	0.28571	0.10491	0.25
exa:cervix/19/996	0.08563	0.25	0.3341	0.42857	0.1249	0.19048
exa:cervix/19/997	0.15402	1	0.44542	0.625	0.18762	0.33333

Table 4.7. Short table with Jaro distance used in merge phase with Levenshtein distance used in linking phase.

After all of this verbosity of the results, a more compacted data is require to judge which distance used are the best for the future.
In the Table 4.8, are shown the following information:

- For each row are presented the singular distance used in the merge phase and the distance used in the linking phase.
- For each column are presented each measures used to evaluated the automatic extraction entities lists.
- Each value represent the count of how much, for all of the diagnosis, if the relative distances used in both merge phase and the linking phase obtain the maximal value for each diagnosis, if this is the cases is given 1 point. In a tie case is given both distance 1 point.

Table 4.8. Table shown which method obtain the best measures.

Merge	Linking	Count	$Score_{Lev}$	$Score_{Jaro}$	Precision	Recall	F_1
All	Leven.	50	50	0	0	50	0
All	Jaro	50	0	50	0	50	0
Cosine	Leven.	49	49	0	22	49	16
Cosine	Jaro	49	0	49	22	49	16
Hamming	Leven.	49	49	0	15	49	5
Hamming	Jaro	49	0	49	15	49	5
Leven.	Leven.	49	49	0	22	49	7
Leven.	Jaro	49	0	49	22	49	7
Dam.-Lev.	Leven.	49	49	0	32	49	25
Dam.-Lev.	Jaro	49	0	49	32	49	25
Jaro	Leven.	49	49	0	15	49	6
Jaro	Jaro	49	0	49	15	49	6

4.3 Analysis of the results

In this section all the result presented in section 4.2 and in the Appendix B will be analysed in a technical way.

Starting with the RDFs, it is possible to analyse from them, that form the Levenshtein RDF are being extracted 8 statements and in the Jaro RDF are 9, meanwhile in the ground truth are presents 13 statements.
Another fact is that all of the automatic statements, are the same in the manual statements.

The last thing that could probably deduct is that the Jaro one is better than the Levenshtein, but could be probably a case.

After that, from Figure 4.1 and from Figure 4.2 the trends of the plots are the same in both the figures, with a difference on the values.
Another notable information that is possible to see how with a more loose threshold the precision will drop and the recall will rise, with the exception of few threshold values.
Remember that too see how to loose the threshold, in the cosine and the Jaro distances is needed to decrease the threshold value, meanwhile in the other distances is needed to increase it.

From the tables presents in section 4.2 and in the Appendix B, is possible deduct the following facts:

- From some diagnosis are more easy to extract information, which the measures are relative high, instead of other diagnosis where harder, which the measures are relative low.
- It is possible to notice that in some field that when the average of a measure are pretty low and the maximal measure are a lot higher the deviation standard will be higher instead when the maximal value is similar to the average the deviation standard are lower.
- When the precision could reach the max value, some times could be when the automatic list contain only one element and that element is selected giving precision equal to 1, so the result would be altered.
- When the recall could reach the max value, some times could be when the automatic list contain all the possible elements, given a higher value respect the more restricted lists, so the result would be altered. Same with the count measure.
- To balance both the previous point, the F_1 measure could balance both the problematic thing of the previous 2 points, so it seem that it will give a better index of valuation, but the value are pretty low.

In the end, using the Table 4.8 as a method to evaluate the distance used in both the merge and linking phase and the measures is possible to assume that:

- Using the Levenshtein distance or the Jaro distance in the linking phase is irrelevant to choose the best list.
- The count, score and recall measures are all useless to select the best list.

- So the best measure to use seems to be both the F_1 and the precision, but considered what said before, the precision measure seems unstable, so the best measure to consider is the F_1 measure.
- After examined the previous point, is possible to deduct that the best distance to use in the merge phase is the *Damerau Levenshtein* distance, followed from the *cosine* distance.

Chapter 5

Conclusions

In the present thesis, we try to answer the following research questions:

1. Is it possible to automatically extract medical entities and statements, from several different clinical reports' diagnoses?
2. With respect to the manual extraction method, which is used for the construction of the ground-truth of reference, are the performances of automatic extraction methods competitive?

To answer the first question, first of all we used a multi-stages method to extract manually the entities in a systematic manner. First, we designed the ontology starting from the entities present in the UMLS database and in well-established medical ontologies (e.g., MONDO, NCBITaxon, and UBERON).
Secondly, we extracted the medical entities composing the statements (i.e., triples consisting of subject, predicate, and object) using several state-of-the-art knowledge extraction tools (e.g., MetaMap and TAGME). The entities extracted were later merged in a single list of distinct entities.
Then, we defined some translation rules mapping our dataset's entries to the new ontology-based entities. To achieve this, we relied on a relational PostgreSQL database to store our statements so that they could be converted afterwards in the form of structured RDF files.

Hence, the answer for the first question is affirmative, since using the tools for entities extraction, coupled with the adoption of several metrics of distance necessary for unifying the extraction process from different tools and linking the entities to the ontology, it is possible to generate statements from automatically extracted medical entities coming from reports' diagnoses.
Instead, for what concerns the second question, it is possible to obtain high values in terms of precision and recall measures, but typically not for both the measures at the same time. To observe this it is possible to see from the low results for the F_1 measure (see tables on section 4.2 and in Appendix B).

As seen in section 4.3, the main reason for this low performance could be attributed to the heterogeneous nature of the free-text clinical diagnoses, that poses hindrances to automatic knowledge extraction algorithms. For instance, we observed that some diagnoses are particularly challenging since they are typically short and affected by noise, thus making the extraction of the correct entities almost unfeasible.

As future work, it could be useful to investigate a hybrid approach combining both supervised and unsupervised methods to improve the performance of the knowledge extraction process. For instance, more sophisticated approaches could involve recent advancements on deep learning and natural language processing, including transformer and attention-based neural models.

Appendix A

Full ontology

Like it said in chapter 3, in this appendix is present the entire ontology used in the project of this thesis. The ontology refers only to the cervix cancer disease.

The entire figure is been split in 2 parts (on 2 pages), because of its size.

The left part (Figure A.1) contains the ExaMode use cases box and the diagnosis box.
Meanwhile the right part (Figure A.2) contains the procedure box, the anatomical box, the namespaces used in this project and a legend that explains the meaning of the colour of the entities and other information.

It is possible to see how the several boxes are connected with each other.

For more information on how is it been defined, refer to the chapter 3.

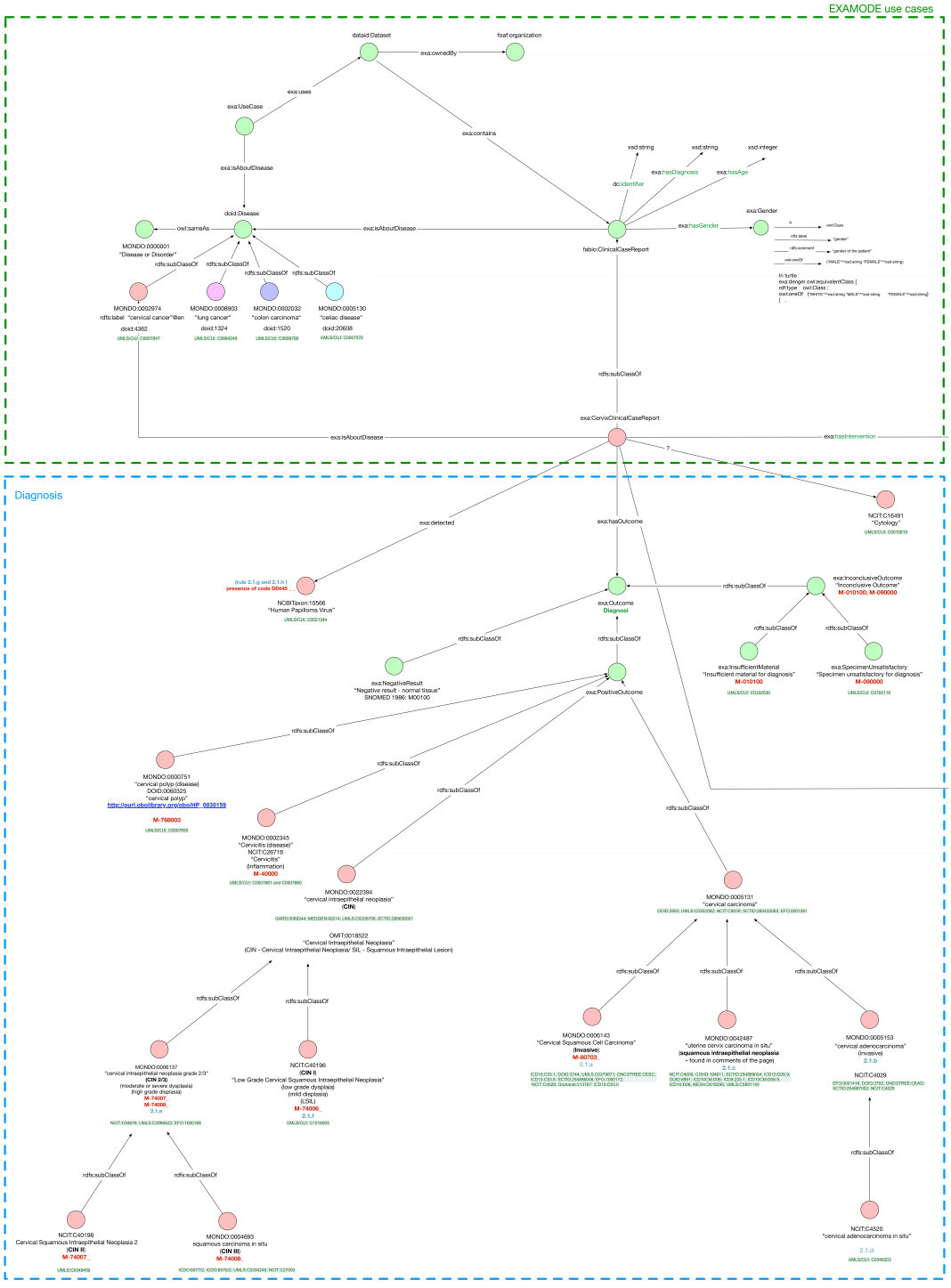

Figure A.1. Left side of the full ontology of this project.

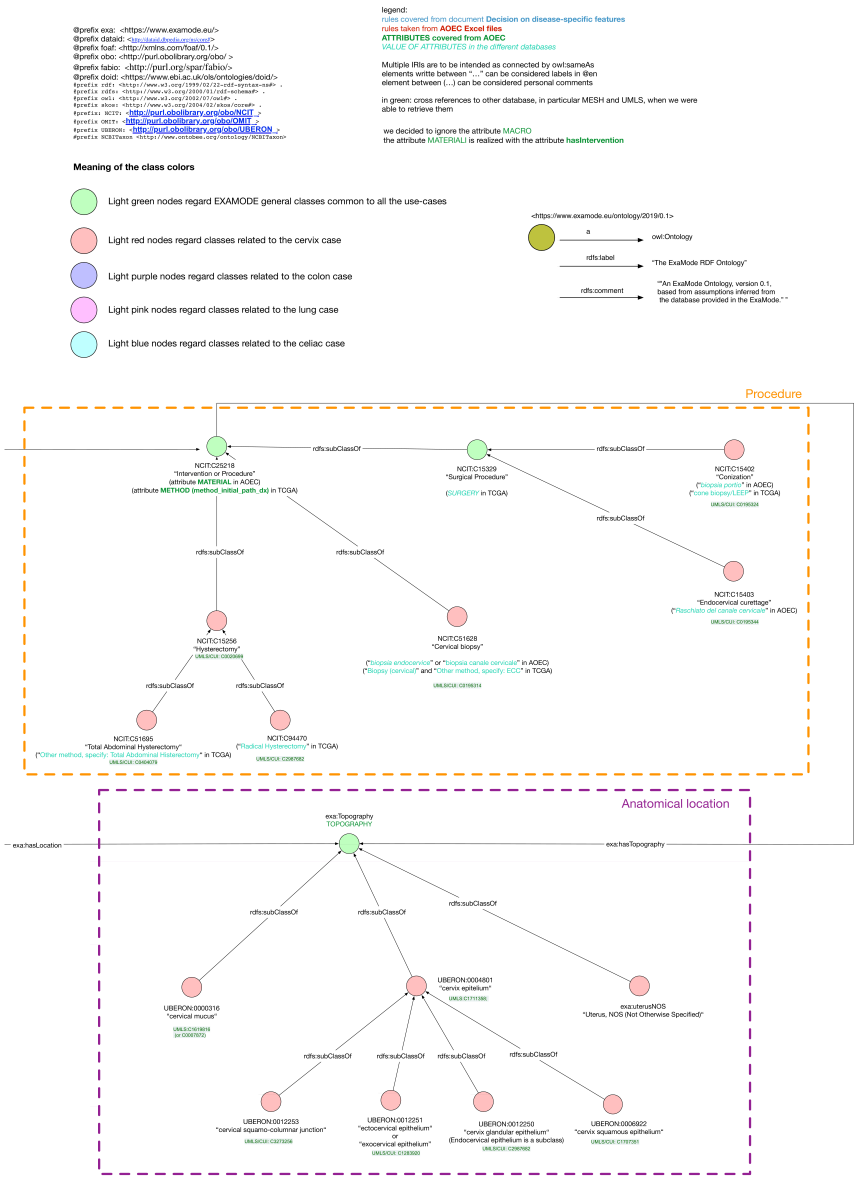

Figure A.2. Right side of the full ontology of this project.

Appendix B

Full result

As has said in section 4.2, for a typographic reason, in this section are included the full version of the results, meanwhile in the chapter 4 are included the shorter version of the results.

The tables present here are described as follow:

- A first mode to split the tables is with the distance used during the linking phase, which are the use of the Levenshtein distance or the Jaro distance.
- For a major readability the tables are been split again in the left side and the right side.
- The left side contain, for its column, the name of the diagnosis and the measures of count and the relative score (($Score_{Lev}$ when the Levenshtein distance is being used in the linking phase, while ($Score_{Jaro}$ when the Jaro distance is being used in the same phase).
- The right side contain, for its column, the precision, recall and the F_1 measures.
- For each row is presented a different diagnosis.
- For each measure is calculated he average value, the deviation standard and the maximal value from all the measure values obtain from the different threshold values, meanwhile is also presents the threshold value corresponding to the maximal value.

Id diagnosis	Count Avg	Count Dev	Count Max	Count At	Score Avg	Score Dev	Score Max	Score At
exa:cervix/19/1000*	1.4333	0.88889	3	0.2	0.0641	0.03846	0.12698	0.2
exa:cervix/19/1002	3.5	1.0952	5	0.34142	0.12931	0.02356	0.2	1
exa:cervix/19/1004	0.78261	0.34026	1	0.62426	0.0686	0.03117	0.11111	0.34142
exa:cervix/19/1006	2.4667	0.78222	3	0.56042	0.11521	0.01942	0.15897	0.34142
exa:cervix/19/1008*	2.8	0.67429	4	0.27071	0.14676	0.01549	0.2	1
exa:cervix/19/1010_1A	1.4583	0.58681	2	0.44142	0.1942	0.05725	0.25	0.64495
exa:cervix/19/1010_2B	1.3548	0.66597	2	0.44142	0.15539	0.0789	0.25	0.64495
exa:cervix/19/1011*	2.7273	0.60606	4	0.27071	0.16561	0.04004	0.225	0.64495
exa:cervix/19/1012	1.8077	0.97633	3	0.34142	0.16925	0.05938	0.25	0.64495
exa:cervix/19/1013	2.1579	1.2188	4	0.25	0.16401	0.03789	0.2	0.94142
exa:cervix/19/1014	1.8421	1.2022	4	0.34142	0.16067	0.06641	0.25	0.64495
exa:cervix/19/1015	0.85	1.02	3	0.2	0.05215	0.06258	0.15	0.34142
exa:cervix/19/1016	1.4706	0.97578	3	0.25774	0.06682	0.03931	0.1	0.4
exa:cervix/19/1017	2.88	1.0944	5	0.25	0.18219	0.02968	0.225	0.64495
exa:cervix/19/1018	2.05	1.055	4	0.2	0.1805	0.0556	0.25	0.64495
exa:cervix/19/1020	1.9167	1.0139	4	0.2	0.17009	0.05214	0.25	0.64495
exa:cervix/19/1022*	1.5385	1.432	4	0.2	0.07247	0.05575	0.14762	0.34142
exa:cervix/19/1023	0.57143	0.4898	1	0.4	0.09286	0.07959	0.2	0.34142
exa:cervix/19/1024	2.9091	1.0331	4	0.4	0.13008	0.0333	0.2	0.94142
exa:cervix/19/1025*	2.4583	0.63194	3	0.44142	0.19502	0.01873	0.225	0.64495
exa:cervix/19/1789	1.3636	0.57851	2	0.44495	0.06764	0.01427	0.08333	0.71962
exa:cervix/19/1790	1.5172	0.66587	2	0.54142	0.1412	0.07073	0.25	0.64495
exa:cervix/19/1792	1.7931	0.97503	3	0.32071	0.08598	0.04436	0.2	0.6633
exa:cervix/19/1793	2.1936	0.83247	3	0.47525	0.07483	0.01953	0.12179	0.65689
exa:cervix/19/1794	2	0.76923	3	0.5	0.12264	0.03454	0.2	0.8
exa:cervix/19/1795	2.5	0.71429	3	0.6	0.10758	0.02135	0.2	0.84142
exa:cervix/19/1797*	2.4286	1.0204	4	0.34142	0.10692	0.01899	0.2	0.84142
exa:cervix/19/1798	2.9643	1.0459	4	0.48284	0.12069	0.02297	0.2	0.84142
exa:cervix/19/1799	2.2273	0.49174	3	0.32071	0.10613	0.01955	0.16667	0.88284
exa:cervix/19/1800	2.45	1.085	5	0.2	0.12561	0.02687	0.2	0.7
exa:cervix/19/1802	2.6364	1.0661	5	0.2	0.13386	0.03112	0.2	0.7
exa:cervix/19/1803	2.2632	0.88643	4	0.34142	0.12695	0.02673	0.2	0.7
exa:cervix/19/1807	0.74194	0.81374	2	0.32845	0.02649	0.02905	0.06346	0.32845
exa:cervix/19/1808	0.82143	0.76276	2	0.32845	0.03063	0.02844	0.06346	0.32845
exa:cervix/19/1809	1.4444	0.92593	3	0.32845	0.06052	0.02017	0.08333	0.34142
exa:cervix/19/1811	0.52	0.4992	1	0.4	0.04667	0.0448	0.09091	0.4
exa:cervix/19/1812*	0.54545	0.49587	1	0.4	0.0489	0.04445	0.09091	0.4
exa:cervix/19/279	2.6333	0.56222	3	0.62426	0.11735	0.01481	0.2	0.84142
exa:cervix/19/280	2.3103	0.72533	4	0.27071	0.10567	0.01361	0.2	0.84142
exa:cervix/19/281	0.80952	0.30839	1	0.64142	0.06227	0.02372	0.07692	0.64142
exa:cervix/19/282*	1.9524	0.54875	3	0.27071	0.12037	0.03009	0.2	0.84142
exa:cervix/19/966	2.5625	0.54688	3	0.56042	0.10633	0.0129	0.16667	0.85689
exa:cervix/19/969	3.2273	1.2583	5	0.34142	0.12773	0.02078	0.2	1
exa:cervix/19/975	1.3667	0.63333	2	0.37071	0.07482	0.01995	0.1	0.62426
exa:cervix/19/985	2.425	1.2675	5	0.2	0.08581	0.01362	0.1	0.76569
exa:cervix/19/990	2.0385	1.0355	3	0.48284	0.06539	0.01574	0.09307	0.2
exa:cervix/19/993	0.66667	0.59259	2	0.2	0.03059	0.02719	0.06478	0.2
exa:cervix/19/995	1.5	0.63889	2	0.56042	0.08561	0.03897	0.13846	0.34142
exa:cervix/19/996	2.4595	0.73046	3	0.5	0.08846	0.02098	0.14286	0.88284
exa:cervix/19/997	3.3954	1.1022	5	0.32845	0.13325	0.02437	0.2	1

Table B.1. Left full table with cosine distance used in merge phase with Levenshtein distance used in linking phase.

