Universidade do Minho
Escola de Engenharia

Andreia Dóris Pedras Rodrigues

Extracting knowledge from documents related with *Candida* invasive fungal infections in iron overload context

Janeiro de 2021

Universidade do Minho
Escola de Engenharia

Andreia Dóris Pedras Rodrigues

Extracting knowledge from documents related with *Candida* invasive fungal infections in iron overload context

Dissertação de Mestrado
Mestrado em Bioinformática

Trabalho realizado sob a orientação dos
Professor Doutor Miguel Rocha
Doutora Catarina Pimentel

Janeiro de 2021

Acknowledgements

Gostaria de agradecer aos meus orientadores, Prof. Miguel Rocha e Dr.ª Catarina Pimentel, por toda a ajuda durante este último ano. Aprendi (e ainda estou a aprender) muito com eles.

Também agradeço muito ao Ruben Rodrigues, por ter estado sempre disponível para me esclarecer todas as dúvidas que me surgiram. A ajuda dele foi indispensável para a realização deste trabalho.

Dedico este trabalho aos meus pais e ao meu companheiro, por todo o apoio que sempre me deram e por nunca deixarem de acreditar em mim.

Statement of Integrity

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

Abstract

Invasive fungal infections caused by *Candida* are associated with high mortality and morbidity rates in hospitalized patients. Iron plays a major role in these infections, as they are exacerbated under iron overload conditions. In this context, it is important to understand the association between iron levels and invasive fungal infections, as it can serve as an indicator of the severity of the disease, and eventually it can help establish measures to improve treatment efficacy.

Nowadays, manually inferring these associations from biomedical documents is a time-consuming task, due to the high amount of available scientific text data. As such, these tasks naturally benefit from the Biomedical Text Mining field, which includes a wide variety of methods for automatic extraction of high-quality information from biomedical text documents.

In this work, relevant documents related to iron overload and fungal infections were retrieved from PubMed to build a corpus. Then, both Named Entity Recognition and Relation Extraction processes were executed using the @Note text mining tool. Finally, relevant sentences were manually extracted and a curated dataset with documents containing those sentences was created.

Since the number of publications obtained about *Candida* and iron overload was very low, the analysis was made taking into account all fungi. A total of 15 publications were considered relevant and 168 relevant associations were extracted.

Although associations of iron levels with both severity of infection and treatment efficacy were not extracted, it was possible to conclude that, in many cases, iron overload is a predictor for fungal infections, and patients' iron levels highly affect treatment efficacy.

The Biomedical Text Mining process described in the present thesis enabled the creation of a dataset of relevant biomedical publications containing interesting associations between fungal infections, drugs and associated diseases in a clinical context of iron overload, although in the future this process could be improved, especially regarding dictionaries, in order to obtain a higher number of relevant publications.

Keywords: Biomedical text mining; Invasive fungal infections; Iron overload.

Resumo

As infeções fúngicas invasivas causadas por *Candida* estão associadas a elevadas taxas de mortalidade e morbilidade em doentes hospitalizados. O ferro tem um papel importante neste tipo de infeções, visto que estas são exacerbadas em condições de excesso de ferro. Neste contexto, é extremamente importante compreender a associação entre os níveis de ferro e infeções fúngicas invasivas, pois pode servir como indicador da severidade da doença e, eventualmente, ajudar a estabelecer medidas para melhorar a eficácia de tratamento.

Atualmente, inferir manualmente este tipo de associações de documentos biomédicos revela-se uma tarefa bastante demorada, devido ao elevado volume de dados de texto científico disponíveis. Como tal, estas tarefas beneficiam claramente da área da mineração de textos biomédicos, que inclui uma ampla variedade de métodos para extração de informação de alta qualidade de documentos de texto biomédicos.

No presente trabalho, foram identificados, inicialmente, documentos relevantes que associam o ferro com infeções fúngicas invasivas para construir um *corpus*. De seguida, os processos de *Reconhecimento de entidades nomeadas* e *Extração de relações* foram realizados usando a ferramenta de mineração de textos @Note. Finalmente, as frases mais relevantes foram extraídas e foi criado um *corpus* curado de documentos contendo essas mesmas frases.

Visto que o número de publicações obtidas relacionadas com *Candida* e excesso de ferro foi muito baixo, a análise foi feita tendo em conta todos os fungos. Um total de 15 publicações foram consideradas relevantes e 168 associações foram extraídas.

Embora não tivesse sido possível extrair associações entre níveis de ferro e a eficácia do tratamento/severidade da infeção, foi possível concluir que o excesso de ferro prevê o surgimento de infeções fúngicas em muitos casos, e que os níveis de ferro dos pacientes afetam fortemente a eficácia do tratamento.

O processo de mineração de textos biomédicos no presente trabalho possibilitou a criação de um corpus de publicações biomédicas relevantes contendo associações interessantes entre infeções fúngicas, fármacos e doenças associadas, no contexto clínico de excesso de ferro, embora este processo pudesse ser melhorado no futuro, especialmente no que diz respeito aos dicionários, para que seja possível a obtenção de um maior número de publicações relevantes.

Palavras-chave: Mineração de textos biomédicos; Infeções fúngicas invasivas; Excesso de ferro.

Contents

List of Figures

List of Tables

Acronyms

BioTM – Biomedical Text Mining

BSI - Bloodstream Infection

Fe – Iron

ICU – Intensive Care Unit

IE – Information Extraction

IFI – Invasive Fungal Infection

IR – Information Retrieval

MeSH – Medical Subject Headings

ML – Machine Learning

NER – Named Entity Recognition

NLP – Natural Language Processing

POS – Part-of-speech

RE – Relation Extraction

1. Introduction

1.1. Context and motivation

Invasive fungal infections (IFIs) caused by the yeast *Candida*, also named invasive candidiasis, represent the most common fungal disease among hospitalized patients receiving immunosuppressive or intensive antibacterial therapies, and are associated with high morbidity and mortality rates. Among hospital-acquired (nosocomial) infections, *Candida* species represent the tenth most frequent pathogen in Europe, while in the United States, invasive candidiasis is the fourth leading cause of nosocomial bloodstream infections [1, 2]. Iron has been reported to play a major role in fungal infections, as it is an essential element for all fungal pathogens and the infection is exacerbated under iron overload conditions [3, 4]. Thus, patients in a medical context of iron-loading might eventually experience disease relapse, which has significant clinical implications. Therefore, it is of utmost importance to better understand the associations between iron levels in patients with IFIs and both treatment efficacy and severity of infection.

However, manually inferring these associations from biomedical documents is nearly impossible, since there are tens of thousands of text documents related to fungal infections in PubMed. Therefore, tasks such as correlating data or extracting relationships from text documents benefit from Text Mining automatic methods, which highly increase their speed and efficiency. Biomedical Text Mining (BioTM), a Text Mining subfield, has gained significant attention in the scientific community in recent years, since it helps researchers to deal with large amounts of textual data through the creation of tools for extraction of high-quality information in an automated and efficient way [5].

The task of recognizing and extracting biomedical terms (bio-entities) from biomedical documents is known as Named Entity Recognition (NER), which includes different approaches such as rule-based, dictionary-based tagging and Machine Learning (ML) based approaches [6-8]. Relation Extraction (RE) is a task that aims to extract bio-events, i.e. relations between bio-entities, and includes, among others, co-occurrence-based, syntactic-based and ML-based approaches [9-11].

1.2. Objectives

The present work aims to generate and extract relevant knowledge from a set of textual documents that can show associations between iron levels in patients with IFIs and treatment efficacy or severity of infection.

In detail, the main scientific and technological objectives are:

- review the state of the art regarding IFIs and their associations with iron, as well as relevant BioTM topics and software tools;
- explore available BioTM tools in the host group which may be used in the project;
- identify documents related with IFIs caused by *Candida*, or fungi in general, in the context of iron overload, creating a corpus that will be the core of the project;
- annotate the documents using available NER and co-occurrence based RE tools;
- extract from the corpus relevant sentences containing the most relevant annotations and curate these results to create a curated dataset.