Prec Avg	Prec Dev	Prec Max	Prec At	Rec Avg	Rec Dev	Rec Max	Rec At	F_1 Avg	F_1 Dev	F_1 Max	F_1 At
0.05746	0.03547	0.14286	0.54142	0.20476	0.12698	0.42857	0.2	0.08736	0.05311	0.19048	0.54142
0.16096	0.06914	1	1	0.38889	0.12169	0.55556	0.34142	0.18307	0.02874	0.26087	0.72426
0.03666	0.02204	0.09091	0.62426	0.13043	0.05671	0.16667	0.62426	0.05532	0.02912	0.11765	0.62426
0.12362	0.05071	0.25	0.72426	0.35238	0.11175	0.42857	0.56042	0.16946	0.05091	0.3	0.56042
0.15492	0.07699	1	1	0.35	0.08429	0.5	0.27071	0.17331	0.03884	0.25	0.72426
0.12197	0.04888	0.33333	0.64495	0.20833	0.08383	0.28571	0.44142	0.13928	0.03508	0.21053	0.44142
0.07007	0.03021	0.16667	0.64495	0.16935	0.08325	0.25	0.44142	0.09216	0.03438	0.14815	0.44142
0.23191	0.12718	1	0.8	0.30303	0.06734	0.44444	0.27071	0.21137	0.04591	0.33333	0.64495
0.10595	0.0357	0.25	0.64495	0.25824	0.13948	0.42857	0.34142	0.13673	0.04049	0.2	0.34142
0.22963	0.12261	1	0.94142	0.23977	0.13543	0.44444	0.25	0.181	0.04557	0.26667	0.25
0.14153	0.05401	0.25	0.64495	0.26316	0.17175	0.57143	0.34142	0.17238	0.07496	0.28571	0.4
0.03803	0.04563	0.11765	0.4	0.10625	0.1275	0.375	0.2	0.05518	0.06621	0.16	0.4
0.1047	0.06159	0.22222	0.38214	0.18382	0.12197	0.375	0.25774	0.12895	0.07632	0.23529	0.38214
0.21639	0.05509	0.5	0.88284	0.288	0.10944	0.5	0.25	0.21552	0.03222	0.27027	0.25
0.16952	0.07216	0.5	0.64495	0.22778	0.11722	0.44444	0.2	0.16332	0.04041	0.27273	0.4
0.10305	0.03257	0.15789	0.4	0.23958	0.12674	0.5	0.2	0.13753	0.04897	0.22222	0.4
0.08276	0.06366	0.21429	0.34142	0.21978	0.20456	0.57143	0.2	0.11656	0.09133	0.28571	0.34142
0.04267	0.03658	0.11111	0.4	0.09524	0.08163	0.16667	0.4	0.05791	0.04964	0.13333	0.4
0.31411	0.09178	1	0.94142	0.36364	0.12913	0.5	0.4	0.29216	0.0586	0.4	0.4
0.27035	0.12205	1	0.94142	0.30729	0.07899	0.375	0.44142	0.24024	0.04507	0.33333	0.64495
0.17976	0.06082	0.33333	0.71962	0.27273	0.1157	0.4	0.44495	0.20285	0.05005	0.30769	0.44495
0.09652	0.04601	0.2	0.54142	0.21675	0.09512	0.28571	0.54142	0.12777	0.0508	0.23529	0.54142
0.08258	0.04076	0.16667	0.60825	0.29885	0.1625	0.5	0.32071	0.12566	0.06122	0.22222	0.60825
0.12266	0.04811	0.28571	0.65689	0.31336	0.11892	0.42857	0.47525	0.16551	0.05155	0.28571	0.65689
0.11794	0.03559	0.2	0.67589	0.28571	0.10989	0.42857	0.5	0.15942	0.04488	0.23529	0.67589
0.19916	0.08323	0.5	0.84142	0.3125	0.08929	0.375	0.6	0.21524	0.05985	0.35294	0.6
0.12609	0.04905	0.33333	0.84142	0.30357	0.12755	0.5	0.34142	0.15641	0.03352	0.25	0.67589
0.18915	0.05299	0.5	0.84142	0.37054	0.13074	0.5	0.48284	0.21741	0.03283	0.28571	0.48284
0.18903	0.08901	0.5	0.88284	0.31818	0.07025	0.42857	0.32071	0.20632	0.03904	0.33333	0.72426
0.22625	0.07732	0.5	0.59424	0.35	0.155	0.71429	0.2	0.24967	0.063	0.36364	0.59424
0.22622	0.07366	0.4	0.59424	0.37662	0.1523	0.71429	0.2	0.2589	0.06123	0.375	0.5
0.22846	0.07952	0.5	0.59424	0.32331	0.12663	0.57143	0.34142	0.24656	0.06435	0.36364	0.59424
0.02443	0.0268	0.07407	0.32845	0.09274	0.10172	0.25	0.32845	0.0383	0.04201	0.11429	0.32845
0.02912	0.02704	0.07407	0.32845	0.11735	0.10897	0.28571	0.32845	0.04615	0.04285	0.11765	0.32845
0.11356	0.05182	0.2	0.62426	0.24074	0.15432	0.5	0.32845	0.1476	0.06765	0.28571	0.32845
0.02132	0.02047	0.05882	0.4	0.104	0.09984	0.2	0.4	0.03513	0.03372	0.09091	0.4
0.02501	0.02274	0.06667	0.4	0.09091	0.08264	0.16667	0.4	0.03887	0.03533	0.09524	0.4
0.21032	0.11209	1	0.84142	0.43889	0.0937	0.5	0.62426	0.22881	0.05872	0.36364	0.67589
0.18092	0.09505	0.5	0.84142	0.28879	0.09067	0.5	0.27071	0.18962	0.04203	0.33333	0.67589
0.05919	0.03592	0.16667	0.64142	0.13492	0.0514	0.16667	0.64142	0.07757	0.03848	0.16667	0.64142
0.18965	0.08363	0.5	0.84142	0.27891	0.07839	0.42857	0.27071	0.19661	0.04394	0.30769	0.67589
0.16255	0.07025	0.5	0.85689	0.36607	0.07813	0.42857	0.56042	0.19667	0.04653	0.33333	0.72426
0.1293	0.05795	1	1	0.35859	0.13981	0.55556	0.34142	0.1507	0.01525	0.2	1
0.06917	0.02729	0.16667	0.62426	0.22778	0.10556	0.33333	0.37071	0.10025	0.0335	0.16667	0.62426
0.07005	0.01469	0.125	0.76569	0.30313	0.15844	0.625	0.2	0.10735	0.02538	0.14925	0.2
0.12059	0.04323	0.2	0.62426	0.29121	0.14793	0.42857	0.48284	0.16111	0.05574	0.27273	0.48284
0.03524	0.03132	0.11111	0.39712	0.11111	0.09877	0.33333	0.2	0.05209	0.0463	0.13333	0.39712
0.05863	0.02849	0.16667	0.72426	0.21429	0.09127	0.28571	0.56042	0.08632	0.03441	0.16	0.56042
0.11403	0.04312	0.33333	0.88284	0.35135	0.10435	0.42857	0.5	0.156	0.04107	0.23529	0.72426
0.16224	0.07761	1	1	0.42442	0.13778	0.625	0.32845	0.18827	0.03234	0.3	0.64142

Table B.2. Right full table with cosine distance used in merge phase with Levenshtein distance used in linking phase.

Id diagnosis	Count Avg	Count Dev	Count Max	Count At	Score Avg	Score Dev	Score Max	Score At
exa:cervix/19/1000*	1.8857	0.55673	3	15.4	0.12161	0.01249	0.16667	3
exa:cervix/19/1002	3.4565	1.2391	5	12.8	0.15633	0.02726	0.2	0
exa:cervix/19/1004	0.91429	0.15673	1	3	0.09524	0.01633	0.11111	3
exa:cervix/19/1006	2.3548	0.58273	3	10.4	0.16382	0.03185	0.2	5.6
exa:cervix/19/1008*	2.5	1	4	12.8	0.19173	0.01694	0.21667	10.4
exa:cervix/19/1010_1A	1.625	0.53125	2	6.8	0.18383	0.05134	0.225	6.8
exa:cervix/19/1010_2B	1.25	0.91667	2	7	0.10692	0.0836	0.25	6.8
exa:cervix/19/1011*	3.1282	0.98356	4	8.8	0.18238	0.03217	0.23333	7
exa:cervix/19/1012	2.1333	0.80889	3	10.8	0.17594	0.03607	0.225	6.8
exa:cervix/19/1013	2.7333	0.98667	4	14.2	0.14543	0.02588	0.2	1
exa:cervix/19/1014	2.85	1.01	4	10.6	0.16015	0.03895	0.21667	6.8
exa:cervix/19/1015	1.5333	1.0311	3	14	0.1135	0.06054	0.2	5
exa:cervix/19/1016	1.5263	1.0194	3	14	0.06744	0.04259	0.10556	7
exa:cervix/19/1017	3.2647	1.308	5	14	0.16511	0.01756	0.2	2
exa:cervix/19/1018	2.4194	1.0448	4	13.8	0.14671	0.03285	0.2	5
exa:cervix/19/1020	2.4242	1.3388	4	14	0.14345	0.05217	0.2	5
exa:cervix/19/1022*	2.96	0.624	4	14	0.14187	0.03455	0.225	5.4
exa:cervix/19/1023	0.78947	0.33241	1	5	0.15789	0.06648	0.2	5
exa:cervix/19/1024	2.4762	1.2154	4	13.4	0.12635	0.03574	0.2	0.8
exa:cervix/19/1025*	2.3939	0.73462	3	7	0.19998	0.01669	0.23333	7
exa:cervix/19/1789	0.23529	0.38754	2	25	0.00945	0.01557	0.06548	25
exa:cervix/19/1790	1.8077	0.32544	2	7	0.14574	0.04553	0.18056	7
exa:cervix/19/1792	0.56667	0.79333	3	17	0.03456	0.04838	0.14286	10.6
exa:cervix/19/1793	1.7222	0.83333	3	16.8	0.12208	0.04747	0.25	4.4
exa:cervix/19/1794	1.6897	0.85612	3	11	0.18316	0.02668	0.225	10
exa:cervix/19/1795	2.4483	0.49465	3	10.2	0.1292	0.02343	0.2	2.8
exa:cervix/19/1797*	1.9302	1.0946	4	18.4	0.16864	0.05168	0.25	6.2
exa:cervix/19/1798	3.0323	0.87409	4	10	0.13355	0.02515	0.2	3.6
exa:cervix/19/1799	1.7667	0.71556	3	13.4	0.10941	0.02011	0.16667	2.6
exa:cervix/19/1800	2.84	1.1808	5	17	0.16259	0.03092	0.20556	10.2
exa:cervix/19/1802	2.8	1.2	5	17	0.14897	0.03535	0.2	2.8
exa:cervix/19/1803	2.6364	1.0248	4	10.6	0.16656	0.02712	0.20556	10.2
exa:cervix/19/1807	0.60526	0.76454	2	18.2	0.02515	0.03177	0.07692	13.6
exa:cervix/19/1808	0.53125	0.73047	2	17.4	0.02109	0.029	0.07692	14
exa:cervix/19/1809	1.421	0.96953	3	13.8	0.07177	0.03777	0.11111	6
exa:cervix/19/1811	0.61765	0.47232	1	7.8	0.05348	0.04089	0.09091	7.8
exa:cervix/19/1812*	0.55172	0.49465	1	7.8	0.04781	0.04286	0.09091	7.8
exa:cervix/19/279	1.9429	0.9698	3	10.2	0.19094	0.03737	0.25	6
exa:cervix/19/280	2	1.1539	4	17.8	0.16661	0.04209	0.225	10
exa:cervix/19/281	0.42308	0.48817	1	8.8	0.03254	0.03755	0.07692	8.8
exa:cervix/19/282*	1.6364	0.63636	3	17.8	0.18488	0.03336	0.225	10
exa:cervix/19/966	2.2059	0.56055	3	10.4	0.14816	0.0264	0.2	4.4
exa:cervix/19/969	3.125	1.3125	5	12.8	0.16226	0.02914	0.2	0
exa:cervix/19/975	0.75	0.91667	2	14	0.03236	0.03955	0.1	12.2
exa:cervix/19/985	1.7736	1.5842	5	23	0.07976	0.05418	0.18056	10.4
exa:cervix/19/990	1.25	0.89063	3	15.2	0.07074	0.03979	0.14286	8.8
exa:cervix/19/993	1.0769	0.49704	2	13.6	0.08064	0.0359	0.11111	4.8
exa:cervix/19/995	1.1702	0.63558	2	10.4	0.15413	0.07026	0.225	10.4
exa:cervix/19/996	1.3721	0.85668	3	12.6	0.08023	0.03143	0.14286	4.4
exa:cervix/19/997	3.2857	1.2536	5	13.4	0.16079	0.02737	0.2	0

Table B.3. Left full table with Hamming distance used in merge phase with Levenshtein distance used in linking phase.

Prec Avg	Prec Dev	Prec Max	Prec At	Rec Avg	Rec Dev	Rec Max	Rec At	F_1 Avg	F_1 Dev	F_1 Max	F_1 At
0.11533	0.08664	1	3	0.26939	0.07953	0.42857	15.4	0.11584	0.02544	0.25	3
0.14418	0.08015	1	0	0.38406	0.13768	0.55556	12.8	0.15834	0.02611	0.27273	5
0.0418	0.02074	0.14286	3	0.15238	0.02612	0.16667	3	0.06246	0.02459	0.15385	3
0.1257	0.06351	0.5	4.6	0.33641	0.08325	0.42857	10.4	0.1588	0.03905	0.33333	5
0.12638	0.07674	1	0	0.3125	0.125	0.5	12.8	0.13343	0.0217	0.22222	0
0.10801	0.04153	0.33333	5	0.23214	0.07589	0.28571	6.8	0.13625	0.03559	0.21053	6.8
0.03428	0.02476	0.08333	7	0.15625	0.11458	0.25	7	0.05574	0.04026	0.125	7
0.16408	0.06454	1	1	0.34758	0.10928	0.44444	8.8	0.18301	0.02701	0.25	8.8
0.10612	0.03126	0.33333	4.4	0.30476	0.11556	0.42857	10.8	0.14452	0.0302	0.2	4.4
0.17589	0.09026	1	1	0.3037	0.10963	0.44444	14.2	0.1745	0.01824	0.21053	4.6
0.18797	0.0492	0.33333	5	0.40714	0.14429	0.57143	10.6	0.24339	0.05392	0.33333	6.8
0.06119	0.03374	0.10714	14	0.19167	0.12889	0.375	14	0.09127	0.05206	0.16667	14
0.07966	0.0506	0.14286	7	0.19079	0.12742	0.375	14	0.11125	0.07195	0.19355	14
0.18584	0.07671	1	2	0.32647	0.1308	0.5	14	0.19006	0.0229	0.23529	7.8
0.1047	0.03082	0.25	5	0.26882	0.11608	0.44444	13.8	0.1392	0.0367	0.2069	6.8
0.08098	0.0349	0.13636	7	0.30303	0.16736	0.5	14	0.12588	0.05858	0.2	7
0.17622	0.04431	0.33333	5.4	0.42286	0.08914	0.57143	14	0.23593	0.03645	0.31579	6.4
0.07444	0.04224	0.2	5	0.13158	0.0554	0.16667	5	0.09084	0.04205	0.18182	5
0.15757	0.03632	0.33333	0.8	0.30952	0.15193	0.5	13.4	0.18936	0.0472	0.26667	13.4
0.18012	0.08003	1	0.8	0.29924	0.09183	0.375	7	0.18408	0.03165	0.26087	7
0.01548	0.02549	0.125	25	0.04706	0.07751	0.4	25	0.02328	0.03834	0.19048	25
0.11096	0.04523	0.25	6.2	0.25824	0.04649	0.28571	7	0.14408	0.03615	0.25	7
0.0153	0.02142	0.07692	17	0.09444	0.13222	0.5	17	0.02631	0.03683	0.13333	17
0.0902	0.03532	0.33333	3.8	0.24603	0.11905	0.42857	16.8	0.11604	0.02965	0.2	3.8
0.14366	0.09644	1	2	0.24138	0.1223	0.42857	11	0.13532	0.03578	0.25	2
0.18488	0.12848	1	2.8	0.30603	0.06183	0.375	10.2	0.18582	0.05323	0.4	2.8
0.05896	0.02	0.16667	3.4	0.24128	0.13683	0.5	18.4	0.08901	0.03053	0.14286	3.4
0.19017	0.0848	0.66667	4.8	0.37903	0.10926	0.5	10	0.2158	0.03829	0.36364	4.8
0.13982	0.0945	1	2.6	0.25238	0.10222	0.42857	13.4	0.13573	0.02899	0.25	2.6
0.21898	0.08467	0.66667	4.4	0.40571	0.16869	0.71429	17	0.25	0.05046	0.4	4.4
0.21256	0.09861	0.66667	3.6	0.4	0.17143	0.71429	17	0.23753	0.04987	0.4	3.6
0.24676	0.10198	0.66667	4.6	0.37662	0.1464	0.57143	10.6	0.26254	0.05313	0.4	4.6
0.01509	0.01906	0.05263	18.2	0.07566	0.09557	0.25	18.2	0.02512	0.03174	0.08696	18.2
0.01346	0.01851	0.05405	17.4	0.07589	0.10435	0.28571	17.4	0.02285	0.03142	0.09091	17.4
0.09224	0.04963	0.15789	13.8	0.23684	0.16159	0.5	13.8	0.13033	0.07251	0.24	13.8
0.02405	0.01839	0.05556	7.8	0.12353	0.09446	0.2	7.8	0.03997	0.03056	0.08696	7.8
0.02303	0.02065	0.05882	7.8	0.09195	0.08244	0.16667	7.8	0.03656	0.03277	0.08696	7.8
0.13055	0.09594	1	3.2	0.32381	0.16163	0.5	10.2	0.13434	0.03809	0.28571	3.2
0.11986	0.05361	0.5	3.4	0.25	0.14423	0.5	17.8	0.13483	0.04137	0.2	3.4
0.01593	0.01838	0.04545	8.8	0.07051	0.08136	0.16667	8.8	0.02594	0.02993	0.07143	8.8
0.15592	0.1064	1	3.2	0.23377	0.09091	0.42857	17.8	0.14049	0.02908	0.25	3.2
0.14323	0.10333	1	3	0.31513	0.08008	0.42857	10.4	0.14444	0.03877	0.28571	5
0.11833	0.07111	1	0	0.34722	0.14583	0.55556	12.8	0.12777	0.01619	0.2	0
0.02095	0.0256	0.06667	14	0.125	0.15278	0.33333	14	0.03581	0.04377	0.11111	14
0.0301	0.02265	0.06667	13.4	0.2217	0.19802	0.625	23	0.05267	0.04064	0.12048	23
0.04601	0.02766	0.09091	15.2	0.17857	0.12723	0.42857	15.2	0.07198	0.04352	0.15	15.2
0.06293	0.02794	0.14286	4.8	0.17949	0.08284	0.33333	13.6	0.08936	0.03789	0.15385	4.8
0.0325	0.01538	0.09091	4.6	0.16717	0.0908	0.28571	10.4	0.05218	0.02409	0.11111	4.6
0.03535	0.01692	0.08333	4.4	0.19601	0.12238	0.42857	12.6	0.05817	0.02761	0.10526	4.4
0.1309	0.06513	1	0	0.41071	0.15671	0.625	13.4	0.15216	0.01865	0.22222	0

Table B.4. Right full table with Hamming distance used in merge phase with Levenshtein distance used in linking phase.

Id diagnosis	Count Avg	Count Dev	Count Max	Count At	Score Avg	Score Dev	Score Max	Score At
exa:cervix/19/1000*	1.8571	0.54694	3	14.2	0.11896	0.01598	0.16667	2.4
exa:cervix/19/1002	3.4783	1.1966	5	11.4	0.1588	0.03619	0.2	0
exa:cervix/19/1004	0.90323	0.17482	1	2.6	0.09462	0.01831	0.11111	2.6
exa:cervix/19/1006	2.3636	0.57851	3	8.4	0.15162	0.03179	0.2	3.8
exa:cervix/19/1008*	2.3902	0.8959	4	11.8	0.18743	0.02327	0.21667	8.2
exa:cervix/19/1010_1A	1.7619	0.40816	2	5.2	0.18271	0.05803	0.225	5.2
exa:cervix/19/1010_2B	1.3226	0.87409	2	5.4	0.11431	0.08211	0.25	5.2
exa:cervix/19/1011*	3.1389	0.90895	4	8.4	0.17924	0.0335	0.23333	5.6
exa:cervix/19/1012	2.2333	0.81778	3	8	0.16698	0.03993	0.225	5.2
exa:cervix/19/1013	2.7097	0.83455	4	11.6	0.14374	0.01928	0.2	0.8
exa:cervix/19/1014	2.85	1.025	4	8.4	0.15015	0.04943	0.21667	5.2
exa:cervix/19/1015	1.5862	1.0654	3	11.4	0.10018	0.06218	0.2	5
exa:cervix/19/1016	1.8182	0.71901	3	11	0.08776	0.02393	0.11111	4
exa:cervix/19/1017	3.4444	1.2531	5	11.4	0.16094	0.01544	0.2	1.8
exa:cervix/19/1018	2.48	1.0688	4	11	0.12882	0.04264	0.225	5.2
exa:cervix/19/1020	2.3714	1.4188	4	11.4	0.12616	0.07209	0.225	5
exa:cervix/19/1022*	2.76	1.0048	4	11	0.12808	0.03786	0.225	5.4
exa:cervix/19/1023	0.78947	0.33241	1	5	0.15789	0.06648	0.2	5
exa:cervix/19/1024	2.3478	1.3195	4	10.6	0.13718	0.04352	0.2	0.8
exa:cervix/19/1025*	2.5667	0.66444	3	5.2	0.20146	0.01902	0.23333	5.6
exa:cervix/19/1789	0.6875	0.85938	2	12.6	0.02567	0.03209	0.08333	11.8
exa:cervix/19/1790	1.8333	0.29167	2	5.8	0.16591	0.03598	0.25	5.2
exa:cervix/19/1792	0.625	0.78125	3	14.6	0.04563	0.05704	0.14286	8.4
exa:cervix/19/1793	1.7	0.78667	3	15.6	0.11759	0.04555	0.25	4
exa:cervix/19/1794	1.8667	0.98222	3	7.4	0.16873	0.03613	0.2	1.6
exa:cervix/19/1795	2.4516	0.56608	3	7.4	0.13159	0.03614	0.2	1.8
exa:cervix/19/1797*	2.1316	1.1177	4	16.2	0.17323	0.05383	0.25	3.4
exa:cervix/19/1798	2.931	1.0511	4	7.4	0.14001	0.02778	0.2	1.8
exa:cervix/19/1799	1.7407	0.78738	3	11.6	0.10842	0.02708	0.16667	2.2
exa:cervix/19/1800	2.625	1.2292	5	15.2	0.16527	0.03378	0.20556	7.4
exa:cervix/19/1802	2.84	1.0464	5	15.2	0.15007	0.03709	0.2	2.6
exa:cervix/19/1803	2.5	1.15	4	8.2	0.16665	0.03318	0.20556	7.4
exa:cervix/19/1807	0.54545	0.72727	2	13.6	0.02279	0.03038	0.07692	11.6
exa:cervix/19/1808	0.54545	0.72727	2	15.6	0.02279	0.03038	0.07692	11.6
exa:cervix/19/1809	1.35	0.95	3	11.8	0.06801	0.04081	0.11111	5.6
exa:cervix/19/1811	0.58621	0.48514	1	6.8	0.05068	0.04194	0.09091	6.8
exa:cervix/19/1812*	0.5	0.5	1	7	0.04356	0.04356	0.09091	7
exa:cervix/19/279	1.8529	0.85294	3	9	0.20004	0.0529	0.25	3.4
exa:cervix/19/280	2.1786	1.0485	4	15	0.15481	0.04732	0.225	7.4
exa:cervix/19/281	0.41667	0.48611	1	7.8	0.03205	0.03739	0.07692	7.8
exa:cervix/19/282*	1.76	0.4864	3	15	0.18349	0.03714	0.225	6.2
exa:cervix/19/966	2.2895	0.56094	3	8.2	0.1319	0.02874	0.2	4
exa:cervix/19/969	3.24	1.2496	5	11.4	0.15378	0.03463	0.2	0
exa:cervix/19/975	0.89189	0.77137	2	12	0.05137	0.04443	0.1	7.2
exa:cervix/19/985	1.8636	1.6364	5	17.6	0.06811	0.05263	0.18056	8.2
exa:cervix/19/990	1.2414	0.7824	3	15.2	0.07804	0.03814	0.14286	7.2
exa:cervix/19/993	1.0741	0.48011	2	12	0.08285	0.0347	0.11111	4
exa:cervix/19/995	1.275	0.6525	2	8.2	0.15029	0.0597	0.225	8.2
exa:cervix/19/996	2.125	0.83125	3	7.8	0.08986	0.02201	0.14286	2.4
exa:cervix/19/997	3.3019	1.251	5	11.4	0.15966	0.03196	0.2	0

Table B.5. Left full table with Levenshtein distance used in merge phase with Levenshtein distance used in linking phase.

Prec Avg	Prec Dev	Prec Max	Prec At	Rec Avg	Rec Dev	Rec Max	Rec At	F_1 Avg	F_1 Dev	F_1 Max	F_1 At
0.09233	0.04984	0.5	2.4	0.26531	0.07813	0.42857	14.2	0.11146	0.02418	0.22222	2.4
0.13854	0.07439	1	0	0.38647	0.13296	0.55556	11.4	0.15427	0.01719	0.22222	3.6
0.04284	0.02365	0.14286	2.6	0.15054	0.02914	0.16667	2.6	0.06274	0.02695	0.15385	2.6
0.12021	0.05797	0.4	4.2	0.33766	0.08264	0.42857	8.4	0.15737	0.03989	0.33333	4.2
0.12415	0.07624	1	0	0.29878	0.11199	0.5	11.8	0.13014	0.02056	0.22222	0
0.11658	0.05125	0.28571	5.2	0.2517	0.05831	0.28571	5.2	0.15112	0.04911	0.28571	5.2
0.03751	0.0242	0.09091	5.4	0.16532	0.10926	0.25	5.4	0.06044	0.039	0.13333	5.4
0.17216	0.08306	1	0.8	0.34877	0.10099	0.44444	8.4	0.18245	0.02555	0.27273	5.2
0.12002	0.04158	0.33333	4	0.31905	0.11683	0.42857	8	0.15694	0.0314	0.21429	8
0.19932	0.12217	1	0.8	0.30108	0.09273	0.44444	11.6	0.18683	0.02456	0.30769	2.8
0.19437	0.07256	0.4	4.8	0.40714	0.14643	0.57143	8.4	0.25105	0.07628	0.375	5.2
0.05911	0.03693	0.10345	11.4	0.19828	0.13317	0.375	11.4	0.0903	0.0572	0.16216	11.4
0.11384	0.03527	0.2	4	0.22727	0.08988	0.375	11	0.14557	0.04429	0.21053	5.6
0.20745	0.10204	1	1.8	0.34444	0.12531	0.5	11.4	0.20494	0.02057	0.28571	2.8
0.09977	0.03708	0.2	4	0.27556	0.11876	0.44444	11	0.13773	0.04578	0.24	5.6
0.07611	0.04349	0.13636	5.8	0.29643	0.17735	0.5	11.4	0.11977	0.06844	0.2	5.8
0.14913	0.04359	0.25	5.6	0.39429	0.14354	0.57143	11	0.20868	0.05962	0.31579	5.6
0.07087	0.03689	0.16667	5	0.13158	0.0554	0.16667	5	0.08946	0.04063	0.16667	5
0.15007	0.04131	0.33333	0.8	0.29348	0.16493	0.5	10.6	0.18053	0.05681	0.27586	10.6
0.1848	0.0895	1	0.8	0.32083	0.08306	0.375	5.2	0.19061	0.03796	0.3	5.2
0.04951	0.06189	0.15385	12.6	0.1375	0.17188	0.4	12.6	0.07269	0.09087	0.22222	12.6
0.13748	0.08309	0.5	5.2	0.2619	0.04167	0.28571	5.8	0.15468	0.04991	0.30769	5.8
0.01729	0.02161	0.07692	14.6	0.10417	0.13021	0.5	14.6	0.0296	0.037	0.13333	14.6
0.08958	0.03141	0.25	3	0.24286	0.11238	0.42857	15.6	0.11743	0.02747	0.18182	3
0.1559	0.10026	1	1.6	0.26667	0.14032	0.42857	7.4	0.14905	0.03711	0.25	1.6
0.20863	0.14624	1	1.8	0.30645	0.07076	0.375	7.4	0.19198	0.04587	0.36364	2.4
0.07592	0.03216	0.33333	1.6	0.26645	0.13972	0.5	16.2	0.10093	0.02739	0.18182	1.6
0.2038	0.09958	1	1.8	0.36638	0.13139	0.5	7.4	0.20498	0.02733	0.28571	4.2
0.11452	0.05783	0.5	2.2	0.24868	0.11248	0.42857	11.6	0.13007	0.03079	0.22222	2.2
0.21109	0.05808	0.5	2.2	0.375	0.1756	0.71429	15.2	0.23873	0.04651	0.30769	8.2
0.21461	0.1001	0.66667	3.4	0.40571	0.14949	0.71429	15.2	0.24091	0.04545	0.4	3.4
0.22428	0.06177	0.5	2.2	0.35714	0.16429	0.57143	8.2	0.24466	0.05185	0.33333	8.2
0.01319	0.01759	0.05	13.6	0.06818	0.09091	0.25	13.6	0.02209	0.02945	0.08333	13.6
0.01379	0.01839	0.05263	15.6	0.07792	0.1039	0.28571	15.6	0.02342	0.03122	0.08889	15.6
0.0888	0.05383	0.15385	6.2	0.225	0.15833	0.5	11.8	0.1247	0.07618	0.23077	11.8
0.02193	0.01815	0.05556	6.8	0.11724	0.09703	0.2	6.8	0.03672	0.03039	0.08696	6.8
0.01996	0.01996	0.05556	7	0.08333	0.08333	0.16667	7	0.03203	0.03203	0.08333	7
0.11839	0.07887	1	1.6	0.30882	0.14216	0.5	9	0.12761	0.02987	0.28571	1.6
0.12601	0.05567	0.5	1.8	0.27232	0.13106	0.5	15	0.14308	0.03086	0.2	1.8
0.01559	0.01819	0.04545	7.8	0.06944	0.08102	0.16667	7.8	0.02541	0.02964	0.07143	7.8
0.17338	0.11332	1	1.8	0.25143	0.06949	0.42857	15	0.1569	0.02877	0.25	1.8
0.12373	0.07993	0.5	2.6	0.32707	0.08013	0.42857	8.2	0.14775	0.04864	0.36364	3.6
0.11366	0.06839	1	0	0.36	0.13884	0.55556	11.4	0.12561	0.01447	0.2	0
0.02677	0.02315	0.0625	12	0.14865	0.12856	0.33333	12	0.04514	0.03904	0.10526	12
0.03054	0.02473	0.06667	17.6	0.23295	0.20455	0.625	17.6	0.05377	0.04415	0.12048	17.6
0.04881	0.02485	0.1	4	0.17734	0.11177	0.42857	15.2	0.07453	0.03798	0.13953	15.2
0.06327	0.02856	0.14286	4	0.17901	0.08002	0.33333	12	0.0895	0.0381	0.15385	4
0.03543	0.01494	0.11111	3.6	0.18214	0.09321	0.28571	8.2	0.0565	0.02219	0.125	3.6
0.06589	0.01903	0.16667	2.4	0.30357	0.11875	0.42857	7.8	0.10088	0.02074	0.15385	2.4
0.12904	0.06354	1	0	0.41274	0.15637	0.625	11.4	0.15211	0.01729	0.22222	0

Table B.6. Right full table with Levenshtein distance used in merge phase with Levenshtein distance used in linking phase.

Id diagnosis	Count Avg	Count Dev	Count Max	Count At	Score Avg	Score Dev	Score Max	Score At
exa:cervix/19/1000*	1.8571	0.54694	3	14.2	0.11896	0.01598	0.16667	2.4
exa:cervix/19/1002	3.4783	1.1966	5	11.4	0.1588	0.03619	0.2	0
exa:cervix/19/1004	0.90323	0.17482	1	2.6	0.09462	0.01831	0.11111	2.6
exa:cervix/19/1006	2.3824	0.58131	3	8.4	0.15077	0.03165	0.2	3.8
exa:cervix/19/1008*	2.3902	0.8959	4	11.8	0.18602	0.02384	0.21667	8.2
exa:cervix/19/1010_1A	1.7619	0.40816	2	5.2	0.18271	0.05803	0.225	5.2
exa:cervix/19/1010_2B	1.3226	0.87409	2	5.4	0.11431	0.08211	0.25	5.2
exa:cervix/19/1011*	3.1176	0.93426	4	8.4	0.17821	0.03326	0.23333	5.6
exa:cervix/19/1012	2.2333	0.81778	3	8	0.16698	0.03993	0.225	5.2
exa:cervix/19/1013	2.7097	0.83455	4	11.6	0.14374	0.01928	0.2	0.8
exa:cervix/19/1014	2.85	1.025	4	8.4	0.14866	0.04883	0.21667	5.4
exa:cervix/19/1015	1.5862	1.0654	3	11.4	0.10018	0.06218	0.2	5
exa:cervix/19/1016	1.8182	0.71901	3	11	0.08776	0.02393	0.11111	4
exa:cervix/19/1017	3.4444	1.2531	5	11.4	0.16094	0.01544	0.2	1.8
exa:cervix/19/1018	2.48	1.0688	4	11	0.12882	0.04264	0.225	5.2
exa:cervix/19/1020	2.3714	1.4188	4	11.4	0.12616	0.07209	0.225	5
exa:cervix/19/1022*	2.76	1.0048	4	11	0.12808	0.03786	0.225	5.4
exa:cervix/19/1023	0.78947	0.33241	1	5	0.15789	0.06648	0.2	5
exa:cervix/19/1024	2.3478	1.3195	4	10.6	0.13718	0.04352	0.2	0.8
exa:cervix/19/1025*	2.5667	0.66444	3	5.2	0.20146	0.01902	0.23333	5.6
exa:cervix/19/1789	0.6875	0.85938	2	12.6	0.02567	0.03209	0.08333	11.8
exa:cervix/19/1790	1.8333	0.29167	2	5.8	0.16591	0.03598	0.25	5.2
exa:cervix/19/1792	0.625	0.78125	3	14.6	0.04563	0.05704	0.14286	8.4
exa:cervix/19/1793	1.7	0.78667	3	15.6	0.11759	0.04555	0.25	4
exa:cervix/19/1794	1.8667	0.98222	3	7.4	0.16873	0.03613	0.2	1.6
exa:cervix/19/1795	2.4516	0.56608	3	7.4	0.13159	0.03614	0.2	1.8
exa:cervix/19/1797*	2.1538	1.1085	4	16.2	0.17236	0.05325	0.25	3.4
exa:cervix/19/1798	2.9	1.0533	4	7.4	0.14053	0.02751	0.2	1.8
exa:cervix/19/1799	1.7143	0.78571	3	11.6	0.10852	0.02622	0.16667	2.2
exa:cervix/19/1800	2.68	1.2416	5	15.2	0.164	0.03415	0.20556	7.4
exa:cervix/19/1802	2.92	1.0496	5	15.2	0.14708	0.03826	0.2	2.6
exa:cervix/19/1803	2.5	1.15	4	8.2	0.16665	0.03318	0.20556	7.4
exa:cervix/19/1807	0.52941	0.71626	2	13.6	0.02212	0.02992	0.07692	11.6
exa:cervix/19/1808	0.54545	0.72727	2	15.6	0.02279	0.03038	0.07692	11.6
exa:cervix/19/1809	1.35	0.95	3	11.8	0.06801	0.04081	0.11111	5.6
exa:cervix/19/1811	0.58621	0.48514	1	6.8	0.05068	0.04194	0.09091	6.8
exa:cervix/19/1812*	0.5	0.5	1	7	0.04356	0.04356	0.09091	7
exa:cervix/19/279	1.8529	0.85294	3	9	0.20004	0.0529	0.25	3.4
exa:cervix/19/280	2.1071	1.051	4	15	0.15758	0.04778	0.225	7.4
exa:cervix/19/281	0.41667	0.48611	1	7.8	0.03205	0.03739	0.07692	7.8
exa:cervix/19/282*	1.75	0.5	3	15	0.18536	0.03618	0.225	6.2
exa:cervix/19/966	2.282	0.55227	3	8.2	0.13136	0.02844	0.2	4
exa:cervix/19/969	3.2449	1.2703	5	11.4	0.15367	0.03524	0.2	0
exa:cervix/19/975	0.89189	0.77137	2	12	0.05137	0.04443	0.1	7.2
exa:cervix/19/985	1.907	1.6301	5	17.6	0.06969	0.05187	0.18056	8.2
exa:cervix/19/990	1.2667	0.78667	3	15.2	0.07933	0.03799	0.14286	7.2
exa:cervix/19/993	1.0769	0.49704	2	12	0.08177	0.03537	0.11111	4
exa:cervix/19/995	1.2683	0.64247	2	8.2	0.14939	0.05794	0.225	8.2
exa:cervix/19/996	2.125	0.83125	3	7.8	0.08986	0.02201	0.14286	2.4
exa:cervix/19/997	3.3019	1.251	5	11.4	0.15966	0.03196	0.2	0

Table B.7. Left full table with Levenshtein Damerau distance used in merge phase with Levenshtein distance used in linking phase.