1.3. Thesis structure

This thesis is divided into 5 chapters, which comprises the present introduction, followed by the state of the art, methods, results and discussion, and conclusions and future work.

The State-of-the-Art chapter includes two main topics: Invasive Fungal Infections and Biomedical Text Mining. The former consists of an overview of Invasive Fungal Infections caused by *Candida* and their relationship with iron, while in the latter, methods and tools of Information Retrieval and Information Extraction are presented. Named Entity Recognition and Relation Extraction approaches are then described in more detail for application in the present work.

In the Methods chapter, the BioTM pipeline used to obtain a relevant set of publications is described, as well as the tools used throughout this work.

The Results and Discussion chapter includes the outcomes regarding the publications retrieved from PubMed, the annotation process, the extraction of relevant associations and the relevant documents obtained, as well as their discussion.

The final chapter comprises conclusions about the BioTM process used and the publications obtained in this work, as well as future work considerations.

2. State of the Art

2.1. Invasive Fungal Infections

Fungal infections can be caused by primary or opportunistic pathogens. Primary pathogens are naturally able to establish an infection in healthy hosts and are often associated with superficial and cutaneous infections affecting skin, hair and nails. Differently, opportunistic pathogens do not usually infect healthy hosts. Instead, they take advantage of certain situations in the host, such as a weakened immune system, an altered microbiota, or a break in protective barriers, causing mucosal infections and invasive fungal infections (IFIs) [12]. In IFIs, fungal pathogens reach the host bloodstream and can colonize any major organ of the human body. These infections can cause symptoms ranging from a simple fever to a septic shock and are associated with an elevated mortality rate (at least 1.5 million deaths worldwide each year) [12].

Invasive candidiasis is an IFI caused by opportunistic pathogens belonging to the genus *Candida* and is the most common fungal disease among hospitalized patients receiving immunosuppressive or intensive antibacterial therapies. In Europe, *Candida* species occupy a noticeable tenth position in the rank of the most frequent pathogens causing nosocomial infections [1]. In the United States, concerning hospital acquired bloodstream infections (BSIs), *Candida* species rank in the fourth position, being the deadliest pathogens of all [2].

Nosocomial BSIs caused by *Candida*, also known as candidemia, are the most common form of invasive candidiasis and a serious problem in Intensive Care Units (ICUs) [13]. It was shown, in a recent study involving several countries and patients with nosocomial BSIs admitted to ICU, that a significant percentage of those infections were of fungal origin, with *Candida albicans* being the most frequent species followed by *Candida glabrata* and *Candida parapsilosis* [14].

Candidemia is linked to both high mortality rates and hospitalization costs [15]. Additionally, there is an alarming proportion of patients with candidemia who receive inadequate antifungal therapy. Studies have been reporting that some patients receive incomplete therapy or no therapy at all [15], and, in other cases, the administration of antifungal therapy is delayed due to a failure in recognizing at-risk patients, which further increases the length of hospital stay as well as hospitalization costs [16].

Although much research has been made in order to develop new therapeutic agents to treat invasive candidiasis, currently available drugs belong to four molecule classes: fluoropyrimidines, polyenes, azoles, and echinocandins, with the last ones being the most recent and the preferred drugs for the treatment of invasive candidiasis [17].

Iron (Fe) is an essential nutrient for several cellular metabolic processes and its presence can play an important role in the human immune response. During infection, mechanisms of defence are activated in the host to prevent pathogens from accessing essential nutrients, such as Fe. Host responses for Fe-withhold, such as mechanisms for controlling host Fe metabolism that decrease free ionic Fe levels in tissue fluids, can decrease microbes' pathogenicity [3].

One of those mechanisms involves the expression Fe-binding proteins, such as transferrin and lactoferrin, which maintain a low-Fe environment by binding to Fe and decreasing its availability to extracellular microbes [4]. These proteins have high affinity for ferric iron, and are only 30–40% saturated under normal conditions, whereas in patients with Fe-overload complications, the level of saturation of serum transferrin with iron is abnormally high [4]. In addition, free iron can also be stored intracellularly in ferritin, a multimeric protein that is able to convert ferrous Fe into ferric Fe through its ferroxidase activity. This mechanism of Fe storage by ferritin also confers host protection against infections [3].

Contrarily, the shutdown of those defence mechanisms caused by freely available Fe can lead to rapid fungal growth infections in tissue fluids, which exacerbates the infection [4]. Accordingly, Fe-overload conditions are associated with poor clinical outcomes in many infectious diseases [3].

Several health conditions contribute to an increase in the levels of Fe during infection, including cancer, haemochromatosis and hepatic disease [4]. A study has shown that patients suffering from acute leukaemia, which is often accompanied by iron overload, were susceptible to infection caused by *Candida albicans* [19]. Additionally, it has been reported that two high-affinity iron permease genes are essential virulence factors in *C. albicans* [20]. Together, these findings support the fact that availability of iron plays an important role in fungal infections. Therefore, understanding the effect of iron levels in patients with IFIs is critical, since it has important implications in disease severity and may affect treatment efficacy.

Manually inferring these associations from biomedical documents, such as publications, dissertations and clinical trials, can be a time-consuming task due to the high amount of available data. For instance, a recent PubMed search for abstracts that contain the words "fungal infection" returned a dataset of more than 140 thousand documents, and a more narrowed search for abstracts mentioning "iron overload" returned more than 18 thousand documents. Any attempts to correlate data or extract relationships across documents are almost impossible to achieve due to the large number of resulting documents. As such, those tasks naturally benefit from Biomedical Text Mining (BioTM) automatic methods, which increase their speed and efficiency.

2.2. Biomedical Text Mining

The amount of textual data that is created every day is continuously growing, and the vast majority of it is written in natural language, an unstructured format which computers cannot simply process and understand as humans can [21]. As a result, the recent field of Text Mining has emerged, concerning automated processing and analysis of text, which allows the extraction of meaningful information from free text.

In the biomedical fields, specialized literature is growing at an increasing rate, with PubMed database, for instance, having around 40 thousand new records added each month [22]. This text information overload creates a challenge for researchers to retrieve relevant publications from literature databases and extract relevant information from those publications. Thus, BioTM, a Text Mining subfield, has gained significant attention in the scientific community, since it allows researchers to deal with large amounts of text data by applying tools for extraction of high-quality information in an automated and efficient way [5].

Text Mining includes a variety of methods and algorithms for text analysis that are often related to distinct fields, such as Natural Language Processing (NLP), data mining, statistics and ML. A typical BioTM analysis involves many distinct steps (Figure 1), which are described in more detail in the following sections.

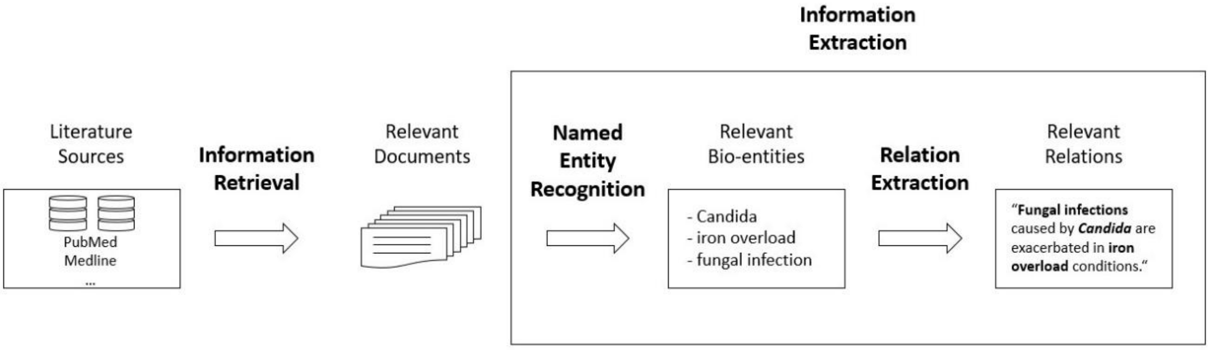


Figure 1 – Overview of a typical BioTM workflow. Firstly, Information Retrieval is used to get relevant documents on a given subject of interest from literature sources. The documents are then used for Named Entity Recognition. Lastly, Relation Extraction task detects relationships between relevant bio-entities (adapted from [23]).

2.2.1. Information Retrieval

The first step in BioTM is Information Retrieval (IR), which concerns the retrieval of relevant documents from a collection of datasets, usually a database, on a given subject of interest [23]. IR mostly focuses on making text information easily accessible for the user rather than analysing it and does not involve text processing or transformation. This process is often done by querying a set of keywords in a database, such as MEDLINE [23]. MEDLINE is a bibliographic database that contains more than 25 million references to journal articles in the life sciences. A distinctive feature of MEDLINE is that its records are indexed with National Library of Medicine's Medical Subject Headings (MeSH) [24]. MeSH terms consist of a comprehensive controlled vocabulary with the purpose of facilitating searching at various levels of specificity.

As the majority of text mining processes aims at discovering patterns across large document collections, an important element in any text mining process is the creation of a document collection, often named as a corpus, which consists of a group of text documents and can range from thousands to millions of documents [22]. PubMed, which gives free access to more than 30 million scientific literature citations from the MEDLINE database, life science journals, and online books [25], is an example of a large document collection containing abstracts in text format for biomedical literature [22].

PubMed represents the most comprehensive online collection of scientific abstracts, being the most used repository by biomedical researchers [23].

Besides scientific literature, other relevant text resources for biomedical research include patents, clinical trial records, medical records, biomedical related blogs and websites [26-29]. There are several available BioTM tools for IR in the biomedical domain (Table 1). Notably, several IR applications are built on the PubMed repository, mainly because it is open access and provides annotated abstracts with MeSH terms [23].

Table 1 – Overview of different IR tools for the biomedical domain.

Name	**Description**
askMEDLINE [30]	Natural language search tool for MEDLINE/PubMed citations
HelioBLAST [31]	Similarity engine that retrieves MEDLINE text records and ranks them according to their similarity to the submitted query
Medline Ranker [32]	Search tool that ranks MEDLINE abstracts based on a submitted biomedical topic
MiSearch [33]	Search tool that ranks retrieved citations from PubMed
GeneView [34]	Search engine built upon a comprehensively annotated version of PubMed abstracts and PubMed Central free full texts, which enables, for instance, searching for entities using unique database identifiers or ranking documents by the number of specific mentions they contain
PICO [35]	Search tool for MEDLINE/PubMed clinical trials
PubCrawler [36]	Lists new daily entries for MEDLINE/PubMed and GenBank that match specific search parameters
PubFocus [37]	Analyses MEDLINE/PubMed search queries and provides statistical information on publication trends, publishing journals and most prolific authors
PubMatrix [38]	Multiplex comparison tool that allows literature mining of PubMed using any two lists of terms, resulting in a frequency matrix of document hits based on term co-occurrence

PubNet [39]	Maps publications into networks according to how related they are to each other based on at least one PubMed query, allowing for graphical visualization, textual navigation and topological analysis
@Note [40]	BioTM platform with a wide set of methodologies for IR and IE from biomedical literature.

After the IR process, the resulting corpus can be analysed by search algorithms to extract relevant information such as occurrence of specific keywords of interest and relations between them.

2.2.2. Information Extraction

Information Extraction (IE) is the task of automatically extracting structured information or facts from unstructured or semi-structured texts [41, 42], being one of the main steps in text mining.

In the biomedical domain, most of the biomedical literature and clinical information is written in an unstructured text format. As such, IE is usually considered as a pre-processing step for many other BioTM tasks such as question answering, knowledge extraction, hypothesis generation and summarization [43-46].

Usually, document collections need to be transformed into a structured format to be analysed by IE systems. Since information in biomedical literature is written in natural language documents, which is an unstructured format, pre-processing techniques are required to transform it into a structured machine-readable format. To achieve this, NLP methodologies concerning the analysis and representation of natural texts are used in BioTM processes [22].

NLP methodologies include sentence splitting (splitting a raw text into sentences by detecting punctuation, word capitalization and breaks in the text), tokenization (splitting a raw text into tokens, i.e., alphanumeric words), filtering (removing words that appear frequently, with no significant relevance, such as stop words), lemmatization (extracting the lemma/canonical form of each word in the raw text), stemming (similar to lemmatization but ignoring the context of the word in the text, performing only a removal of commoner morphological and inflectional endings from words), part-of-speech (POS) tagging (categorizing each token by their word class, such as noun, verb, adjective, etc.), chunking (generating tree structures according to the sentence token positions that are named with chunk tags, where each tag represents the noun phrase to which they belong) and dependency parsing (identifying grammatical relationships, such as the relation between subject/indirect object tokens and verb tokens) [21].

Named Entity Recognition

In the BioTM domain, NER is an IE task that is used to identify biomedical entities in a corpus and classify them into different categories, such as proteins, genes or diseases [47]. Over the years, many biomedical entity and event extraction approaches have been proposed [48-50]. Nevertheless, the automatic extraction of biomedical entities with high accuracy from natural text remains a challenging task. The large number of related entities due to the progress in biomedical literature poses a challenge to NER systems as they are usually dependent on dictionaries of biological terms, which are often incomplete due to the continuous increase on the number of biomedical terms. Another challenge for NER in biomedical literature is the fact that a given biomedical concept may have more than one synonym, which makes NER systems not being able to recognise the same concept when represented in different forms. The use of acronyms and abbreviations is also an issue to NER systems which often fail to correctly identify biomedical concepts expressed in those forms [21].

Considering the above-mentioned challenges, it is crucial that NER systems have a high performance when analysing large amounts of text. To evaluate NER systems' performance, metrics such as precision, recall and F-score are frequently used.

NER techniques can be divided into three main types of approaches: rule-based, dictionary-based and ML-based approaches.

In rule-based approaches, regular expression patterns are used to identify bio-entities in text. The patterns must be adapted according to the biomedical context [6].

Dictionary-based tagging approaches involve the use of curated dictionaries containing biomedical terms. The dictionary terms are matched to the text and each matched token is annotated with the biological class according to the dictionary term. Multiple dictionaries can be used simultaneously to allow a wider scope. The use of this type of approaches may eventually lead to an ambiguity issue caused by the presence of different words with the same spelling but with distinct meanings, which can only be overcome with manual curation [7].

ML-based approaches for NER can be supervised or semi-supervised, and involve the training of mathematical models on an annotated corpus that is representative of the real-life dataset, which are then used for bio-entity prediction in an unannotated dataset [8]. Some examples of ML approaches for NER in biomedical texts include chemical entities identification, organism name identification, and extraction of cancer stage information from health records [51-53].

ML approaches include classification-based and sequence-based methods. In classification methods, NER is viewed as a classification problem, and classifiers such as Naive Bayes and Support Vector Machines are widely used [54, 55]. On the other hand, in sequence methods, the complete sequence of words is used to predict the most likely tag for a sequence of words, and Hidden Markov Models, Maximum Entropy Markov Models and Conditional Random Fields are the most commonly used approaches [56-58].

There are several available tools for NER tasks. Table 2 provides an overview of different NER systems, as well as their main functionalities.

Table 2 – Overview of different NER tools for the biomedical domain.