Prec Avg	Prec Dev	Prec Max	Prec At	Rec Avg	Rec Dev	Rec Max	Rec At	F_1 Avg	F_1 Dev	F_1 Max	F_1 At
0.09233	0.04984	0.5	2.4	0.26531	0.07813	0.42857	14.2	0.11146	0.02418	0.22222	2.4
0.13815	0.07376	1	0	0.38647	0.13296	0.55556	11.4	0.15404	0.01689	0.22222	3.6
0.04284	0.02365	0.14286	2.6	0.15054	0.02914	0.16667	2.6	0.06274	0.02695	0.15385	2.6
0.11944	0.05667	0.4	4.2	0.34034	0.08304	0.42857	8.4	0.15726	0.03879	0.33333	4.2
0.12412	0.07626	1	0	0.29878	0.11199	0.5	11.8	0.13009	0.0206	0.22222	0
0.11658	0.05125	0.28571	5.2	0.2517	0.05831	0.28571	5.2	0.15112	0.04911	0.28571	5.2
0.03751	0.0242	0.09091	5.4	0.16532	0.10926	0.25	5.4	0.06044	0.039	0.13333	5.4
0.17459	0.08583	1	0.8	0.34641	0.10381	0.44444	8.4	0.18213	0.02521	0.26087	5.2
0.12002	0.04158	0.33333	4	0.31905	0.11683	0.42857	8	0.15694	0.0314	0.21429	8
0.19932	0.12217	1	0.8	0.30108	0.09273	0.44444	11.6	0.18683	0.02456	0.30769	2.8
0.1927	0.07106	0.4	4.8	0.40714	0.14643	0.57143	8.4	0.24994	0.07573	0.35294	5.2
0.05911	0.03693	0.10345	11.4	0.19828	0.13317	0.375	11.4	0.0903	0.0572	0.16216	11.4
0.11384	0.03527	0.2	4	0.22727	0.08988	0.375	11	0.14557	0.04429	0.21053	5.6
0.20745	0.10204	1	1.8	0.34444	0.12531	0.5	11.4	0.20494	0.02057	0.28571	2.8
0.09977	0.03708	0.2	4	0.27556	0.11876	0.44444	11	0.13773	0.04578	0.24	5.6
0.07611	0.04349	0.13636	5.8	0.29643	0.17735	0.5	11.4	0.11977	0.06844	0.2	5.8
0.14913	0.04359	0.25	5.6	0.39429	0.14354	0.57143	11	0.20868	0.05962	0.31579	5.6
0.07087	0.03689	0.16667	5	0.13158	0.0554	0.16667	5	0.08946	0.04063	0.16667	5
0.15007	0.04131	0.33333	0.8	0.29348	0.16493	0.5	10.6	0.18053	0.05681	0.27586	10.6
0.1848	0.0895	1	0.8	0.32083	0.08306	0.375	5.2	0.19061	0.03796	0.3	5.2
0.04951	0.06189	0.15385	12.6	0.1375	0.17188	0.4	12.6	0.07269	0.09087	0.22222	12.6
0.13656	0.08177	0.5	5.2	0.2619	0.04167	0.28571	5.8	0.15407	0.04914	0.30769	5.8
0.01724	0.02156	0.07692	14.6	0.10417	0.13021	0.5	14.6	0.02954	0.03693	0.13333	14.6
0.08958	0.03141	0.25	3	0.24286	0.11238	0.42857	15.6	0.11743	0.02747	0.18182	3
0.1559	0.10026	1	1.6	0.26667	0.14032	0.42857	7.4	0.14905	0.03711	0.25	1.6
0.20863	0.14624	1	1.8	0.30645	0.07076	0.375	7.4	0.19198	0.04587	0.36364	2.4
0.07537	0.03161	0.33333	1.6	0.26923	0.13856	0.5	16.2	0.10079	0.02685	0.18182	1.6
0.20653	0.10045	1	1.8	0.3625	0.13167	0.5	7.4	0.20704	0.02863	0.28571	4.2
0.11253	0.05662	0.5	2.2	0.2449	0.11224	0.42857	11.6	0.1284	0.03141	0.22222	2.2
0.20931	0.05685	0.5	2.2	0.38286	0.17737	0.71429	15.2	0.2395	0.04527	0.30769	8.2
0.21387	0.10045	0.66667	3.4	0.41714	0.14994	0.71429	15.2	0.2419	0.04442	0.4	3.4
0.22428	0.06177	0.5	2.2	0.35714	0.16429	0.57143	8.2	0.24466	0.05185	0.33333	8.2
0.01281	0.01733	0.05	13.6	0.06618	0.08953	0.25	13.6	0.02144	0.02901	0.08333	13.6
0.01379	0.01839	0.05263	15.6	0.07792	0.1039	0.28571	15.6	0.02342	0.03122	0.08889	15.6
0.0888	0.05383	0.15385	6.2	0.225	0.15833	0.5	11.8	0.1247	0.07618	0.23077	11.8
0.02193	0.01815	0.05556	6.8	0.11724	0.09703	0.2	6.8	0.03672	0.03039	0.08696	6.8
0.01996	0.01996	0.05556	7	0.08333	0.08333	0.16667	7	0.03203	0.03203	0.08333	7
0.11839	0.07887	1	1.6	0.30882	0.14216	0.5	9	0.12761	0.02987	0.28571	1.6
0.1285	0.05768	0.5	1.8	0.26339	0.13138	0.5	15	0.14269	0.03057	0.2	1.8
0.01559	0.01819	0.04545	7.8	0.06944	0.08102	0.16667	7.8	0.02541	0.02964	0.07143	7.8
0.17762	0.11627	1	1.8	0.25	0.07143	0.42857	15	0.15867	0.02826	0.25	1.8
0.12357	0.07796	0.5	2.6	0.32601	0.0789	0.42857	8.2	0.14823	0.04806	0.36364	3.6
0.11437	0.06956	1	0	0.36054	0.14114	0.55556	11.4	0.12557	0.01471	0.2	0
0.02677	0.02315	0.0625	12	0.14865	0.12856	0.33333	12	0.04514	0.03904	0.10526	12
0.03125	0.02458	0.06667	17.6	0.23837	0.20376	0.625	17.6	0.05502	0.0439	0.12048	17.6
0.04934	0.02441	0.1	4	0.18095	0.11238	0.42857	15.2	0.07556	0.03754	0.13953	15.2
0.06233	0.02894	0.14286	4	0.17949	0.08284	0.33333	12	0.0885	0.0388	0.15385	4
0.03533	0.01464	0.11111	3.6	0.18118	0.09178	0.28571	8.2	0.05637	0.02175	0.125	3.6
0.06586	0.01905	0.16667	2.4	0.30357	0.11875	0.42857	7.8	0.10083	0.0207	0.15385	2.4
0.12904	0.06354	1	0	0.41274	0.15637	0.625	11.4	0.15211	0.01729	0.22222	0

Table B.8. Right full table with Levenshtein Damerau distance used in merge phase with Levenshtein distance used in linking phase.

Id diagnosis	Count Avg	Count Dev	Count Max	Count At	Score Avg	Score Dev	Score Max	Score At
exa:cervix/19/1000*	1.8431	0.73972	3	0.70656	0.10647	0.01788	0.16667	0.88016
exa:cervix/19/1002	3.8571	0.63673	5	0.6946	0.12816	0.02115	0.2	1
exa:cervix/19/1004	0.95455	0.08678	1	0.92952	0.10404	0.01093	0.11111	0.92952
exa:cervix/19/1006	2.6	0.56889	3	0.79897	0.13686	0.01951	0.2	0.82621
exa:cervix/19/1008*	2.9833	0.39556	4	0.6946	0.15907	0.00974	0.21667	0.82621
exa:cervix/19/1010_1A	1.8387	0.27055	2	0.82872	0.18159	0.05137	0.25	0.8975
exa:cervix/19/1010_2B	1.7059	0.41522	2	0.78405	0.16166	0.06605	0.25	0.89905
exa:cervix/19/1011*	3.0417	0.35938	4	0.69396	0.15009	0.05005	0.225	0.90536
exa:cervix/19/1012	1.9487	0.68376	3	0.69086	0.16739	0.06756	0.25	0.90127
exa:cervix/19/1013	2.4634	0.77097	4	0.6241	0.13795	0.02892	0.2	0.96364
exa:cervix/19/1014	2.0741	0.83402	4	0.69086	0.19899	0.04667	0.25	0.8975
exa:cervix/19/1015	1.4054	0.79474	3	0.64074	0.1325	0.05162	0.2	0.80039
exa:cervix/19/1016	1.5926	0.85322	3	0.64673	0.08348	0.03092	0.11111	0.75707
exa:cervix/19/1017	3.1957	1.0208	5	0.66134	0.18051	0.03371	0.25	0.8975
exa:cervix/19/1018	2.3809	0.79819	4	0.65936	0.17065	0.03482	0.25	0.90127
exa:cervix/19/1020	2.4091	0.92769	4	0.6563	0.18384	0.03708	0.25	0.8975
exa:cervix/19/1022*	1.7812	1.2598	4	0.61751	0.08737	0.0546	0.14545	0.70072
exa:cervix/19/1023	0.7619	0.36281	1	0.82544	0.14762	0.07029	0.2	0.80547
exa:cervix/19/1024	2.7586	0.97265	4	0.70292	0.11107	0.02314	0.2	0.96364
exa:cervix/19/1025*	2.6757	0.47334	3	0.82872	0.1944	0.01866	0.225	0.90127
exa:cervix/19/1789	1.1177	0.83045	2	0.7243	0.04552	0.03213	0.08333	0.74796
exa:cervix/19/1790	1.6765	0.43772	2	0.78212	0.17265	0.05609	0.25	0.8975
exa:cervix/19/1792	1	0.95238	3	0.69637	0.03975	0.03785	0.125	0.70481
exa:cervix/19/1793	2.3659	0.80428	3	0.77862	0.08827	0.01771	0.12564	0.72754
exa:cervix/19/1794	2.2	0.59429	3	0.79817	0.14507	0.02797	0.225	0.87385
exa:cervix/19/1795	2.9	0.185	3	0.87475	0.12268	0.03099	0.2	0.92393
exa:cervix/19/1797*	2.1833	0.50222	4	0.61801	0.14655	0.02692	0.225	0.87761
exa:cervix/19/1798	3.2439	0.88519	4	0.76666	0.12345	0.01772	0.225	0.88377
exa:cervix/19/1799	1.8056	0.50309	3	0.6736	0.10943	0.02931	0.16667	0.87702
exa:cervix/19/1800	2.8276	0.92747	5	0.66738	0.13248	0.02386	0.25	0.87385
exa:cervix/19/1802	2.75	0.82813	5	0.66738	0.11949	0.03031	0.25	0.87385
exa:cervix/19/1803	2.6	0.912	4	0.70873	0.14009	0.02442	0.25	0.8765
exa:cervix/19/1807	0.34043	0.55048	2	0.67561	0.01158	0.01872	0.06346	0.67561
exa:cervix/19/1808	0.42222	0.56296	2	0.66083	0.01786	0.02382	0.06346	0.66083
exa:cervix/19/1809	1.6087	0.70321	3	0.67654	0.08016	0.01511	0.1	0.84373
exa:cervix/19/1811	0.63158	0.46537	1	0.75765	0.05522	0.04069	0.09091	0.75765
exa:cervix/19/1812*	0.55882	0.49308	1	0.75765	0.04857	0.04286	0.09091	0.75765
exa:cervix/19/279	2.5	0.54167	3	0.77594	0.13661	0.01919	0.225	0.87385
exa:cervix/19/280	2.1579	0.41551	4	0.6563	0.13243	0.01637	0.225	0.87385
exa:cervix/19/281	0.59375	0.48242	1	0.779	0.04567	0.03711	0.07692	0.779
exa:cervix/19/282*	2.0625	0.17578	3	0.64551	0.13964	0.01444	0.225	0.89251
exa:cervix/19/966	2.8269	0.30621	3	0.82902	0.13025	0.02282	0.16667	0.91263
exa:cervix/19/969	3.5802	0.83524	5	0.69531	0.12475	0.01891	0.2	1
exa:cervix/19/975	0.52273	0.66529	2	0.66577	0.0337	0.04289	0.1	0.70982
exa:cervix/19/985	1.9342	0.97957	5	0.67148	0.08455	0.02999	0.25	0.82794
exa:cervix/19/990	1.7436	0.67587	3	0.67099	0.06011	0.01541	0.09307	0.66561
exa:cervix/19/993	1	0.12903	2	0.63528	0.06322	0.02183	0.11111	0.85694
exa:cervix/19/995	1.8125	0.32813	2	0.82639	0.12718	0.02235	0.20833	0.82633
exa:cervix/19/996	2.3387	0.78928	3	0.76659	0.08917	0.02151	0.2	0.86919
exa:cervix/19/997	3.5634	0.76294	5	0.67258	0.14135	0.02166	0.21667	0.81838

Table B.9. Left full table with Jaro distance used in merge phase with Levenshtein distance used in linking phase.