Name	**Description**
GeneValorization [59]	Provides bibliographic overviews in the form of a matrix containing the number of publications with co-occurrences of gene names and keywords defining a context of study
Anne O'Tate [60]	Gives an overview of the set of articles retrieved by a PubMed query and displays them according to various categories such as important words, phrases or MeSH pairs found in titles or abstracts
MEDIE [61]	Semantic search tool that retrieves biomedical correlations from MEDLINE
PubReMiner [62]	Displays the results of a PubMed query into frequency tables, which can be added/excluded from the query to optimize the results
Reflect [63]	Tags gene, protein, and small molecules in any web page by querying a URL

Whatizit [64]	Text processing system that identifies molecular biology terms and links them to publicly available databases
LINNAEUS [65]	An open source, stand-alone software system for recognition and normalization of species mentions, using a dictionary-based approach
Neji [66]	An open source framework optimized for biomedical NER using dictionary, rule and ML-based methods, with integrated NLP modules
GNAT [67]	A library and web service for gene NER and normalization in biomedical articles; mentions of genes and proteins in the articles are linked to Entrez Gene identifiers
ABNER [68]	An open source software tool for molecular biology NER, using a ML system for automatic tagging of genes, proteins and other bio-entities
Génie [69]	Takes a biological topic related to a gene function as input, evaluates MEDLINE for relevance to that subject, and then ranks all the genes of a requested organism according to the relevance of their MEDLINE records.
BeFree [70]	Text mining tool that contains a module for NER based on dictionary methods to find and uniquely identify bio-entity mentions in the literature

Relation Extraction

In the biomedical domain, RE concerns the recognition and extraction of bio-events and relationships among biomedical entities, in text documents. For instance, the identification of the verb "enhance" plus a gene entity allows the extraction of a gene expression enhancement event, while gene–disease associations and interactions between proteins are other examples of biomedical relationships. The associations are often between two entities, although they can include more than two entities. The vast majority of existing RE methods focus on extracting events and relationships within a sentence, although there are cases where a given relationship may span across more than one sentence [21]. RE approaches include co-occurrence-based, syntactic-based and ML-based methods [71]. Generally, ensembles of different techniques are much more effective than using just a single technique for RE tasks [72].

In syntactic-based approaches, relationships between entities are identified by applying syntactic rules to each sentence in the text [10]. These methods are phrase based and are able to detect triples in text. In contrast to co-occurrence methods, they provide information about the type of relationship between two entities and usually have a higher precision than co-occurrence methods [23].

ML-based approaches are frequently used for RE. Like ML methods for NER, they also require the creation of high-quality annotated data for training and assessing the performance of RE systems [21]. Common ML approaches include feature-based and kernel-based methods [11].

In feature-based methods, for each pair of entity mentions, a set of features is generated and a classifier (or an ensemble of classifiers) is then trained to classify any new relation instance. Some important features are word-based features, phrase chunking-based features and semantic-based features [11].

In kernel-based methods, kernel functions are created to compute similarities between representations of two relation instances. Several kernel-based RE systems have different representations for relation instances, such as sequences, syntactic trees, dependency trees, dependency graph paths, and composite kernels that combine individual kernels [11]. Several tools for distinct RE tasks are available with various functionalities (Table 3).

Table 3 – Overview of different RE tools for the biomedical domain.

Name	**Description**
Quertle [73]	Allows a semantic search in multiple biomedical databases and runs a query via relationships between concepts, enabling retrieval of more pertinent results and navigation by key concepts
Chilibot [74]	Identifies relationships between genes, proteins or any keywords queried by the user by mining PubMed, and displays them as a graph
Coremine Medical™ [75]	Presents results about health, medicine and biology in a dashboard format comprised of panels containing various categories of information ranging from introductory sources to the latest scientific articles
FACTA+ [76]	Text search engine that finds and allows visualization of indirect associations between biomedical concepts from MEDLINE abstracts

STRING [77]	Database of known and predicted protein-protein interactions by collecting, scoring and integrating all publicly available sources of protein-protein interaction information
PPInterFinder [78]	Extracts human protein–protein interactions from MEDLINE abstracts using relation keyword co-occurrences with protein names
CoCiter [79]	Analyses gene–gene, gene–term or term–term associations by evaluating the co-citation significance of a gene set with any other user defined gene sets, or to any free terms, and by accessing NCBI Gene and MEDLINE
LAITOR [80]	Text mining software that finds co-occurrence of biological entities (gene/protein terms) together with bio-interactions and concept terms from customized dictionaries
BeFree [81]	Identifies relationships between bio-entities by their co-occurrence in sentences, which then are processed by a RE module based on Support Vector Machines to predict the correct associations
EventMine [82]	A ML-based system that extracts events from documents that already contain named entity annotations, such as genes and proteins
@Note [40]	Uses co-occurrence and semantic rules to extract relationships between bio-entities

2.2.3. Biomedical Text Mining Applications

The vast majority of BioTM approaches to date has focused on extracting information regarding molecular processes and diseases, such as protein–protein interactions [83, 84], gene–gene relationships [85], phosphorylation events [86], metabolic and signalling pathways [87, 88], protein–compound interactions [89], gene–disease associations [90, 91], gene–protein interactions [92], associations between genetic markers and diseases [93], and drug-related knowledge that includes the discovery of novel drug targets, drug-–side effects, drug–drug interactions, drug–disease and drug–indications interactions [94]. A great amount of BioTM approaches concerns the molecular oncology area, although other areas include cardiovascular diseases, synapse biology and brain disease-associated genes, retinal diseases, and asthma candidate genes, among others [95].

While most existing BioTM approaches aim to extract information related to biological processes from scientific publications, some efforts have also been made in extracting relevant medical information from clinical records, such as temporal relations, status, assertion/risk and co-morbidities [96]. Many clinically-oriented challenge tasks have been introduced in the last decade, which consisted on, for instance, recognizing clinical concepts such as medical problems, tests, treatments, medication and dosage [97, 98], detecting temporal elations [99], determining smoking status [100], predicting obesity and its co-morbidities [101], and predicting heart disease risks [102].

Some BioTM tools have also been developed with the purpose of extraction of clinical information [103]. Table 4 provides an overview of the most widely used tools in the clinical context.

Regardless of the work done in BioTM in both biological and clinical contexts, further study regarding medical entities recognition is required as this is essential to link molecular and medical observations, and thus improving the association between laboratory research and clinical applications described in the literature.

Table 4 – Overview of different BioTM tools for the clinical context.

Name	**Description**
cTakes [104]	Open-source NLP system based on UIMA framework for extraction of information from electronic health records unstructured clinical text
MetaMap [105]	National Institutes of Health (NIH)-developed NLP tool that maps biomedical text to UMLS concepts
MedLEE [106]	NLP system that extracts, structures, and encodes clinical information from narrative clinical notes
KMCI [107]	NLP system that identifies biomedical concepts and maps them to UMLS concepts
HITEx [108]	Open-source NLP tool based on the GATE framework for various tasks such as principal diagnoses extraction and smoking status extraction

MedEx [109]	NLP tool used to recognize drug names, dose, route, and frequency from free-text clinical records
MedXN [110]	A tool to extract comprehensive medication information from clinical narratives and normalize it to RxNorm
MedTime [111]	A tool to extract temporal information from clinical narratives and normalize it to the TIMEX3 standard
MedTagger [112]	Open-source NLP pipeline based on UIMA framework for indexing based on dictionaries, information extraction, and machine learning–based named entity recognition from clinical text

3. Methods

3.1. Pipeline

In order to extract relevant associations between iron overload and IFIs from biomedical literature, a BioTM pipeline was created. The essential steps involved in this pipeline are depicted in Figure 2.

In the IR step, documents related with *Candida* in the context of iron overload were retrieved from PubMed to create a corpus that will be the core of the project. Initially, a PubMed search with the keyword "candida" was done, to obtain all publications related to *Candida*. However, the number of publications related to both *Candida* and iron overload was expected to be very low, since a PubMed search was previously done using the keywords "candida" and "iron overload", which only returned 17 publications. Therefore, a second analysis was undertaken, including every fungal organism instead of being limited to *Candida* species, to create a second corpus using the term "iron overload" to query PubMed publications related to iron overload.