Prec Avg	Prec Dev	Prec Max	Prec At	Rec Avg	Rec Dev	Rec Max	Rec At	F_1 Avg	F_1 Dev	F_1 Max	F_1 At
0.08448	0.03056	0.33333	0.88016	0.26331	0.10567	0.42857	0.70656	0.11299	0.02205	0.2	0.88016
0.16546	0.09009	1	1	0.42857	0.07075	0.55556	0.6946	0.19732	0.05835	0.36364	0.83035
0.05862	0.0357	0.25	0.92952	0.15909	0.01446	0.16667	0.92952	0.07739	0.03345	0.2	0.92952
0.16039	0.07545	0.5	0.84094	0.37143	0.08127	0.42857	0.79897	0.19617	0.05276	0.33333	0.82605
0.16039	0.09558	1	1	0.37292	0.04944	0.5	0.6946	0.1829	0.05593	0.35294	0.82639
0.18667	0.11019	1	0.8975	0.26267	0.03865	0.28571	0.82872	0.17868	0.04535	0.30769	0.82872
0.11313	0.07173	1	0.89905	0.21324	0.0519	0.25	0.78405	0.11395	0.02822	0.22222	0.89905
0.21219	0.13892	1	0.94921	0.33796	0.03993	0.44444	0.69396	0.20896	0.06039	0.375	0.84796
0.12059	0.04683	0.5	0.90127	0.27839	0.09768	0.42857	0.69086	0.14493	0.02355	0.22222	0.90127
0.17548	0.08258	1	0.96364	0.27371	0.08566	0.44444	0.6241	0.1734	0.02241	0.25	0.82804
0.21353	0.09993	1	0.8975	0.2963	0.11915	0.57143	0.69086	0.20149	0.02797	0.25	0.8975
0.06811	0.02791	0.14286	0.82184	0.17568	0.09934	0.375	0.64074	0.09371	0.03818	0.14815	0.73439
0.10761	0.04208	0.2	0.75707	0.19907	0.10665	0.375	0.64673	0.13294	0.05257	0.21053	0.70681
0.1963	0.08542	1	0.8975	0.31957	0.10208	0.5	0.66134	0.19677	0.02436	0.28571	0.83051
0.14833	0.06264	0.5	0.90127	0.26455	0.08869	0.44444	0.65936	0.16242	0.02349	0.25	0.86444
0.1132	0.03245	0.22222	0.83704	0.30114	0.11596	0.5	0.6563	0.15205	0.02842	0.23529	0.83704
0.08753	0.05497	0.17647	0.69223	0.25446	0.17997	0.57143	0.61751	0.12795	0.08309	0.25	0.69223
0.07211	0.04264	0.2	0.82544	0.12698	0.06047	0.16667	0.82544	0.08814	0.04438	0.18182	0.82544
0.20974	0.06452	0.5	0.90033	0.34483	0.12158	0.5	0.70292	0.23896	0.04321	0.33333	0.90033
0.2395	0.1379	1	0.96364	0.33446	0.05917	0.375	0.82872	0.22776	0.05531	0.36364	0.90127
0.10109	0.07136	0.22222	0.7243	0.22353	0.16609	0.4	0.7243	0.13737	0.09697	0.28571	0.7243
0.15396	0.09053	1	0.8975	0.2395	0.06253	0.28571	0.78212	0.14881	0.03109	0.25	0.8975
0.03076	0.0293	0.09091	0.69637	0.16667	0.15873	0.5	0.69637	0.05169	0.04923	0.15385	0.69637
0.12858	0.04276	0.33333	0.87928	0.33798	0.1149	0.42857	0.77862	0.16991	0.03853	0.26087	0.77862
0.18039	0.09252	1	0.95294	0.31429	0.0849	0.42857	0.79817	0.18601	0.02914	0.28571	0.87385
0.25175	0.17285	1	0.92393	0.3625	0.02312	0.375	0.87475	0.24253	0.08311	0.5	0.87475
0.10373	0.05928	0.33333	0.9406	0.27292	0.06278	0.5	0.61801	0.13044	0.04361	0.28571	0.87761
0.23428	0.12495	1	0.90319	0.40549	0.11065	0.5	0.76666	0.2399	0.04172	0.4	0.88377
0.12641	0.05449	0.5	0.87702	0.25794	0.07187	0.42857	0.6736	0.14717	0.02968	0.22222	0.87702
0.23126	0.07356	0.5	0.87385	0.40394	0.1325	0.71429	0.66738	0.26154	0.04619	0.375	0.80317
0.19795	0.07501	0.42857	0.83333	0.39286	0.1183	0.71429	0.66738	0.24082	0.0676	0.42857	0.83333
0.2502	0.08053	0.5	0.8765	0.37143	0.13029	0.57143	0.70873	0.26478	0.04654	0.375	0.80317
0.00805	0.01301	0.05	0.67561	0.04255	0.06881	0.25	0.67561	0.01353	0.02188	0.08333	0.67561
0.01123	0.01497	0.04878	0.66083	0.06032	0.08042	0.28571	0.66083	0.01889	0.02519	0.08333	0.66083
0.14516	0.04312	0.33333	0.84373	0.26812	0.1172	0.5	0.67654	0.17447	0.04292	0.25	0.76766
0.02646	0.0195	0.06667	0.75765	0.12632	0.09307	0.2	0.75765	0.04329	0.0319	0.1	0.75765
0.0245	0.02162	0.0625	0.75765	0.09314	0.08218	0.16667	0.75765	0.03845	0.03393	0.09091	0.75765
0.17712	0.11448	1	0.9406	0.41667	0.09028	0.5	0.77594	0.19909	0.05942	0.44444	0.87385
0.16398	0.10014	0.66667	0.87385	0.26974	0.05194	0.5	0.6563	0.1735	0.05289	0.36364	0.87385
0.02796	0.02272	0.07143	0.779	0.09896	0.0804	0.16667	0.779	0.04318	0.03508	0.1	0.779
0.23052	0.16197	1	0.90319	0.29464	0.02511	0.42857	0.64551	0.20945	0.06854	0.44444	0.89251
0.18277	0.11471	1	0.91263	0.40385	0.04374	0.42857	0.82902	0.20718	0.07339	0.42857	0.82902
0.13286	0.06637	1	1	0.39781	0.0928	0.55556	0.69531	0.16274	0.03901	0.3	0.82794
0.01441	0.01834	0.05405	0.66577	0.08712	0.11088	0.33333	0.66577	0.02468	0.03141	0.09302	0.66577
0.04329	0.01619	0.07143	0.82794	0.24178	0.12245	0.625	0.67148	0.07121	0.02638	0.12658	0.67148
0.08788	0.03104	0.16667	0.90033	0.24908	0.09655	0.42857	0.67099	0.12358	0.03699	0.21053	0.82556
0.08382	0.04939	0.33333	0.85694	0.16667	0.02151	0.33333	0.63528	0.09951	0.03819	0.22222	0.85694
0.07325	0.04053	0.22222	0.82639	0.25893	0.04688	0.28571	0.82639	0.10491	0.04384	0.25	0.82639
0.08563	0.0287	0.25	0.91279	0.3341	0.11275	0.42857	0.76659	0.1249	0.02909	0.19048	0.83634
0.15402	0.08011	1	1	0.44542	0.09537	0.625	0.67258	0.18762	0.04749	0.33333	0.82794

Table B.10. Right full table with Jaro distance used in merge phase with Levenshtein distance used in linking phase.

Id diagnosis	Count Avg	Count Dev	Count Max	Count At	Score Avg	Score Dev	Score Max	Score At
exa:cervix/19/1000*	0.7	0.60667	3	0	0.32516	0.2818	0.60215	0.2
exa:cervix/19/1002	2.5	0.83333	4	0	0.56608	0.16398	0.69205	0.55939
exa:cervix/19/1004	2.6522	1.4934	5	0.2	0.46999	0.20434	0.61121	0.62426
exa:cervix/19/1006	2.3667	0.78444	4	0	0.61794	0.13043	0.74704	0.55939
exa:cervix/19/1008*	2.0286	1.0563	4	0	0.57201	0.23044	0.74704	0.55939
exa:cervix/19/1010_1A	1.4583	0.58681	2	0.44142	0.73241	0.13078	0.8669	0.64495
exa:cervix/19/1010_2B	2.3226	0.69927	3	0.44142	0.67195	0.06454	0.74641	0.64495
exa:cervix/19/1011*	1.6364	0.69972	4	0.2	0.73819	0.12121	0.8669	0.64495
exa:cervix/19/1012	1.6154	0.70414	3	0.26667	0.70263	0.16215	0.8669	0.64495
exa:cervix/19/1013	0.73684	0.77562	3	0.2	0.28714	0.30225	0.64851	0.25
exa:cervix/19/1014	1.3158	0.6482	2	0.44142	0.68307	0.21571	0.8669	0.64495
exa:cervix/19/1015	0.4	0.56	2	0.2	0.19851	0.27791	0.67926	0.34142
exa:cervix/19/1016	0.17647	0.29066	1	0.25774	0.12762	0.2102	0.7232	0.25774
exa:cervix/19/1017	1.84	1.0592	4	0.2	0.64625	0.2068	0.8669	0.64495
exa:cervix/19/1018	1.5	0.55	2	0.44142	0.66825	0.11919	0.8669	0.64495
exa:cervix/19/1020	1.4583	0.63194	2	0.44142	0.68183	0.17844	0.8669	0.64495
exa:cervix/19/1022*	0.53846	0.8284	3	0.2	0.15538	0.23905	0.6842	0.2
exa:cervix/19/1023	0.35714	0.45918	1	0.34142	0.24259	0.3119	0.67926	0.34142
exa:cervix/19/1024	0.31818	0.43388	1	0.34142	0.19752	0.26934	0.62077	0.34142
exa:cervix/19/1025*	1.7083	0.84028	4	0.2	0.69393	0.17348	0.8669	0.64495
exa:cervix/19/1789	0	0	0	0.77321	0	0	0	0.77321
exa:cervix/19/1790	1.3103	0.68728	3	0	0.66485	0.23957	0.8669	0.64495
exa:cervix/19/1792	2.3448	1.0868	4	0.2	0.53382	0.22089	0.72082	0.6633
exa:cervix/19/1793	1.6452	0.59105	3	0	0.61974	0.16823	0.79408	0.65939
exa:cervix/19/1794	1.2308	0.60947	3	0	0.6113	0.20222	0.79408	0.65939
exa:cervix/19/1795	2.2143	0.68878	4	0	0.6944	0.06185	0.74704	0.65939
exa:cervix/19/1797*	2.3143	0.74449	3	0.44142	0.67709	0.07738	0.74704	0.65939
exa:cervix/19/1798	2.3929	0.6352	4	0	0.70785	0.03478	0.74704	0.65939
exa:cervix/19/1799	0.59091	0.53719	2	0	0.35247	0.32159	0.67926	0.34142
exa:cervix/19/1800	2.25	0.775	4	0.2	0.65687	0.13137	0.74704	0.58868
exa:cervix/19/1802	2.0909	0.66942	4	0.2	0.66368	0.12067	0.74704	0.58868
exa:cervix/19/1803	2.1579	0.70914	3	0.44142	0.65461	0.13781	0.74704	0.58868
exa:cervix/19/1807	1.129	0.8283	3	0.28165	0.40992	0.26446	0.67671	0.32845
exa:cervix/19/1808	2.0714	1.2194	4	0.28165	0.47925	0.17116	0.65952	0.6
exa:cervix/19/1809	0.83333	0.27778	1	0.62426	0.48506	0.16169	0.58207	0.62426
exa:cervix/19/1811	1.6	0.48	2	0.44142	0.60499	0.02416	0.62702	0.34142
exa:cervix/19/1812*	1.5	0.63636	2	0.44142	0.52835	0.1441	0.62702	0.34142
exa:cervix/19/279	2.6	0.58667	3	0.58284	0.70682	0.01032	0.74704	0.65939
exa:cervix/19/280	2.3448	0.68014	4	0.2	0.70528	0.05012	0.74704	0.65939
exa:cervix/19/281	1.3809	0.93424	3	0.2	0.49532	0.28304	0.71206	0.58284
exa:cervix/19/282*	2.2857	0.87075	4	0.2	0.67096	0.06645	0.74704	0.65939
exa:cervix/19/966	2.2812	0.89844	3	0.56042	0.58881	0.18442	0.73648	0.34142
exa:cervix/19/969	2.1136	0.90186	5	0	0.57706	0.18561	0.74704	0.55939
exa:cervix/19/975	2.6333	0.78667	4	0.2	0.55551	0.14814	0.67094	0.62426
exa:cervix/19/985	2.925	1.0012	4	0.44142	0.65669	0.0985	0.73908	0.55939
exa:cervix/19/990	0.61538	0.80473	2	0.33333	0.21876	0.28607	0.67926	0.34142
exa:cervix/19/993	1.2778	1.1667	4	0.2	0.35754	0.31781	0.66207	0.39712
exa:cervix/19/995	2.25	0.95833	3	0.56042	0.57038	0.19013	0.72445	0.34142
exa:cervix/19/996	2.1351	0.60774	3	0.34142	0.65077	0.07741	0.72457	0.4
exa:cervix/19/997	2.1628	0.86425	4	0	0.58488	0.19452	0.79408	0.65939

Table B.11. Left full table with cosine distance used in merge phase with Jaro distance used in linking phase.

Prec Avg	Prec Dev	Prec Max	Prec At	Rec Avg	Rec Dev	Rec Max	Rec At	F_1 Avg	F_1 Dev	F_1 Max	F_1 At
0.02261	0.01959	0.06	0	0.1	0.08667	0.42857	0	0.03635	0.03151	0.10526	0
0.08552	0.04098	0.1875	0.65689	0.27778	0.09259	0.44444	0	0.12589	0.05154	0.24	0.65689
0.10706	0.04655	0.18182	0.62426	0.44203	0.2489	0.83333	0.2	0.16703	0.07262	0.26667	0.34142
0.11255	0.04381	0.25	0.64142	0.3381	0.11206	0.57143	0	0.15795	0.04311	0.26667	0.64142
0.06696	0.02919	0.125	0.56042	0.25357	0.13204	0.5	0	0.10209	0.0437	0.18182	0.44142
0.12197	0.04888	0.33333	0.64495	0.20833	0.08383	0.28571	0.44142	0.13928	0.03508	0.21053	0.44142
0.16319	0.07278	0.5	0.82426	0.29032	0.08741	0.375	0.44142	0.17877	0.03734	0.28571	0.64495
0.10107	0.04233	0.33333	0.64495	0.18182	0.07775	0.44444	0.2	0.11297	0.02734	0.16667	0.44142
0.10133	0.04088	0.25	0.64495	0.23077	0.10059	0.42857	0.26667	0.12799	0.03496	0.2	0.44142
0.03215	0.03384	0.09524	0.25	0.08187	0.08618	0.33333	0.2	0.045	0.04737	0.13333	0.25
0.11436	0.04735	0.25	0.64495	0.18797	0.0926	0.28571	0.44142	0.1328	0.04557	0.2	0.44142
0.01599	0.02238	0.0625	0.2	0.05	0.07	0.25	0.2	0.02389	0.03345	0.1	0.2
0.00788	0.01298	0.05556	0.25774	0.02206	0.03633	0.125	0.25774	0.01155	0.01902	0.07692	0.25774
0.10534	0.04176	0.17647	0.4	0.184	0.10592	0.4	0.2	0.12595	0.05207	0.22222	0.4
0.14678	0.07734	0.5	0.64495	0.16667	0.06111	0.22222	0.44142	0.13123	0.03576	0.2	0.44142
0.08648	0.03218	0.14286	0.64495	0.18229	0.07899	0.25	0.44142	0.11221	0.03723	0.18182	0.44142
0.0189	0.02907	0.10345	0.2	0.07692	0.11834	0.42857	0.2	0.03027	0.04657	0.16667	0.2
0.0211	0.02713	0.07143	0.34142	0.05952	0.07653	0.16667	0.34142	0.03106	0.03993	0.1	0.34142
0.01736	0.02367	0.06667	0.34142	0.03977	0.05424	0.125	0.34142	0.02401	0.03274	0.08696	0.34142
0.12206	0.04116	0.25	0.64495	0.21354	0.10503	0.5	0.2	0.14282	0.04387	0.21053	0.44142
0	0	0	0.77321	0	0	0	0.77321	0	0	0	0.77321
0.07823	0.03121	0.16667	0.64495	0.18719	0.09818	0.42857	0	0.10483	0.03892	0.17391	0.44142
0.11981	0.06217	0.27273	0.62426	0.3908	0.18113	0.66667	0.2	0.17691	0.08288	0.35294	0.62426
0.09482	0.04279	0.2	0.67589	0.23502	0.08444	0.42857	0	0.1268	0.04367	0.23529	0.56449
0.06978	0.02647	0.14286	0.8	0.17582	0.08707	0.42857	0	0.09524	0.03377	0.14815	0.44142
0.1743	0.07239	0.5	0.84142	0.27679	0.0861	0.5	0	0.18567	0.03415	0.28571	0.67589
0.12891	0.0552	0.33333	0.84142	0.28929	0.09306	0.375	0.44142	0.1586	0.04827	0.25	0.67589
0.17437	0.0832	0.5	0.84142	0.29911	0.0794	0.5	0	0.18735	0.03306	0.30769	0.7266
0.02663	0.02421	0.06667	0.44142	0.08442	0.07674	0.28571	0	0.03991	0.03628	0.09524	0
0.22167	0.08897	0.5	0.59424	0.32143	0.11071	0.57143	0.2	0.24029	0.06405	0.36364	0.59424
0.19121	0.07711	0.4	0.59424	0.2987	0.09563	0.57143	0.2	0.21236	0.05251	0.33333	0.59424
0.22711	0.08787	0.5	0.59424	0.30827	0.10131	0.42857	0.44142	0.2424	0.06498	0.36364	0.59424
0.04473	0.02957	0.09375	0.28165	0.14113	0.10354	0.375	0.28165	0.06555	0.04305	0.15	0.28165
0.09183	0.03599	0.15	0.4	0.29592	0.1742	0.57143	0.28165	0.13481	0.0575	0.22222	0.4
0.07905	0.04081	0.2	0.62426	0.13889	0.0463	0.16667	0.62426	0.09631	0.03987	0.18182	0.62426
0.16711	0.11582	1	0.68944	0.32	0.096	0.4	0.44142	0.16624	0.04558	0.33333	0.68944
0.09562	0.03888	0.2	0.57889	0.25	0.10606	0.33333	0.44142	0.13077	0.04392	0.22222	0.44142
0.20729	0.10974	1	0.84142	0.43333	0.09778	0.5	0.58284	0.22489	0.0548	0.36364	0.67589
0.1843	0.09312	0.5	0.84142	0.2931	0.08502	0.5	0.2	0.19401	0.04169	0.33333	0.67589
0.0737	0.04211	0.125	0.58284	0.23016	0.15571	0.5	0.2	0.10816	0.06181	0.18182	0.44142
0.20774	0.07081	0.5	0.84142	0.32653	0.12439	0.57143	0.2	0.2224	0.04178	0.30769	0.67589
0.10379	0.04806	0.2	0.68284	0.32589	0.12835	0.42857	0.56042	0.15117	0.06081	0.26087	0.56042
0.05758	0.02092	0.11111	0.64142	0.23485	0.10021	0.55556	0	0.08924	0.03061	0.14815	0.64142
0.14584	0.07307	0.33333	0.62426	0.43889	0.13111	0.66667	0.2	0.2066	0.08288	0.4	0.58284
0.09488	0.02883	0.16667	0.72426	0.36563	0.12516	0.5	0.44142	0.14242	0.03481	0.21429	0.56042
0.02207	0.02886	0.08696	0.33333	0.08791	0.11496	0.28571	0.33333	0.03508	0.04587	0.13333	0.33333
0.0643	0.05715	0.15385	0.34142	0.21296	0.19444	0.66667	0.2	0.09677	0.08602	0.23529	0.2
0.08331	0.04273	0.16667	0.64142	0.32143	0.1369	0.42857	0.56042	0.12648	0.05697	0.24	0.56042
0.10055	0.04171	0.33333	0.88284	0.30502	0.08682	0.42857	0.34142	0.1352	0.03163	0.23529	0.72426
0.07251	0.03168	0.16667	0.64142	0.27035	0.10803	0.5	0	0.10915	0.04096	0.2	0.64142

Table B.12. Right full table with cosine distance used in merge phase with Jaro distance used in linking phase.

Id diagnosis	Count Avg	Count Dev	Count Max	Count At	Score Avg	Score Dev	Score Max	Score At
exa:cervix/19/1000*	1.8571	0.98776	3	10.6	0.24057	0.16496	0.40143	10.6
exa:cervix/19/1002	3.3478	1.0492	4	5	0.42859	0.12444	0.51904	16
exa:cervix/19/1004	3.6571	1.631	5	7.6	0.51431	0.14695	0.63919	4.8
exa:cervix/19/1006	3.0323	0.87409	4	10.4	0.48664	0.09943	0.68963	5.4
exa:cervix/19/1008*	2.8913	1.0397	4	10.4	0.42315	0.14295	0.56154	10.4
exa:cervix/19/1010_1A	1.625	0.53125	2	6.8	0.65449	0.10908	0.77308	6.8
exa:cervix/19/1010_2B	2.1389	0.76543	3	9.4	0.55016	0.15282	0.67138	5
exa:cervix/19/1011*	2.718	0.7048	4	19.2	0.64176	0.06804	0.74131	7
exa:cervix/19/1012	1.8	0.53333	3	19.4	0.65103	0.17361	0.77308	6.8
exa:cervix/19/1013	1.3667	0.75556	3	17.6	0.52095	0.13892	0.64851	14.2
exa:cervix/19/1014	1.6	0.56	2	6.8	0.67701	0.1354	0.77308	6.8
exa:cervix/19/1015	1.4333	0.79333	2	5.2	0.46122	0.24598	0.67926	5
exa:cervix/19/1016	0.21053	0.33241	1	14	0.15225	0.2404	0.7232	14
exa:cervix/19/1017	2.4412	1.0761	4	16.8	0.57803	0.17075	0.70666	6.8
exa:cervix/19/1018	1.6452	0.52653	2	6.2	0.64487	0.12821	0.77308	6.8
exa:cervix/19/1020	1.4546	0.69421	2	7	0.55263	0.20096	0.77308	7
exa:cervix/19/1022*	1.28	0.5632	3	20	0.62861	0.10058	0.70049	14
exa:cervix/19/1023	0.78947	0.33241	1	5	0.53626	0.22579	0.67926	5
exa:cervix/19/1024	0.47619	0.49887	1	7	0.2956	0.30968	0.62077	7
exa:cervix/19/1025*	2.5152	0.85767	4	19.4	0.66264	0.09129	0.74131	7
exa:cervix/19/1789	0	0	0	7.4	0	0	0	7.4
exa:cervix/19/1790	2.6923	0.52071	3	7	0.50141	0.05101	0.67926	6.2
exa:cervix/19/1792	2.1333	0.96889	4	12.6	0.60033	0.08118	0.66061	3
exa:cervix/19/1793	2.6111	0.66975	3	5	0.41957	0.09324	0.66349	4.8
exa:cervix/19/1794	2.4483	0.8371	3	5	0.43642	0.0737	0.59545	3.6
exa:cervix/19/1795	3.1724	0.74197	4	10.2	0.5122	0.06588	0.7	2.8
exa:cervix/19/1797*	2.6279	0.58843	3	4.6	0.62093	0.08664	0.70988	14.2
exa:cervix/19/1798	3.0968	0.81582	4	10	0.48466	0.08643	0.7	3.6
exa:cervix/19/1799	1.6	0.64	2	5	0.2717	0.10868	0.33963	5
exa:cervix/19/1800	2.48	0.8576	4	15.4	0.67813	0.05425	0.74872	10.2
exa:cervix/19/1802	2.12	1.104	4	15.4	0.67372	0.0539	0.73963	10.2
exa:cervix/19/1803	2.5455	0.66116	3	5.8	0.65359	0.05942	0.72445	14.4
exa:cervix/19/1807	1.8947	0.52909	3	19.2	0.56698	0.06137	0.65952	4
exa:cervix/19/1808	2.5625	0.91406	4	19.6	0.58378	0.03185	0.65952	0.8
exa:cervix/19/1809	0.31579	0.43213	1	11.2	0.18381	0.25153	0.58207	11.2
exa:cervix/19/1811	1.7647	0.35986	2	5.4	0.61473	0.0188	0.62702	5.4
exa:cervix/19/1812*	1.7241	0.39952	2	5.4	0.61261	0.02087	0.62702	5.4
exa:cervix/19/279	2.6286	0.59429	3	5.6	0.67789	0.02108	0.70988	14.2
exa:cervix/19/280	2.5385	0.81065	4	16.2	0.67951	0.05227	0.74872	10
exa:cervix/19/281	1.6539	1.1154	3	9	0.52049	0.24023	0.69566	8.2
exa:cervix/19/282*	3	0.36364	4	16.2	0.67139	0.00592	0.7	3.2
exa:cervix/19/966	2.4706	0.84083	3	5.6	0.59151	0.17397	0.73648	16.4
exa:cervix/19/969	4.2292	1.2205	5	5	0.46216	0.09941	0.60175	3.4
exa:cervix/19/975	2.1667	1.5648	4	12.2	0.41734	0.25504	0.61836	12.2
exa:cervix/19/985	2.8868	1.147	4	8.8	0.60504	0.11503	0.69745	16.4
exa:cervix/19/990	0.71875	0.62891	2	15.4	0.35196	0.30796	0.67926	6.4
exa:cervix/19/993	2.3462	1.2426	4	13.6	0.49709	0.19119	0.64224	13.6
exa:cervix/19/995	2.383	0.91897	3	5	0.57656	0.17174	0.70808	16.8
exa:cervix/19/996	2.2093	0.73553	3	9.8	0.56864	0.07935	0.66077	12.6
exa:cervix/19/997	3.4898	0.85381	4	5	0.44017	0.09142	0.60175	3.4

Table B.13. Left full table with Hamming distance used in merge phase with Jaro distance used in linking phase.

Prec Avg	Prec Dev	Prec Max	Prec At	Rec Avg	Rec Dev	Rec Max	Rec At	F_1 Avg	F_1 Dev	F_1 Max	F_1 At
0.05429	0.02219	0.08696	6.6	0.26531	0.14111	0.42857	10.6	0.08873	0.03776	0.14286	10.6
0.10616	0.0605	0.30769	5	0.37198	0.11657	0.44444	5	0.15328	0.06818	0.36364	5
0.13123	0.04341	0.21739	7.6	0.60952	0.27184	0.83333	7.6	0.2119	0.06924	0.34483	7.6
0.12737	0.0349	0.21429	6	0.43318	0.12487	0.57143	10.4	0.18646	0.0386	0.28571	6
0.09317	0.03956	0.25	4.6	0.36141	0.12996	0.5	10.4	0.14027	0.04929	0.3	4.6
0.10801	0.04153	0.33333	5	0.23214	0.07589	0.28571	6.8	0.13625	0.03559	0.21053	6.8
0.07959	0.03357	0.22222	4.2	0.26736	0.09568	0.375	9.4	0.11527	0.03713	0.23529	4.2
0.11748	0.04069	0.23077	5.6	0.30199	0.07831	0.44444	19.2	0.15874	0.04034	0.27273	5.6
0.0832	0.03245	0.16667	5.4	0.25714	0.07619	0.42857	19.4	0.12172	0.04031	0.21053	5.4
0.05671	0.02295	0.125	4	0.15185	0.08395	0.33333	17.6	0.07845	0.03058	0.13333	17.6
0.11319	0.04228	0.33333	5	0.22857	0.08	0.28571	6.8	0.14104	0.03879	0.22222	6.8
0.06308	0.03502	0.15385	5.2	0.17917	0.09917	0.25	5.2	0.09162	0.04891	0.19048	5.2
0.00861	0.0136	0.04348	14	0.02632	0.04155	0.125	14	0.01297	0.02048	0.06452	14
0.09739	0.0329	0.16667	5	0.24412	0.10761	0.4	16.8	0.13352	0.04564	0.21429	6.8
0.08037	0.03452	0.25	5	0.1828	0.0585	0.22222	6.2	0.10133	0.02768	0.17391	6.2
0.0515	0.02013	0.09091	7	0.18182	0.08678	0.25	7	0.07884	0.02953	0.13333	7
0.07011	0.02345	0.16667	5.4	0.18286	0.08046	0.42857	20	0.09678	0.02982	0.16216	20
0.07444	0.04224	0.2	5	0.13158	0.0554	0.16667	5	0.09084	0.04205	0.18182	5
0.02057	0.02155	0.05263	7	0.05952	0.06236	0.125	7	0.03049	0.03194	0.07407	7
0.14429	0.04627	0.25	3	0.31439	0.10721	0.5	19.4	0.18419	0.04598	0.3	5.6
0	0	0	7.4	0	0	0	7.4	0	0	0	7.4
0.16054	0.06058	0.33333	7	0.38462	0.07439	0.42857	7	0.21216	0.05582	0.375	7
0.10233	0.02642	0.25	3	0.35556	0.16148	0.66667	12.6	0.14566	0.03108	0.2	3
0.13954	0.07496	0.42857	5	0.37302	0.09568	0.42857	5	0.18987	0.07757	0.42857	5
0.14604	0.05745	0.3	5	0.34975	0.11959	0.42857	5	0.19197	0.06045	0.35294	5
0.18411	0.07016	0.5	2.8	0.39655	0.09275	0.5	10.2	0.21987	0.03841	0.375	5
0.10269	0.04873	0.33333	3.4	0.32849	0.07355	0.375	4.6	0.14342	0.04799	0.28571	3.4
0.18826	0.08016	0.5	3.6	0.3871	0.10198	0.5	10	0.22014	0.04386	0.33333	6
0.07648	0.03766	0.2	5	0.22857	0.09143	0.28571	5	0.11165	0.04877	0.23529	5
0.19163	0.0718	0.5	3	0.35429	0.12251	0.57143	15.4	0.22062	0.03484	0.33333	5
0.14621	0.05337	0.5	2.8	0.30286	0.15771	0.57143	15.4	0.1688	0.04083	0.23529	15.4
0.24337	0.10149	0.5	3.6	0.36364	0.09445	0.42857	5.8	0.26242	0.06439	0.4	5.8
0.08138	0.027	0.2	4	0.23684	0.06614	0.375	19.2	0.11165	0.02189	0.17391	6
0.15879	0.09211	1	0.8	0.36607	0.13058	0.57143	19.6	0.16835	0.02862	0.25	0.8
0.01651	0.0226	0.0625	11.2	0.05263	0.07202	0.16667	11.2	0.02509	0.03433	0.09091	11.2
0.14562	0.0892	1	1.8	0.35294	0.07197	0.4	5.4	0.16191	0.04344	0.33333	1.8
0.13597	0.06057	0.5	3.8	0.28736	0.06659	0.33333	5.4	0.15964	0.03529	0.25	3.8
0.17107	0.09933	1	3.2	0.4381	0.09905	0.5	5.6	0.19265	0.05289	0.35294	5.6
0.17426	0.09454	0.66667	4.6	0.31731	0.10133	0.5	16.2	0.1888	0.03757	0.36364	4.6
0.07561	0.0372	0.13043	9	0.27564	0.1859	0.5	9	0.11602	0.05982	0.2069	9
0.28429	0.18532	1	3.2	0.42857	0.05195	0.57143	16.2	0.27747	0.08202	0.5	5.4
0.09223	0.04534	0.25	5.6	0.35294	0.12012	0.42857	5.6	0.13962	0.0568	0.31579	5.6
0.12145	0.0656	0.33333	5	0.46991	0.13561	0.55556	5	0.18019	0.07734	0.41667	5
0.07138	0.04569	0.13793	12.2	0.36111	0.2608	0.66667	12.2	0.11801	0.07727	0.22857	12.2
0.07665	0.02929	0.22222	3.4	0.36085	0.14338	0.5	8.8	0.118	0.03417	0.23529	3.4
0.02426	0.02123	0.05882	15.4	0.10268	0.08984	0.28571	15.4	0.03903	0.03416	0.09756	15.4
0.13112	0.0509	0.22222	5.4	0.39103	0.2071	0.66667	13.6	0.1903	0.07965	0.3	9.8
0.07905	0.04091	0.2	5	0.34043	0.13128	0.42857	5	0.12138	0.05214	0.27273	5
0.07544	0.02704	0.2	3.2	0.31561	0.10508	0.42857	9.8	0.11204	0.02605	0.21053	4.4
0.11578	0.06535	0.33333	5	0.43622	0.10673	0.5	5	0.17069	0.07617	0.4	5

Table B.14. Right full table with Hamming distance used in merge phase with Jaro distance used in linking phase.

Id diagnosis	Count Avg	Count Dev	Count Max	Count At	Score Avg	Score Dev	Score Max	Score At
exa:cervix/19/1000*	1.6571	1.022	3	10.4	0.23025	0.15789	0.40143	10.4
exa:cervix/19/1002	3.3696	1.0416	4	4	0.45499	0.11869	0.7	3.6
exa:cervix/19/1004	3.7419	1.5442	5	6.6	0.50679	0.16348	0.63957	4
exa:cervix/19/1006	3.1818	0.89256	4	7.6	0.53217	0.08613	0.68963	3.8
exa:cervix/19/1008*	2.7805	1.1612	4	8.2	0.4281	0.14618	0.68963	3.8
exa:cervix/19/1010_1A	1.7619	0.40816	2	5.2	0.66262	0.12621	0.77308	5.2
exa:cervix/19/1010_2B	2.3226	0.87409	3	5.4	0.54817	0.14146	0.67138	4.6
exa:cervix/19/1011*	2.6389	0.79938	4	16.6	0.64343	0.07437	0.77308	5.2
exa:cervix/19/1012	1.7	0.61333	3	17.4	0.63447	0.21149	0.77308	5.2
exa:cervix/19/1013	1.2258	0.78668	3	13.8	0.46207	0.20867	0.64851	11.6
exa:cervix/19/1014	1.6	0.6	2	5.2	0.64774	0.19432	0.77308	5.2
exa:cervix/19/1015	1.3793	0.85612	2	5	0.43209	0.2682	0.62653	5
exa:cervix/19/1016	0.22727	0.35124	1	11	0.16436	0.25402	0.7232	11
exa:cervix/19/1017	2.3889	1.3148	4	13	0.53129	0.23613	0.70666	5.2
exa:cervix/19/1018	1.56	0.6688	2	5.2	0.54413	0.21765	0.77308	5.2
exa:cervix/19/1020	1.4286	0.81633	2	5	0.48637	0.27793	0.77308	5
exa:cervix/19/1022*	1.2	0.672	3	17	0.54819	0.21927	0.70049	11
exa:cervix/19/1023	0.78947	0.33241	1	5	0.53626	0.22579	0.67926	5
exa:cervix/19/1024	0.3913	0.47637	1	7	0.24291	0.29572	0.62077	7
exa:cervix/19/1025*	2.5333	0.85333	4	17.2	0.67418	0.10269	0.77308	5.2
exa:cervix/19/1789	0	0	0	5.8	0	0	0	5.8
exa:cervix/19/1790	2.5417	0.6875	3	6.6	0.55933	0.11252	0.8669	5.2
exa:cervix/19/1792	2.0625	1.0234	4	10.4	0.56851	0.14213	0.66993	5.2
exa:cervix/19/1793	2.4	0.88	3	5	0.46004	0.09948	0.67926	3.2
exa:cervix/19/1794	2.4	0.8	3	5	0.45126	0.10782	0.63735	3.2
exa:cervix/19/1795	3.0323	0.99896	4	7.4	0.5696	0.07997	0.7	1.8
exa:cervix/19/1797*	2.7105	0.50277	3	3.2	0.65719	0.06918	0.70988	8
exa:cervix/19/1798	3.3793	0.85612	4	6	0.5381	0.08062	0.7	1.8
exa:cervix/19/1799	1.4074	0.83402	2	5	0.239	0.14163	0.33963	5
exa:cervix/19/1800	2.4583	0.78472	4	13.4	0.67578	0.05631	0.74872	7.4
exa:cervix/19/1802	2.28	1.0688	4	13.4	0.67936	0.05516	0.73963	7.4
exa:cervix/19/1803	2.45	0.66	3	5.6	0.65669	0.06567	0.72445	9.4
exa:cervix/19/1807	1.8788	0.65381	3	13.4	0.56465	0.07153	0.65952	3.8
exa:cervix/19/1808	2.697	0.87052	4	13.4	0.58339	0.0297	0.65952	0.8
exa:cervix/19/1809	0.25	0.375	1	10.6	0.14552	0.21828	0.58207	10.6
exa:cervix/19/1811	1.931	0.12842	2	3.2	0.62342	0.00671	0.62702	3.2
exa:cervix/19/1812*	1.9643	0.06888	2	3.2	0.62516	0.0036	0.62702	3.2
exa:cervix/19/279	2.6765	0.45675	3	5.2	0.69609	0.01507	0.70988	8
exa:cervix/19/280	2.75	0.80357	4	13.6	0.68296	0.04878	0.74872	7.4
exa:cervix/19/281	1.9167	0.9375	3	7.8	0.59353	0.14838	0.69566	5.8
exa:cervix/19/282*	2.88	0.432	4	13.6	0.67325	0.00738	0.7	1.8
exa:cervix/19/966	2.5526	0.65928	3	5.2	0.64707	0.10217	0.73648	10.4
exa:cervix/19/969	4.2	1.28	5	4.2	0.46355	0.11125	0.6345	3.8
exa:cervix/19/975	2.3514	1.5471	4	10	0.43433	0.25825	0.64034	4.8
exa:cervix/19/985	3.0455	1.2583	4	5.4	0.59121	0.16357	0.72175	10
exa:cervix/19/990	0.65517	0.67776	2	13.8	0.30274	0.31318	0.67926	5.8
exa:cervix/19/993	2.2593	1.1769	4	12	0.49984	0.18513	0.64224	12
exa:cervix/19/995	2.425	0.92	3	3.8	0.56507	0.19777	0.70808	10
exa:cervix/19/996	2.5	0.725	3	4.2	0.59368	0.08905	0.71349	2.4
exa:cervix/19/997	3.4151	0.94909	4	4	0.44778	0.10138	0.68092	3.8

Table B.15. Left full table with Levenshtein distance used in merge phase with Jaro distance used in linking phase.