Next, in the NER step, the abstracts of the publications from both corpora were annotated using different dictionaries. In a first phase, the goal was to obtain every publication mentioning iron overload and *Candida*, or iron overload and any fungal organism. To achieve that, four dictionaries were used to generate the annotations: one to identify organisms in general ("all-organisms" dictionary), two to identify fungal organisms ("only-fungi" and "mycobank" dictionaries) and one to identify iron terms ("iron-terms" dictionary).

The "all-organisms" dictionary was created based on the file available on NCBI Taxonomy [113] containing over 2 million entries for taxa names. The "only-fungi" dictionary contains all NCBI taxonomy fungi taxa names for the fungi subtree, obtained upon searching for fungi on NCBI Taxonomy [114]. Since the NCBI taxonomy database is not an authoritative source for nomenclature or classification, as stated in the disclaimer on its web page, another dictionary, was used to annotate terms related to fungal organisms. The "mycobank" dictionary is based on the MycoBank database [115] which contains the nomenclature for all fungal taxa and includes over 500 thousand entries. The "iron-terms" dictionary is a small dictionary with only nine entries that includes two classes, one for iron and other for iron overload terms, based on MeSH terms for iron and iron overload [116].

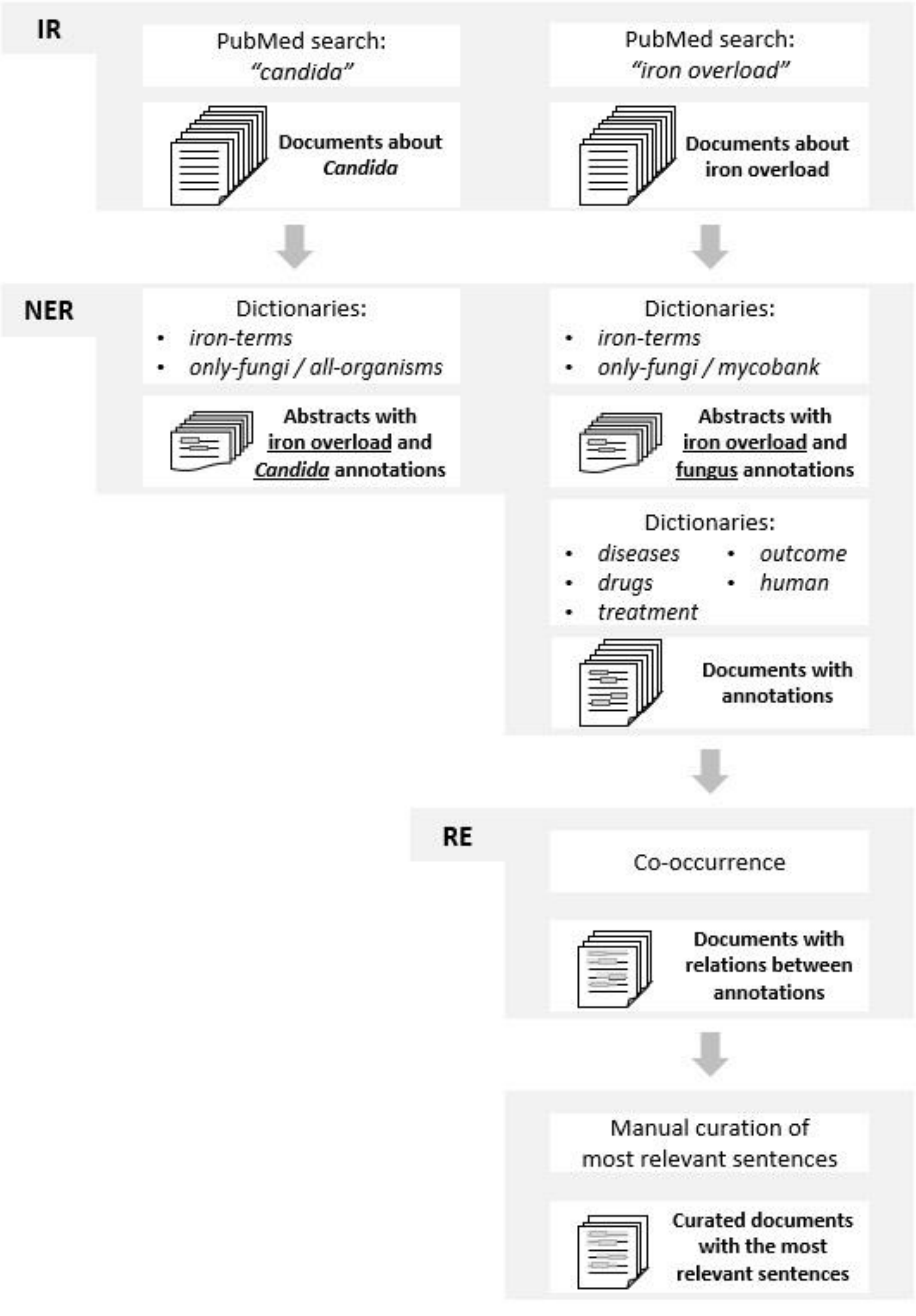

IR
PubMed search: "candida"
PubMed search: "iron overload"
Documents about Candida
Documents about iron overload
NER
Dictionaries: • iron-terms • only-fungi / all-organisms
Dictionaries: • iron-terms • only-fungi / mycobank
Abstracts with iron overload and Candida annotations
Abstracts with iron overload and fungus annotations
Dictionaries: • diseases • drugs • treatment • outcome • human
Documents with annotations
RE
Co-occurrence
Documents with relations between annotations
Manual curation of most relevant sentences
Curated documents with the most relevant sentences

Figure 2 – Overview of the pipeline used in the present work. Firstly, all publications on *Candida* and iron overload are retrieved from PubMed. Both corpora are then annotated by NER using dictionaries to annotate iron overload and *Candida*/fungi bio-entities. Only the documents containing annotations for both iron overload and fungi are selected for a second annotation step, this time using dictionaries of drugs and diseases. Lastly, RE is used to detect relationships between annotated bio-entities, and a dataset of documents containing the most relevant associations related to fungi, iron overload, and drugs or diseases is obtained.

The *Candida* corpus was annotated with two different organism related dictionaries: the "all-organisms" and the "only-fungi" dictionaries. The iron overload corpus was annotated with the "mycobank" dictionary and with the "only-fungi" dictionary. In addition, both corpora were annotated with the "iron-terms" dictionary. Next, four sub-corpora were created based on the previous annotated corpus: from the *Candida* corpus, two sub-corpora resulted – a corpus with all NCBI organisms + iron terms annotations, and a corpus with only fungal organisms + iron terms annotations; from the iron overload corpus, two sub-corpora resulted – a corpus with fugal organisms + iron terms annotations, and a corpus with mycobank fungal organisms + iron terms annotations. Since the *Candida* sub-corpora resulted in a small number of publications that can easily be manually curated, only the iron overload sub-corpora were analysed and curated using the next steps of the BioTM pipeline.

In the second phase of NER analysis, full texts were used, when available, instead of the publication abstracts. To extract bio-entities of interest from iron overload sub-corpora, lexical resources for drugs, diseases, treatment, outcome and human terms were used, in addition to the "iron-terms", "only-fungi" and "all-organisms" dictionaries. The "drugs" dictionary was created based on the DrugBank vocabulary dataset available on DrugBank [117] which includes DrugBank identifiers, names, and synonyms for all drugs. The "diseases" ontology was based on Disease Ontology file available on Disease Ontology's GitHub repository [118], which is a standardized ontology that provides medical vocabulary for human disease concepts [119]. The "treatment" and "outcome" lookup tables were based on the synonyms for the "therapeutics" and "treatment outcome" MeSH categories, respectively [116]. Finally, the "human" lookup table contains synonyms for the word "human".

After the annotation process of the iron overload sub-corpora, relevant associations between annotated bio-entities were extracted. A RE method based on co-occurrence was used, in which every pair of annotated bio-entities that occurred in the same sentence were annotated as a possible relationship.

In the manual curation step, the relevance of each extracted relationship was analysed, to select the most interesting sentences with evidence regarding the scope of the present work. Relevant relationships include associations between iron levels of a patient, the severity of infection and treatment efficacy of a fungal infection. In addition to those, the following associations were also investigated: associations between iron and fungus/fungal infection, in order to find which fungal organisms may cause IFIs in the context of iron overload; associations between iron and disease, in order to find which diseases, associated with iron overload, could be risk factors for IFIs; and associations between fungus/fungal infection and drug, in order to assess existent drugs to treat a given fungal infection.

Finally, a curated dataset of publications containing the aforementioned associations was obtained. In this step, only documents containing relevant relationships were considered, and of those, only publications referring to clinical cases were considered relevant.

3.2. Tools

3.2.1. @Note

The tool used for the BioTM process in the present work was @Note [120], which is a BioTM platform with a wide set of methodologies for both Information Retrieval and Information Extraction tasks. This tool provides an end-user application, which includes interfaces for IR and IE tasks, and for the creation of lexical resources such as dictionaries, ontologies and lookup tables. The IR interface allows for querying specific databases, such as PubMed, managing and updating query results, downloading relevant documents, converting PDF to text, assigning relevance and managing corpora. The IE interface includes NER and RE tasks, both with several methods available.

The PubMed searches to retrieve *Candida* and iron overload related publications were made using the Publication Manager plug-in. The Publication Manager view includes the date, query details, the number of publications and available abstracts for each search.

The Lexical Resources plug-in, which allows for the management of different types of lexical resources, such as dictionaries, lookup tables and ontologies, was used to create the resources that were used later in the annotation process. To create the “only-fungi”, “mycobank” and “iron-terms” dictionaries, the option Dictionaries was used and csv files with their respective terms were imported. Both “all-organisms” and “drugs” dictionaries were imported using the native loaders for NCBI Taxonomy and DrugBank files, respectively. The Ontologies option was used to import the “diseases” ontology file. Finally, the “treatment”, “outcome” and “human” lookup tables were created using the Lookup Tables option.

The corpora obtained from the different PubMed searches were then annotated with the different lexical resources using the NER Lexical Resources Tagger, with case-insensitive and no pre-processing.

After the NER process, relationships between pairs of annotated bio-entities were extracted using the RE Rel@tion Co-occurrence Extraction process. The Mix Entity Pairs Sentence model was selected, which allows for the extraction of relations between all combinations of entity pairs in a given sentence.

The obtained relationships were then exported in csv format for manual curation. For each relationship, the csv file included information about its @note internal annotation ID, the PubMed ID of the publication from which it was extracted, co-occurring bio-entities, the start offset of the sentence from which it was extracted and the sentence itself.

@Note has an associated relational database, in which all the information about PubMed queries, created corpora, lexical resources, and NER and RE processes are stored (Figure 3).

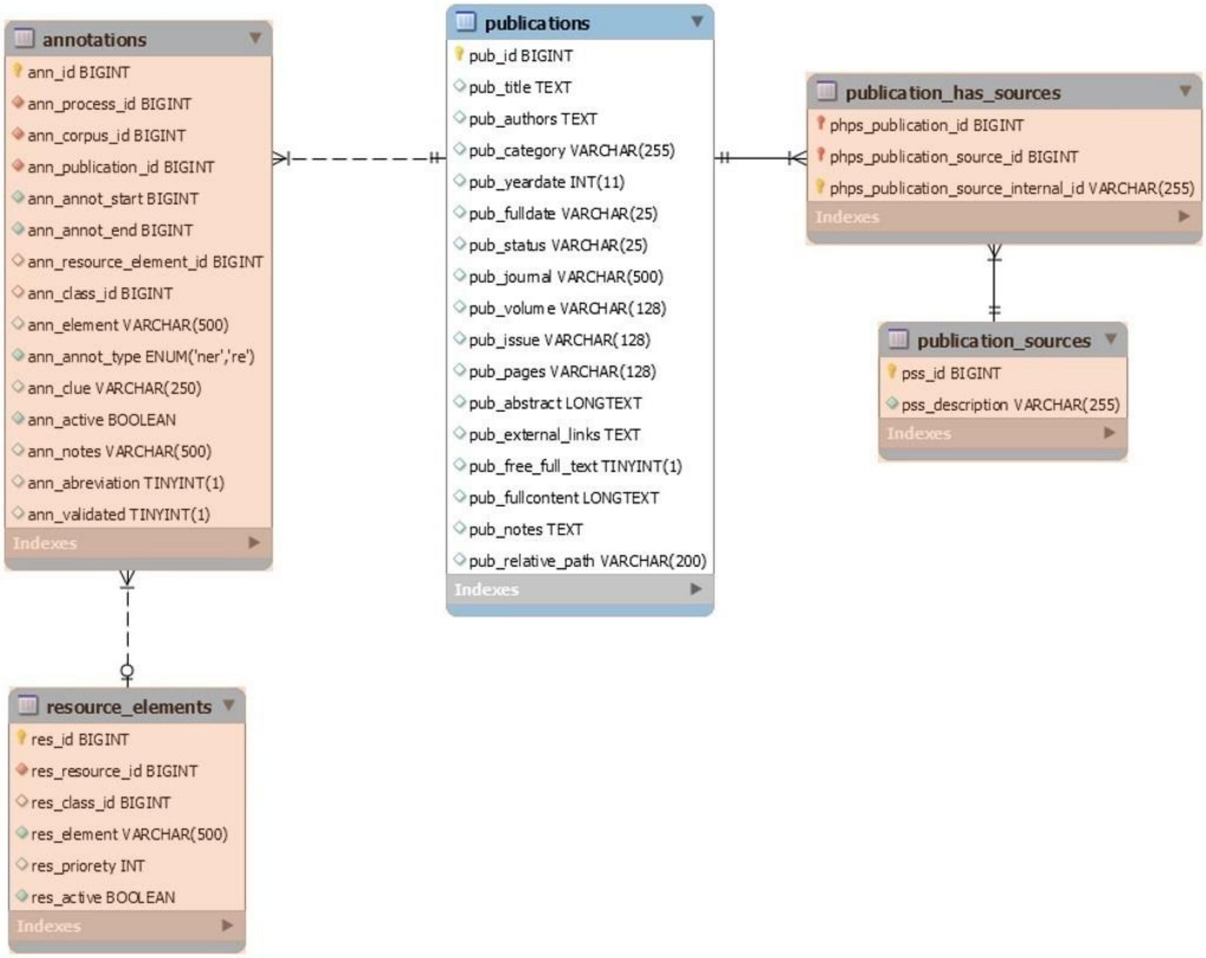


Figure 3 – Tables from @Note associated relational database used in the present work (highlighted in orange) (adapted from http://anote-project.org/wiki/images/3/38/Database.png).

3.2.2. Queries

During the BioTM process, SQL queries were done to easily access the information contained in the database created by @Note[1]. Regarding the *Candida* corpus, the following questions were answered:

- how many publications contained *Candida* annotations, with both dictionaries for organisms;
- how many annotations for each dictionary there were in total;
- how many *Candida* annotations there were in total;
- how many publications contained iron overload annotations;
- how many publications contained iron overload and *Candida* annotations (with both dictionaries) simultaneously;
- which publications contained iron overload and *Candida* annotations, simultaneously.