Prec Avg	Prec Dev	Prec Max	Prec At	Rec Avg	Rec Dev	Rec Max	Rec At	F_1 Avg	F_1 Dev	F_1 Max	F_1 At
0.04588	0.02624	0.08333	5.4	0.23673	0.14601	0.42857	10.4	0.07616	0.04405	0.12903	5.4
0.09492	0.04641	0.25	4	0.3744	0.11573	0.44444	4	0.14445	0.05958	0.32	4
0.1368	0.05319	0.25	4.2	0.62366	0.25737	0.83333	6.6	0.21942	0.08039	0.36364	4.2
0.15653	0.06275	0.66667	3.8	0.45455	0.12751	0.57143	7.6	0.20556	0.03867	0.4	3.8
0.08567	0.03511	0.21429	4	0.34756	0.14515	0.5	8.2	0.13202	0.04976	0.27273	4
0.11658	0.05125	0.28571	5.2	0.2517	0.05831	0.28571	5.2	0.15112	0.04911	0.28571	5.2
0.08058	0.03013	0.15385	4.2	0.29032	0.10926	0.375	5.4	0.12111	0.03831	0.2	5.4
0.11524	0.04236	0.33333	3.2	0.29321	0.08882	0.44444	16.6	0.14918	0.03173	0.23077	5.6
0.0801	0.0371	0.18182	5	0.24286	0.08762	0.42857	17.4	0.11562	0.0461	0.22222	5
0.04609	0.02448	0.08824	13.8	0.1362	0.08741	0.33333	13.8	0.06706	0.03526	0.13953	13.8
0.11142	0.04849	0.22222	5.2	0.22857	0.08571	0.28571	5.2	0.14352	0.05023	0.25	5.2
0.05479	0.03401	0.13333	5	0.17241	0.10702	0.25	5	0.08228	0.05107	0.17391	5
0.0095	0.01469	0.04545	11	0.02841	0.0439	0.125	11	0.01423	0.022	0.06667	11
0.08574	0.03887	0.1875	5.2	0.23889	0.13148	0.4	13	0.12228	0.05955	0.23077	5.2
0.06281	0.03068	0.18182	5.2	0.17333	0.07431	0.22222	5.2	0.08794	0.03699	0.2	5.2
0.04841	0.02798	0.11765	5	0.17857	0.10204	0.25	5	0.07519	0.04296	0.16	5
0.05715	0.02525	0.1	5.4	0.17143	0.096	0.42857	17	0.08382	0.03687	0.16216	17
0.07087	0.03689	0.16667	5	0.13158	0.0554	0.16667	5	0.08946	0.04063	0.16667	5
0.0165	0.02008	0.05	7	0.04891	0.05955	0.125	7	0.02462	0.02997	0.07143	7
0.13232	0.03775	0.25	3	0.31667	0.10667	0.5	17.2	0.17468	0.03869	0.25	5.6
0	0	0	5.8	0	0	0	5.8	0	0	0	5.8
0.16513	0.06841	0.5	5.2	0.3631	0.09821	0.42857	6.6	0.19964	0.04453	0.31579	6.6
0.09008	0.02864	0.2	3	0.34375	0.17057	0.66667	10.4	0.13377	0.04108	0.2	10.4
0.12444	0.05204	0.25	5	0.34286	0.12571	0.42857	5	0.17111	0.05768	0.31579	5
0.14317	0.06026	0.28571	3.2	0.34286	0.11429	0.42857	5	0.18834	0.05758	0.3	5
0.21457	0.10554	1	1.8	0.37903	0.12487	0.5	7.4	0.21803	0.03075	0.30769	3.2
0.11679	0.07203	0.375	3.2	0.33882	0.06285	0.375	3.2	0.15156	0.06159	0.375	3.2
0.25335	0.1351	1	1.8	0.42241	0.10702	0.5	6	0.25159	0.05381	0.36364	3.2
0.05893	0.03506	0.15385	5	0.20106	0.11915	0.28571	5	0.08985	0.05324	0.2	5
0.23111	0.10983	0.66667	3.2	0.35119	0.1121	0.57143	13.4	0.24228	0.04543	0.4	3.2
0.15635	0.05519	0.5	2.6	0.32571	0.15269	0.57143	13.4	0.18313	0.03853	0.25	7.4
0.26003	0.11782	0.66667	3.2	0.35	0.09429	0.42857	5.6	0.26334	0.06209	0.4	3.2
0.07973	0.02336	0.25	3.8	0.23485	0.08173	0.375	13.4	0.10947	0.02107	0.16667	3.8
0.15387	0.07233	1	0.8	0.38528	0.12436	0.57143	13.4	0.17466	0.01898	0.25	0.8
0.01256	0.01884	0.05556	10.6	0.04167	0.0625	0.16667	10.6	0.01929	0.02893	0.08333	10.6
0.16782	0.11705	1	1.6	0.38621	0.02568	0.4	3.2	0.18674	0.07205	0.4	3.2
0.19479	0.13568	1	2.8	0.32738	0.01148	0.33333	3.2	0.19934	0.07492	0.44444	3.2
0.18593	0.11992	1	1.6	0.44608	0.07612	0.5	5.2	0.208	0.0665	0.44444	3.2
0.18242	0.10158	0.66667	3.2	0.34375	0.10045	0.5	13.6	0.19758	0.03918	0.36364	3.2
0.1082	0.03249	0.25	3.4	0.31944	0.15625	0.5	7.8	0.15078	0.04406	0.21429	7.8
0.28649	0.18585	1	1.8	0.41143	0.06171	0.57143	13.6	0.26858	0.06901	0.44444	3.2
0.10875	0.05256	0.33333	3.8	0.36466	0.09418	0.42857	5.2	0.15285	0.0519	0.30769	3.8
0.10283	0.0507	0.27778	4.2	0.46667	0.14222	0.55556	4.2	0.15997	0.0663	0.37037	4.2
0.07871	0.04767	0.13793	10	0.39189	0.25785	0.66667	10	0.12973	0.07924	0.22857	10
0.07137	0.0279	0.16667	3.4	0.38068	0.15728	0.5	5.4	0.11536	0.03839	0.21053	5.4
0.02101	0.02173	0.05882	13.8	0.0936	0.09682	0.28571	13.8	0.03416	0.03534	0.09756	13.8
0.12576	0.04664	0.22222	5	0.37654	0.19616	0.66667	12	0.18211	0.07245	0.26667	5
0.07436	0.04023	0.27273	3.8	0.34643	0.13143	0.42857	3.8	0.11695	0.05458	0.33333	3.8
0.08419	0.03509	0.2	3.2	0.35714	0.10357	0.42857	4.2	0.12718	0.04028	0.23529	3.2
0.10644	0.05587	0.30769	4	0.42689	0.11864	0.5	4	0.16055	0.06961	0.38095	4

Table B.16. Right full table with Levenshtein distance used in merge phase with Jaro distance used in linking phase.

Id diagnosis	Count Avg	Count Dev	Count Max	Count At	Score Avg	Score Dev	Score Max	Score At
exa:cervix/19/1000*	1.6571	1.022	3	10.4	0.23025	0.15789	0.40143	10.4
exa:cervix/19/1002	3.3696	1.0416	4	4	0.45499	0.11869	0.7	3.6
exa:cervix/19/1004	3.7419	1.5442	5	6.6	0.50679	0.16348	0.63957	4
exa:cervix/19/1006	3.2059	0.88754	4	7.6	0.53167	0.08395	0.68963	3.8
exa:cervix/19/1008*	2.7805	1.1612	4	8.2	0.42766	0.14603	0.68963	3.8
exa:cervix/19/1010_1A	1.7619	0.40816	2	5.2	0.66262	0.12621	0.77308	5.2
exa:cervix/19/1010_2B	2.3226	0.87409	3	5.4	0.54817	0.14146	0.67138	4.6
exa:cervix/19/1011*	2.6176	0.83391	4	16.4	0.63981	0.07736	0.77308	5.2
exa:cervix/19/1012	1.7	0.61333	3	17.2	0.63447	0.21149	0.77308	5.2
exa:cervix/19/1013	1.2258	0.78668	3	13.8	0.46207	0.20867	0.64851	11.6
exa:cervix/19/1014	1.6	0.6	2	5.2	0.64774	0.19432	0.77308	5.2
exa:cervix/19/1015	1.3793	0.85612	2	5	0.43209	0.2682	0.62653	5
exa:cervix/19/1016	0.22727	0.35124	1	11	0.16436	0.25402	0.7232	11
exa:cervix/19/1017	2.3889	1.3148	4	12.8	0.53129	0.23613	0.70666	5.2
exa:cervix/19/1018	1.56	0.6688	2	5.2	0.54413	0.21765	0.77308	5.2
exa:cervix/19/1020	1.4286	0.81633	2	5	0.48637	0.27793	0.77308	5
exa:cervix/19/1022*	1.2	0.672	3	17	0.54819	0.21927	0.70049	11
exa:cervix/19/1023	0.78947	0.33241	1	5	0.53626	0.22579	0.67926	5
exa:cervix/19/1024	0.3913	0.47637	1	7	0.24291	0.29572	0.62077	7
exa:cervix/19/1025*	2.5333	0.85333	4	17	0.67418	0.10269	0.77308	5.2
exa:cervix/19/1789	0	0	0	5.8	0	0	0	5.8
exa:cervix/19/1790	2.5417	0.6875	3	6.6	0.55933	0.11252	0.8669	5.2
exa:cervix/19/1792	2.0625	1.0234	4	10.4	0.56851	0.14213	0.66993	5.2
exa:cervix/19/1793	2.4	0.88	3	5	0.46004	0.09948	0.67926	3.2
exa:cervix/19/1794	2.4	0.8	3	5	0.45126	0.10782	0.63735	3.2
exa:cervix/19/1795	3.0323	0.99896	4	7.4	0.5696	0.07997	0.7	1.8
exa:cervix/19/1797*	2.718	0.49178	3	3.2	0.65854	0.06778	0.70988	8
exa:cervix/19/1798	3.3333	0.88889	4	6	0.54315	0.08466	0.7	1.8
exa:cervix/19/1799	1.4286	0.81633	2	5	0.24259	0.13862	0.33963	5
exa:cervix/19/1800	2.52	0.8224	4	13.2	0.67674	0.05414	0.74872	7.4
exa:cervix/19/1802	2.4	1.072	4	13.2	0.67908	0.05507	0.73963	7.4
exa:cervix/19/1803	2.45	0.66	3	5.6	0.65669	0.06567	0.72445	9.4
exa:cervix/19/1807	1.8823	0.63668	3	13.4	0.56505	0.06952	0.65952	3.8
exa:cervix/19/1808	2.697	0.87052	4	13.4	0.58339	0.0297	0.65952	0.8
exa:cervix/19/1809	0.25	0.375	1	10.6	0.14552	0.21828	0.58207	10.6
exa:cervix/19/1811	1.931	0.12842	2	3.2	0.62342	0.00671	0.62702	3.2
exa:cervix/19/1812*	1.9643	0.06888	2	3.2	0.62516	0.0036	0.62702	3.2
exa:cervix/19/279	2.6765	0.45675	3	5.2	0.69609	0.01507	0.70988	8
exa:cervix/19/280	2.7143	0.82653	4	13.4	0.68171	0.04869	0.74872	7.4
exa:cervix/19/281	1.9167	0.9375	3	7.8	0.59353	0.14838	0.69566	5.8
exa:cervix/19/282*	2.875	0.44792	4	13.4	0.67337	0.00764	0.7	1.8
exa:cervix/19/966	2.5641	0.64826	3	5.2	0.64825	0.09973	0.73648	10.4
exa:cervix/19/969	4.1837	1.2995	5	4.2	0.46238	0.11324	0.6345	3.8
exa:cervix/19/975	2.3514	1.5471	4	10	0.43433	0.25825	0.64034	4.8
exa:cervix/19/985	3.0465	1.2861	4	5.2	0.58903	0.16666	0.72175	10
exa:cervix/19/990	0.66667	0.66667	2	13.6	0.31334	0.31334	0.67926	5.8
exa:cervix/19/993	2.3077	1.2308	4	12	0.49766	0.19141	0.64224	12
exa:cervix/19/995	2.439	0.90303	3	3.8	0.56855	0.19414	0.70808	9.8
exa:cervix/19/996	2.5	0.725	3	4.2	0.59368	0.08905	0.71349	2.4
exa:cervix/19/997	3.4151	0.94909	4	4	0.44778	0.10138	0.68092	3.8

Table B.17. Left full table with Levenshtein Damerau distance used in merge phase with Jaro distance used in linking phase.

Prec Avg	Prec Dev	Prec Max	Prec At	Rec Avg	Rec Dev	Rec Max	Rec At	F_1 Avg	F_1 Dev	F_1 Max	F_1 At
0.04588	0.02624	0.08333	5.4	0.23673	0.14601	0.42857	10.4	0.07616	0.04405	0.12903	5.4
0.09453	0.04593	0.25	4	0.3744	0.11573	0.44444	4	0.14422	0.05933	0.32	4
0.1368	0.05319	0.25	4.2	0.62366	0.25737	0.83333	6.6	0.21942	0.08039	0.36364	4.2
0.1556	0.06123	0.66667	3.8	0.45798	0.12679	0.57143	7.6	0.20554	0.03755	0.4	3.8
0.08562	0.03507	0.21429	4	0.34756	0.14515	0.5	8.2	0.13195	0.04971	0.27273	4
0.11658	0.05125	0.28571	5.2	0.2517	0.05831	0.28571	5.2	0.15112	0.04911	0.28571	5.2
0.08058	0.03013	0.15385	4.2	0.29032	0.10926	0.375	5.4	0.12111	0.03831	0.2	5.4
0.11518	0.04221	0.33333	3.2	0.29085	0.09266	0.44444	16.4	0.14818	0.03074	0.23077	5.6
0.0801	0.0371	0.18182	5	0.24286	0.08762	0.42857	17.2	0.11562	0.0461	0.22222	5
0.04609	0.02448	0.08824	13.8	0.1362	0.08741	0.33333	13.8	0.06706	0.03526	0.13953	13.8
0.1103	0.04746	0.2	4.8	0.22857	0.08571	0.28571	5.2	0.14278	0.04957	0.23529	5.2
0.05479	0.03401	0.13333	5	0.17241	0.10702	0.25	5	0.08228	0.05107	0.17391	5
0.0095	0.01469	0.04545	11	0.02841	0.0439	0.125	11	0.01423	0.022	0.06667	11
0.08574	0.03887	0.1875	5.2	0.23889	0.13148	0.4	12.8	0.12228	0.05955	0.23077	5.2
0.06281	0.03068	0.18182	5.2	0.17333	0.07431	0.22222	5.2	0.08794	0.03699	0.2	5.2
0.04841	0.02798	0.11765	5	0.17857	0.10204	0.25	5	0.07519	0.04296	0.16	5
0.05715	0.02525	0.1	5.4	0.17143	0.096	0.42857	17	0.08382	0.03687	0.16216	17
0.07087	0.03689	0.16667	5	0.13158	0.0554	0.16667	5	0.08946	0.04063	0.16667	5
0.0165	0.02008	0.05	7	0.04891	0.05955	0.125	7	0.02462	0.02997	0.07143	7
0.13232	0.03775	0.25	3	0.31667	0.10667	0.5	17	0.17468	0.03869	0.25	5.6
0	0	0	5.8	0	0	0	5.8	0	0	0	5.8
0.16421	0.06725	0.5	5.2	0.3631	0.09821	0.42857	6.6	0.19903	0.04397	0.31579	6.6
0.08999	0.02874	0.2	3	0.34375	0.17057	0.66667	10.4	0.13366	0.04121	0.2	10.4
0.12444	0.05204	0.25	5	0.34286	0.12571	0.42857	5	0.17111	0.05768	0.31579	5
0.14317	0.06026	0.28571	3.2	0.34286	0.11429	0.42857	5	0.18834	0.05758	0.3	5
0.21457	0.10554	1	1.8	0.37903	0.12487	0.5	7.4	0.21803	0.03075	0.30769	3.2
0.1152	0.07116	0.375	3.2	0.33974	0.06147	0.375	3.2	0.15012	0.06105	0.375	3.2
0.25443	0.1321	1	1.8	0.41667	0.11111	0.5	6	0.25209	0.05256	0.36364	3.2
0.06103	0.03534	0.15385	5	0.20408	0.11662	0.28571	5	0.09259	0.05291	0.2	5
0.22853	0.10688	0.66667	3.2	0.36	0.11749	0.57143	13.2	0.24292	0.04431	0.4	3.2
0.15894	0.05383	0.5	2.6	0.34286	0.15314	0.57143	13.2	0.18833	0.03611	0.25	7.4
0.26003	0.11782	0.66667	3.2	0.35	0.09429	0.42857	5.6	0.26334	0.06209	0.4	3.2
0.07914	0.02309	0.25	3.8	0.23529	0.07958	0.375	13.4	0.10915	0.02082	0.16667	3.8
0.15387	0.07233	1	0.8	0.38528	0.12436	0.57143	13.4	0.17466	0.01898	0.25	0.8
0.01256	0.01884	0.05556	10.6	0.04167	0.0625	0.16667	10.6	0.01929	0.02893	0.08333	10.6
0.16782	0.11705	1	1.6	0.38621	0.02568	0.4	3.2	0.18674	0.07205	0.4	3.2
0.19479	0.13568	1	2.8	0.32738	0.01148	0.33333	3.2	0.19934	0.07492	0.44444	3.2
0.18593	0.11992	1	1.6	0.44608	0.07612	0.5	5.2	0.208	0.0665	0.44444	3.2
0.19087	0.10814	0.66667	3.2	0.33929	0.10332	0.5	13.4	0.20229	0.04209	0.36364	3.2
0.1082	0.03249	0.25	3.4	0.31944	0.15625	0.5	7.8	0.15078	0.04406	0.21429	7.8
0.29396	0.18924	1	1.8	0.41071	0.06399	0.57143	13.4	0.27263	0.06884	0.44444	3.2
0.11049	0.05326	0.33333	3.8	0.3663	0.09261	0.42857	5.2	0.15534	0.05325	0.30769	3.8
0.10224	0.05097	0.27778	4.2	0.46485	0.14438	0.55556	4.2	0.15889	0.06638	0.37037	4.2
0.07858	0.04759	0.13793	10	0.39189	0.25785	0.66667	10	0.12958	0.07914	0.22857	10
0.07063	0.02779	0.16667	3.4	0.38081	0.16076	0.5	5.2	0.11428	0.03832	0.21053	5.2
0.02138	0.02138	0.05882	13.6	0.09524	0.09524	0.28571	13.6	0.03477	0.03477	0.09756	13.6
0.12627	0.04866	0.22222	5	0.38462	0.20513	0.66667	12	0.18374	0.07611	0.27273	8.4
0.07488	0.03992	0.27273	3.8	0.34843	0.129	0.42857	3.8	0.1179	0.05433	0.33333	3.8
0.08416	0.03512	0.2	3.2	0.35714	0.10357	0.42857	4.2	0.12713	0.04032	0.23529	3.2
0.10644	0.05587	0.30769	4	0.42689	0.11864	0.5	4	0.16055	0.06961	0.38095	4

Table B.18. Right full table with Levenshtein Damerau distance used in merge phase with Jaro distance used in linking phase.