Regarding the iron overload corpus, the following questions were answered:

- how many annotations for each dictionary class there were in total;
- how many publications contained iron overload and fungus annotations (with both dictionaries for fungi) simultaneously;
- which publications contained iron overload and fungus annotations simultaneously.

Finally, regarding the second step of NER, the following questions were answered:

- how many annotations for each lexical resource there were in total;
- which terms were the most frequently annotated with each lexical resource.

[1] All the resources used in the present work, such as lexical resources, database file, queries done to the database, as well as the resulting corpora, NER annotations and extracted relationships are available at https://drive.google.com/drive/folders/1wxat7XaGxuAT-3zlKXumrnWnOj06-bw-?usp=sharing

4. Results and Discussion

4.1. Corpora Retrieval and Annotation

PubMed search for *Candida* resulted in 60421 documents, with 60418 of them having available abstracts. PubMed search for iron overload retrieved 13077 documents, all of them with available abstracts[2].

After the first step of NER, it was observed that only 59726 out of the 60418 abstracts retrieved from the PubMed search for *Candida* contained annotations when annotated with the "all-organisms" dictionary, and 55333 when annotated with the "only-fungi" dictionary, with a total of 521087 and 171052 annotations, respectively. The same was observed for the PubMed search for iron overload, as only 12996 out of the initial 13077 abstracts were annotated.

The fact that not all publications retrieved from PubMed searches contained annotations could be due to several reasons:

- publications obtained from a PubMed search may not always reflect the keywords used for that search, due to the labels given by the authors to the publications;
- the fact that this first step of NER was done only in the publications' abstracts can also limit the number of resulting publications, as some of them may only refer to terms of interest in their text rather than in the abstract. However, doing this step only in the abstracts helps filtering the most relevant publications, since the most important keywords of a publication often appear in the abstract;
- the dictionaries used to annotate the publications may be incomplete, which may limit the bio-entities that are annotated and, in turn, could limit the number of retrieved publications if they mention bio-entities that are missing in the dictionaries;
- regarding the annotation process of organisms, abstracts may actually contain the word, but the dictionary synonym may be associated with another term (for example, there are cases of organisms species which belonged to the genus *Candida* in the past, but currently belong to another genus);

[2] Both searches were done on 22/05/2020.

- occasionally, there are publications containing misspelled terms, especially regarding taxa names (for example, names of species are not correctly abbreviated).

In the *Candida* corpus, a total of 168938 annotations were obtained with the "only-fungi" dictionary, and 518974 annotations were obtained using the "all-organisms" dictionary, which was expected since the "all-organisms" dictionary annotates every term referring to any organism, whereas the "only-fungi" dictionary only annotates terms for fungal organisms. Another reason for the difference in the number of annotations obtained with the two dictionaries is the presence of words in taxa nomenclature such as "this", "data", "all", "other", etc. in the "all-organisms" dictionary, which are extremely common words with a whole different meaning. Therefore, many documents annotated with this dictionary will be identified as false positives when searching for publications with organisms' annotations.

In total, 118077 *Candida* annotations were obtained with the fungi dictionary, and 162241 were obtained with the "all-organisms" dictionary. This can be explained because the "only-fungi" dictionary did not contained abbreviations for species names, which were only added later for the second NER step. However, despite this limitation and despite the higher number of *Candida* annotations obtained with the "all-organisms" dictionary, a higher number of publications containing *Candida* annotations was found when using the "only-fungi" dictionary (54047), as compared to when using the "all-organisms" dictionary (50554). The reason why this happens is because "only-fungi" dictionary does not contain synonyms, which is especially important in the cases of taxa that changed their nomenclature over the time and some old publications might still refer to their old names. As such, there are cases of many *Candida* species that are not considered *Candida* nowadays (for example*, Candida humicola* has changed its name to *Vanrija humicola*), which means that the "only-fungi" dictionary annotates those species as *Candida,* but the "all-organisms" does not, and therefore some of the publications obtained with the "only-fungi" are possibly false positives.

Only 25 publications of the *Candida* corpus were annotated with the "iron overload" class of the "iron-terms" dictionary, with a total of only 38 iron overload annotations. Of those, 20 documents had annotations for *Candida* using either the "only-fungi" dictionary or the "all-organisms" dictionary. As expected, the number of publications annotated for both *Candida* and iron overload terms was too low.

Due to the dimension of the "all-organisms" dictionary, the NER process proved to be quite heavy computationally when using this dictionary. Additionally, the number of obtained publications related to *Candida* and iron overload was the same with either one of those dictionaries, despite of "only-fungi" dictionary's limitations. For those reasons, the "only-fungi" dictionary was used rather than the "all-organisms" dictionary in the second step of NER.

Regarding the iron overload corpus, 11445 publications containing a total of 31350 iron overload annotations were obtained. 150 publications containing fungi annotations were obtained with the "only-fungi" dictionary, while 527 publications were obtained using the "mycobank" dictionary. 125 documents had annotations simultaneously for iron overload and fungus using the "only-fungi" dictionary, and 448 documents using the "mycobank" dictionary. Given that the number of publications with annotations in general only differs by 10, the clear difference between the number of publications with fungus annotations obtained with the different dictionaries for fungi indicates that either the "only-fungi" dictionary is quite incomplete, or the "mycobank" dictionary annotated too many terms as fungus. In Figure 4, the number of publications obtained with all dictionaries is represented, and in Figure 5, the number of annotations obtained with all the dictionaries is summarized, for both corpora.

In the second step of NER, two corpora were created: one containing the 125 publications obtained with the "only-fungi" dictionary, and another containing the 448 publications obtained with the "mycobank" dictionary. The resulting corpora were annotated for fungi using the "only-fungi" and the "mycobank" dictionaries, respectively, in addition to the other lexical resources.

After the annotation process, the 125 documents annotated with the "only-fungi" dictionary had 1701 annotations for fungi, 595 annotations for iron overload, 7158 annotations for drugs, 2194 annotations for diseases, 494 annotations for treatment, 38 annotations for outcome, and 342 annotations for human, with a total of 15694 annotations The 448 documents annotated with the "mycobank" dictionary had 3736 annotations for fungi, 595 annotations for iron overload, 19338 annotations for drugs, 7229 annotations for diseases, 1454 annotations for treatment, 113 annotations for outcome, and 873 annotations for human, with a total of 43238 annotations (Figure 6).

By analysing which bio-entities are the most commonly annotated by each dictionary (Table 1, Annex), it is possible to observe that there is a clear ambiguity issue, caused by the presence of words in dictionaries with different meanings, which depends on the context of the text (Table 2, Annex). For example, words such as "*Drosophila*", "melanogaster", "*C. elegans*", "*P. aeruginosa*" (in both "only-fungi" and "mycobank" dictionaries), "necrosis", "*Plasmodium*", "*Xenopus*", "omega" (in "mycobank" dictionary), "all", "can" (in "drugs" dictionary), "al", "as", "Fig", "mg", "ng" (in "diseases" ontology) were annotated. However, in most cases, if not in all of them, their meaning in the text where they appear are completely different of their meaning in the dictionaries.