Id diagnosis	Count Avg	Count Dev	Count Max	Count At	Score Avg	Score Dev	Score Max	Score At
exa:cervix/19/1000*	0.88235	0.89965	3	0	0.28887	0.29454	0.60215	0.73095
exa:cervix/19/1002	2.5714	0.6898	4	0	0.64444	0.0966	0.77308	0.81838
exa:cervix/19/1004	3.1818	1.5331	5	0.70948	0.496	0.18036	0.627	0.80984
exa:cervix/19/1006	2.5556	0.61728	4	0	0.69544	0.04648	0.73667	0.82605
exa:cervix/19/1008*	2.4333	0.73222	4	0	0.67275	0.09596	0.77308	0.82621
exa:cervix/19/1010_1A	1.8387	0.27055	2	0.82872	0.78127	0.02762	0.8669	0.8975
exa:cervix/19/1010_2B	2.6275	0.55517	3	0.79577	0.71172	0.05771	0.8669	0.89905
exa:cervix/19/1011*	2.1042	0.46962	4	0.68003	0.697	0.07071	0.8669	0.90536
exa:cervix/19/1012	1.8718	0.45365	3	0.66525	0.76647	0.04865	0.8669	0.90127
exa:cervix/19/1013	0.95122	0.74242	3	0.60673	0.35953	0.28061	0.64101	0.60673
exa:cervix/19/1014	1.6296	0.46639	2	0.82872	0.80662	0.04465	0.8669	0.8975
exa:cervix/19/1015	0.83784	0.31702	2	0.6209	0.54711	0.20702	0.67926	0.80039
exa:cervix/19/1016	0.18519	0.30178	1	0.64673	0.13393	0.21825	0.7232	0.64673
exa:cervix/19/1017	2.4565	1.0974	4	0.70159	0.76498	0.0574	0.8669	0.8975
exa:cervix/19/1018	1.8095	0.31746	2	0.86444	0.68951	0.09245	0.8669	0.90127
exa:cervix/19/1020	1.75	0.40909	2	0.83704	0.68516	0.11451	0.8669	0.8975
exa:cervix/19/1022*	0.96875	0.66602	3	0.61751	0.44172	0.30368	0.6842	0.61751
exa:cervix/19/1023	0.7619	0.36281	1	0.82544	0.51363	0.24459	0.67926	0.80547
exa:cervix/19/1024	0.82759	0.28537	1	0.88333	0.51374	0.17715	0.62077	0.88333
exa:cervix/19/1025*	2.1351	0.85318	4	0.65622	0.73727	0.08765	0.8669	0.90127
exa:cervix/19/1789	0	0	0	0.85219	0	0	0	0.85219
exa:cervix/19/1790	1.9118	0.21453	3	0	0.77505	0.02161	0.8669	0.8975
exa:cervix/19/1792	2.0714	0.67687	4	0.66111	0.59613	0.08516	0.67457	0.70481
exa:cervix/19/1793	1.7805	0.40214	3	0	0.70202	0.07878	0.79408	0.87928
exa:cervix/19/1794	1.6	0.53714	3	0	0.68966	0.08844	0.79408	0.86634
exa:cervix/19/1795	2.7	0.5	4	0	0.71141	0.02102	0.74704	0.87475
exa:cervix/19/1797*	2.6833	0.50667	3	0.84185	0.703	0.04843	0.78345	0.87761
exa:cervix/19/1798	2.8293	0.33195	4	0	0.70042	0.03309	0.78345	0.88377
exa:cervix/19/1799	0.88889	0.24691	2	0	0.57367	0.17235	0.67926	0.81145
exa:cervix/19/1800	2.931	0.79905	4	0.72141	0.70366	0.05116	0.8669	0.87385
exa:cervix/19/1802	2.8125	0.90625	4	0.72141	0.67828	0.08478	0.8669	0.87385
exa:cervix/19/1803	2.52	0.7296	3	0.82872	0.71308	0.05705	0.8669	0.8765
exa:cervix/19/1807	1.766	0.73699	3	0.6854	0.59101	0.06133	0.65952	0.86285
exa:cervix/19/1808	2.4	1.2622	4	0.73016	0.59929	0.03545	0.65952	0.97441
exa:cervix/19/1809	0.6087	0.47637	1	0.77597	0.35431	0.27728	0.58207	0.77597
exa:cervix/19/1811	1.9737	0.05125	2	0.87542	0.62565	0.00268	0.62702	0.87542
exa:cervix/19/1812*	1.9118	0.1609	2	0.84882	0.63163	0.0084	0.67926	0.87542
exa:cervix/19/279	2.75	0.39583	3	0.83571	0.72362	0.01578	0.78345	0.87385
exa:cervix/19/280	2.8684	0.47784	4	0.70189	0.70503	0.03874	0.78345	0.87385
exa:cervix/19/281	1.9375	0.89063	3	0.73328	0.61504	0.11532	0.69566	0.77991
exa:cervix/19/282*	3.0312	0.54492	4	0.70189	0.69078	0.03311	0.78345	0.89251
exa:cervix/19/966	2.6731	0.55325	3	0.8263	0.68434	0.08206	0.73648	0.77222
exa:cervix/19/969	2.7407	0.65752	5	0	0.64143	0.11354	0.76356	0.82794
exa:cervix/19/975	1.9318	1.0837	4	0.65635	0.5088	0.18502	0.64122	0.70982
exa:cervix/19/985	3.1579	1.1523	4	0.80111	0.66554	0.08855	0.77308	0.81838
exa:cervix/19/990	0.89744	0.27613	2	0.64222	0.52558	0.16172	0.62683	0.64222
exa:cervix/19/993	2.0323	0.37877	4	0.63528	0.59309	0.07653	0.64214	0.75062
exa:cervix/19/995	2.5781	0.646	3	0.8166	0.66964	0.08488	0.76356	0.82639
exa:cervix/19/996	2.4677	0.77263	3	0.80984	0.61895	0.0599	0.73117	0.86919
exa:cervix/19/997	2.6197	0.62805	4	0	0.65064	0.07765	0.77308	0.81838

Table B.19. Left full table with Jaro distance used in merge phase with Jaro distance used in linking phase.

Prec Avg	Prec Dev	Prec Max	Prec At	Rec Avg	Rec Dev	Rec Max	Rec At	F_1 Avg	F_1 Dev	F_1 Max	F_1 At
0.02304	0.0235	0.0625	0.73095	0.12605	0.12852	0.42857	0	0.03879	0.03955	0.10526	0
0.08465	0.03737	0.18182	0.83215	0.28571	0.07664	0.44444	0	0.12174	0.04043	0.23077	0.81329
0.12383	0.04762	0.23529	0.77424	0.5303	0.25551	0.83333	0.70948	0.19597	0.07474	0.34783	0.77424
0.15611	0.07109	0.5	0.84094	0.36508	0.08818	0.57143	0	0.19068	0.04594	0.33333	0.82605
0.09447	0.03673	0.22222	0.82639	0.30417	0.09153	0.5	0	0.13331	0.03594	0.23529	0.82639
0.18667	0.11019	1	0.8975	0.26267	0.03865	0.28571	0.82872	0.17868	0.04535	0.30769	0.82872
0.15689	0.07981	1	0.89905	0.32843	0.0694	0.375	0.79577	0.17342	0.04134	0.27273	0.79577
0.12022	0.06284	0.5	0.90536	0.2338	0.05218	0.44444	0.68003	0.13568	0.03498	0.25	0.84796
0.12376	0.05583	0.5	0.90127	0.2674	0.06481	0.42857	0.66525	0.14655	0.03193	0.23529	0.82872
0.03323	0.02593	0.07692	0.60673	0.10569	0.08249	0.33333	0.60673	0.04986	0.03891	0.125	0.60673
0.19308	0.10853	1	0.8975	0.2328	0.06663	0.28571	0.82872	0.17067	0.03113	0.25	0.8975
0.04736	0.02666	0.14286	0.82184	0.10473	0.03963	0.25	0.6209	0.06183	0.02865	0.13333	0.82184
0.00774	0.01262	0.04545	0.64673	0.02315	0.03772	0.125	0.64673	0.0116	0.0189	0.06667	0.64673
0.15589	0.08128	1	0.8975	0.24565	0.10974	0.4	0.70159	0.14791	0.02333	0.2	0.70159
0.1307	0.07426	0.5	0.90127	0.20106	0.03527	0.22222	0.86444	0.13498	0.03909	0.25	0.86444
0.09419	0.04519	0.22222	0.83704	0.21875	0.05114	0.25	0.83704	0.12132	0.03965	0.23529	0.83704
0.04584	0.03154	0.09677	0.61751	0.13839	0.09515	0.42857	0.61751	0.06756	0.04645	0.15789	0.61751
0.07211	0.04264	0.2	0.82544	0.12698	0.06047	0.16667	0.82544	0.08814	0.04438	0.18182	0.82544
0.06052	0.03326	0.16667	0.88333	0.10345	0.03567	0.125	0.88333	0.07264	0.03108	0.14286	0.88333
0.132	0.04476	0.33333	0.90127	0.26689	0.10665	0.5	0.65622	0.15632	0.03576	0.22222	0.82872
0	0	0	0.85219	0	0	0	0.85219	0	0	0	0.85219
0.1824	0.11474	1	0.8975	0.27311	0.03065	0.42857	0	0.17823	0.05084	0.33333	0.82872
0.10528	0.03565	0.25	0.84643	0.34524	0.11281	0.66667	0.66111	0.14939	0.03518	0.25	0.81636
0.1078	0.0511	0.33333	0.87928	0.25436	0.05745	0.42857	0	0.13642	0.04034	0.25	0.83571
0.1166	0.05302	0.5	0.91833	0.22857	0.07673	0.42857	0	0.13263	0.02805	0.22222	0.91833
0.20417	0.10691	1	0.92393	0.3375	0.0625	0.5	0	0.21029	0.04652	0.33333	0.87475
0.12533	0.06586	0.33333	0.9406	0.33542	0.06333	0.375	0.84185	0.1618	0.0553	0.3	0.84185
0.22808	0.14549	1	0.90319	0.35366	0.04149	0.5	0	0.22441	0.06631	0.4	0.88377
0.05348	0.0289	0.16667	0.83789	0.12698	0.03527	0.28571	0	0.0708	0.02998	0.15385	0.83789
0.24759	0.09001	0.5	0.87385	0.41872	0.11415	0.57143	0.72141	0.27795	0.05949	0.46154	0.82872
0.19804	0.06599	0.375	0.81222	0.40179	0.12946	0.57143	0.72141	0.24323	0.05617	0.4	0.81222
0.25845	0.10868	0.5	0.8765	0.36	0.10423	0.42857	0.82872	0.26815	0.07274	0.46154	0.82872
0.08806	0.03188	0.33333	0.86285	0.22074	0.09212	0.375	0.6854	0.11089	0.0218	0.18182	0.86285
0.14415	0.07045	1	0.97441	0.34286	0.18032	0.57143	0.73016	0.16005	0.03315	0.25	0.97441
0.0423	0.03311	0.11111	0.77597	0.10145	0.0794	0.16667	0.77597	0.05875	0.04598	0.13333	0.77597
0.19774	0.1516	1	0.89327	0.39474	0.01025	0.4	0.87542	0.2109	0.09327	0.57143	0.87542
0.19027	0.12482	1	0.87542	0.31863	0.02682	0.33333	0.84882	0.19524	0.06469	0.4	0.84882
0.19413	0.11887	1	0.9406	0.45833	0.06597	0.5	0.83571	0.22297	0.07522	0.44444	0.87385
0.20408	0.10573	0.66667	0.87385	0.35855	0.05973	0.5	0.70189	0.22653	0.06051	0.375	0.84185
0.11688	0.03236	0.25	0.84971	0.32292	0.14844	0.5	0.73328	0.16099	0.04163	0.23077	0.73328
0.28621	0.14872	1	0.90319	0.43304	0.07785	0.57143	0.70189	0.28595	0.05436	0.44444	0.89251
0.1305	0.06508	0.33333	0.8263	0.38187	0.07904	0.42857	0.8263	0.17929	0.0659	0.375	0.8263
0.08024	0.03958	0.21429	0.81707	0.30453	0.07306	0.55556	0	0.11901	0.04578	0.26087	0.81707
0.07544	0.02867	0.11765	0.76678	0.32197	0.18061	0.66667	0.65635	0.11829	0.04708	0.18182	0.65635
0.08695	0.02805	0.16	0.80111	0.39474	0.14404	0.5	0.80111	0.13526	0.03739	0.24242	0.80111
0.0483	0.02492	0.14286	0.88333	0.12821	0.03945	0.28571	0.64222	0.06658	0.02846	0.14286	0.88333
0.1592	0.08265	0.5	0.8224	0.33871	0.06313	0.66667	0.63528	0.19565	0.06509	0.4	0.8224
0.09554	0.04321	0.22222	0.82639	0.3683	0.09229	0.42857	0.8166	0.14105	0.04805	0.26087	0.8166
0.09171	0.03506	0.25	0.91279	0.35253	0.11038	0.42857	0.80984	0.13401	0.03827	0.24	0.80984
0.09112	0.04197	0.2	0.86444	0.32746	0.07851	0.5	0	0.13188	0.04515	0.25	0.81329

Table B.20. Right full table with Jaro distance used in merge phase with Jaro distance used in linking phase.

Bibliography

[1] Hee J An et al. «Prevalence of human papillomavirus DNA in various histological subtypes of cervical adenocarcinoma: a population-based study». In: *Modern Pathology* 18.4 (2005), pp. 528–534.

[2] Luiz H Araujo et al. «Cancer of the lung: non–small cell lung cancer and small cell lung cancer». In: *Abeloff's clinical oncology*. Elsevier, 2020, pp. 1108–1158.

[3] Alan R Aronson and François-Michel Lang. «An overview of MetaMap: historical perspective and recent advances». In: *Journal of the American Medical Informatics Association* 17.3 (2010), pp. 229–236.

[4] DeepL. *DeepL Translate*. URL: https://www.deepl.com/.

[5] ExaMode. *ExaMode: Extreme-scale Analytics via Multimodal Ontology Discovery and Enhancement*. URL: https://www.examode.eu/.

[6] Alessio Fasano et al. «Prevalence of Celiac Disease in At-Risk and Not-At-Risk Groups in the United States: A Large Multicenter Study». In: *Archives of Internal Medicine* 163.3 (Feb. 2003), pp. 286–292. ISSN: 0003-9926. DOI: 10.1001/archinte.163.3.286. eprint: https://jamanetwork.com/journals/jamainternalmedicine/articlepdf/215079/ioi20641.pdf. URL: https://doi.org/10.1001/archinte.163.3.286.

[7] Paolo Ferragina and Ugo Scaiella. «TAGME: On-the-Fly Annotation of Short Text Fragments (by Wikipedia Entities)». In: *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. CIKM '10. Toronto, ON, Canada: Association for Computing Machinery, 2010, pp. 1625–1628. ISBN: 9781450300995. DOI: 10.1145/1871437.1871689. URL: https://doi.org/10.1145/1871437.1871689.

[8] FutureBridge. *Digital Pathology - Transforming the Future of Lab Testing*. URL: https://www.futurebridge.com/industry/perspectives-life-sciences/digital-pathology/.

[9] Google. *Google Translate*. URL: https://translate.google.com/.

[10] Cyril Goutte and Eric Gaussier. «A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation». In: *Advances in Information Retrieval*. Ed. by David E. Losada and Juan M. Fernández-Luna. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 345–359. ISBN: 978-3-540-31865-1.

[11] J.P. Gunasegaran. *Textbook of Histology and Practical guide*. Elsevier Health Sciences, 2010. ISBN: 9788131236215. URL: https://books.google.it/books?id=qATbAgAAQBAJ.

[12] IARC. *International Agency for Research on Cancer*. URL: https://www.iarc.who.int/.

[13] IEO. *Istituto Europeo di Oncologia*. URL: http://www.ieo.it/it/.

[14] National Cancer Institute. *Comprehensive Cancer Information*. URL: https://www.cancer.gov/.

[15] S Lohi et al. «Increasing prevalence of coeliac disease over time». In: *Alimentary pharmacology and therapeutics* 26.9 (2007), pp. 1217–1225.

[16] Niccolò Marini et al. «Unleashing the potential of digital pathology data by training computer-aided diagnosis models without human annotations». In: *NPJ digital medicine* 5.1 (2022), pp. 1–18.

[17] Henry M Marshall et al. «Screening for lung cancer with low-dose computed tomography: a review of current status». In: *Journal of thoracic disease* 5.Suppl 5 (2013), S524.

[18] Mark Neumann et al. «ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing». In: *Proceedings of the 18th BioNLP Workshop and Shared Task*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 319–327. DOI: 10.18653/v1/W19-5034. eprint: arXiv:1902.07669. URL: https://www.aclweb.org/anthology/W19-5034.

[19] NIH. *Unified Medical Language System (UMLS)*. URL: https://www.nlm.nih.gov/research/umls/index.html.

[20] Wei Shen, Jianyong Wang, and Jiawei Han. «Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions». In: *IEEE Transactions on Knowledge and Data Engineering* 27.2 (2015), pp. 443–460. DOI: 10.1109/TKDE.2014.2327028.

[21] American Cancer Society. *American Cancer Society | Information and Resources about for Cancer: Breast, Colon, Lung, Prostate, Skin*. URL: https://www.cancer.org/.

[22] James Turk. *Functions - jellyfish*. URL: https://jamesturk.github.io/jellyfish/functions/.

[23] James Turk. *jellyfish*. URL: https://jamesturk.github.io/jellyfish/.

[24] W3C. *W3C Semantic Web FAQ*. URL: https://www.w3.org/2001/sw/SW-FAQ.

[25] W3C. *W3C: Linked Data*. URL: https://www.w3.org/standards/semanticweb/data.

[26] W3School. *XML RDF*. URL: https://www.w3schools.com/xml/xml_rdf.asp.

[27] Lisheng Wei, Quan Gan, and Tao Ji. «Cervical cancer histology image identification method based on texture and lesion area features». In: *Computer Assisted Surgery* 22.sup1 (2017), pp. 186–199.

[28] Wikipedia. *Cosine similarity - Wikipedia*. URL: https://en.wikipedia.org/wiki/Cosine_similarity.

[29] Wikipedia. *Jaro–Winkler distance - Wikipedia*. URL: https://en.wikipedia.org/wiki/Jaro%E2%80%93Winkler_distance.

[30] Wikipedia. *Ontology - Wikipedia*. URL: https://en.wikipedia.org/wiki/Ontology.

[31] Wikipedia. *Ontology (information science) - Wikipedia*. URL: https://en.wikipedia.org/wiki/Ontology_(information_science).

[32] Wikipedia. *Semantic Web - Wikipedia*. URL: https://en.wikipedia.org/wiki/Semantic_Web.

[33] Wikipedia. *Wikipedia*. URL: https://www.wikipedia.org/.

[34] RH Young. «WHO classification of tumours of female reproductive organs». In: *Kurman RJ Carcangiu ML Herrington CS Young RH Monodermal teratomas and somatic-type tumours arising from a dermoid cyst* (2014), pp. 63–66.

[35] Mark D. Zarella et al. «A Practical Guide to Whole Slide Imaging: A White Paper From the Digital Pathology Association». In: *Archives of Pathology and Laboratory Medicine* 143.2 (Oct. 2018), pp. 222–234. ISSN: 0003-9985. DOI: 10.5858/arpa.2018-0343-RA. eprint: https://meridian.allenpress.com/aplm/article-pdf/143/2/222/1448830/arpa_2018-0343-ra.pdf. URL: https://doi.org/10.5858/arpa.2018-0343-RA.