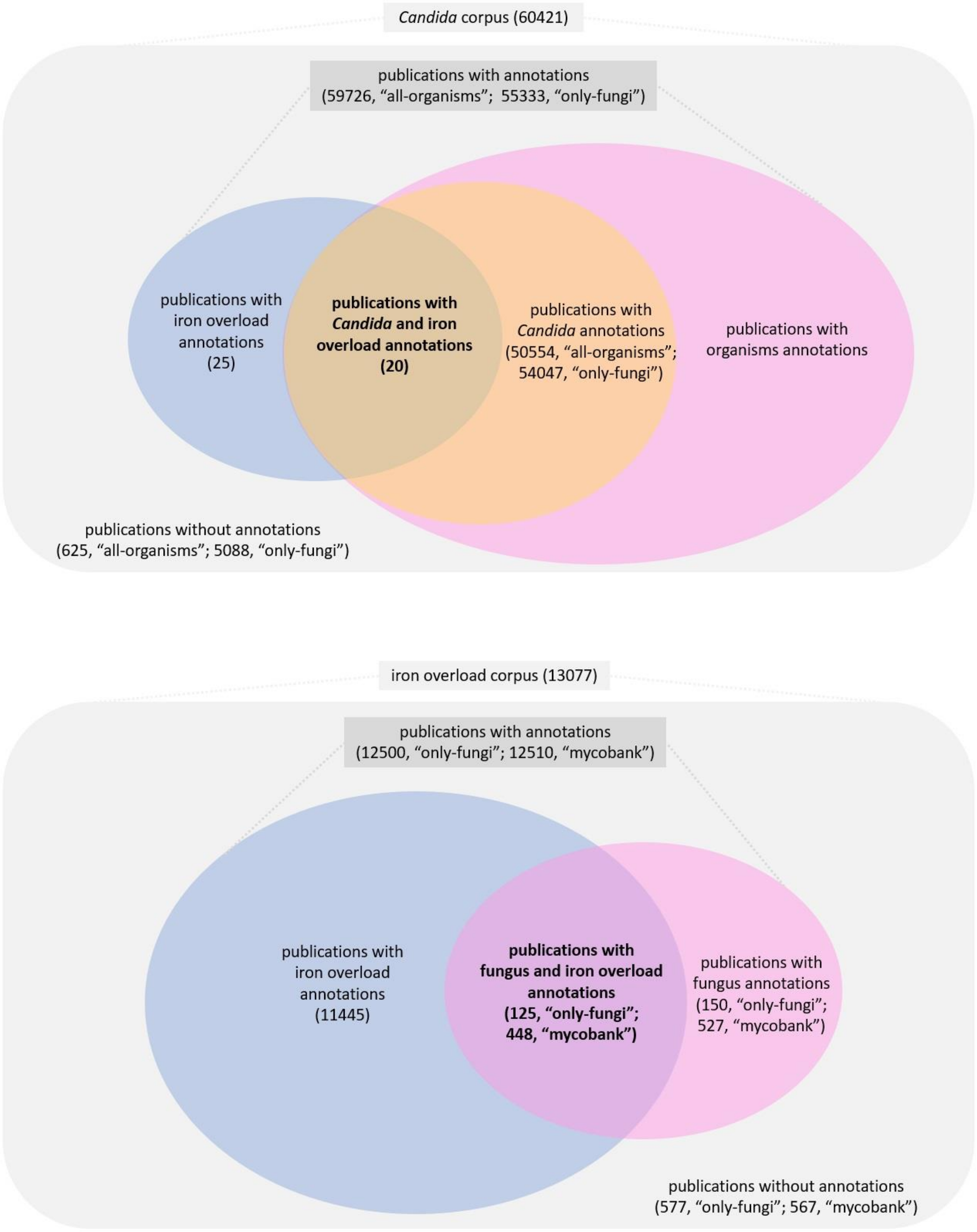

Candida corpus (60421)
publications with annotations
(59726, "all-organisms"; 55333, "only-fungi")
publications with iron overload annotations (25)
publications with *Candida* and iron overload annotations (20)
publications with *Candida* annotations (50554, "all-organisms"; 54047, "only-fungi")
publications with organisms annotations
publications without annotations
(625, "all-organisms"; 5088, "only-fungi")
iron overload corpus (13077)
publications with annotations
(12500, "only-fungi"; 12510, "mycobank")
publications with iron overload annotations (11445)
publications with fungus and iron overload annotations (125, "only-fungi"; 448, "mycobank")
publications with fungus annotations (150, "only-fungi"; 527, "mycobank")
publications without annotations
(577, "only-fungi"; 567, "mycobank")

Figure 4 – Number of publications with annotations for all dictionaries, for both *Candida* and iron overload corpora, in the first step of NER. In brackets, the number of publications obtained with each dictionary for organisms terms is indicated, whenever applicable.

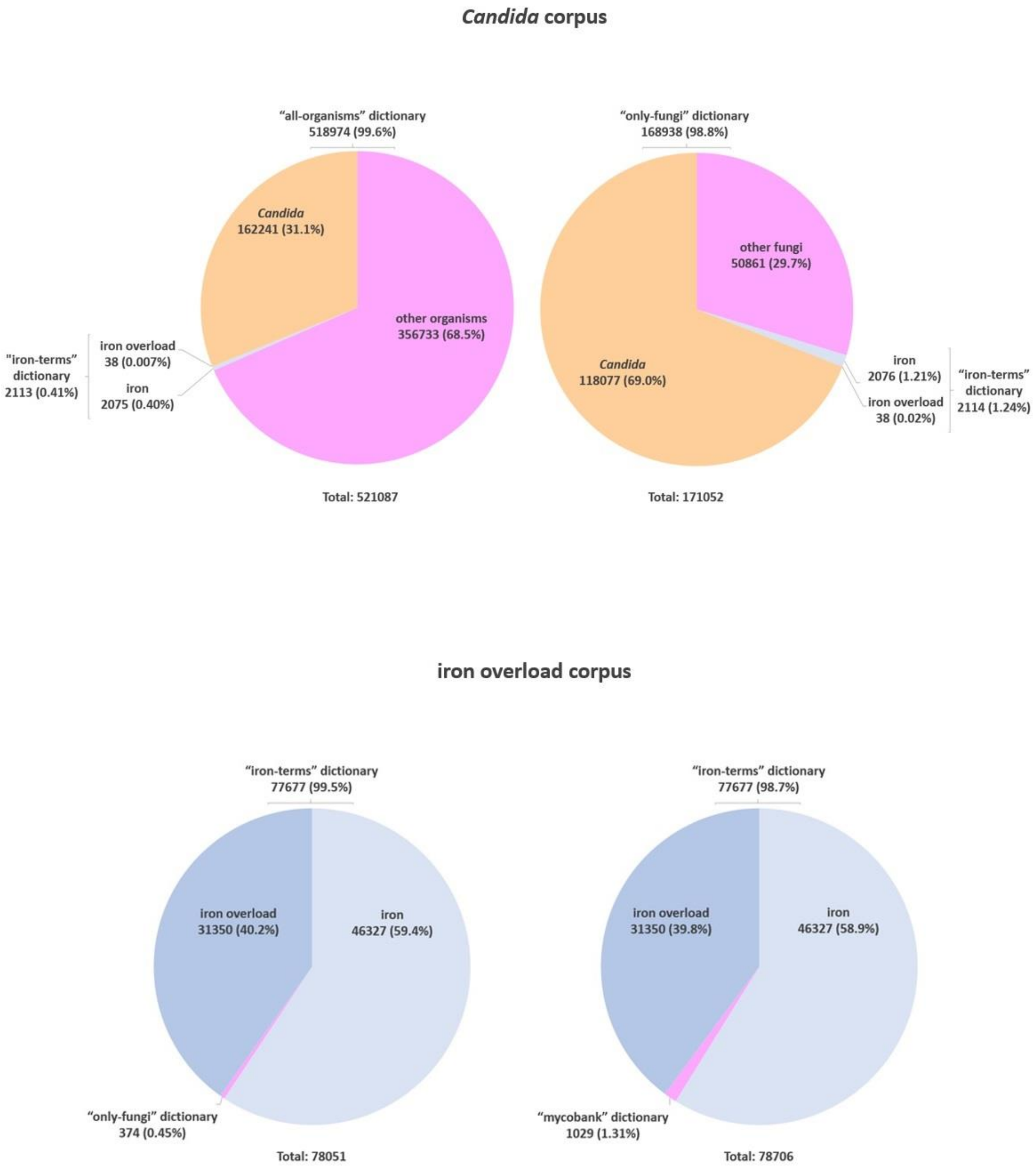
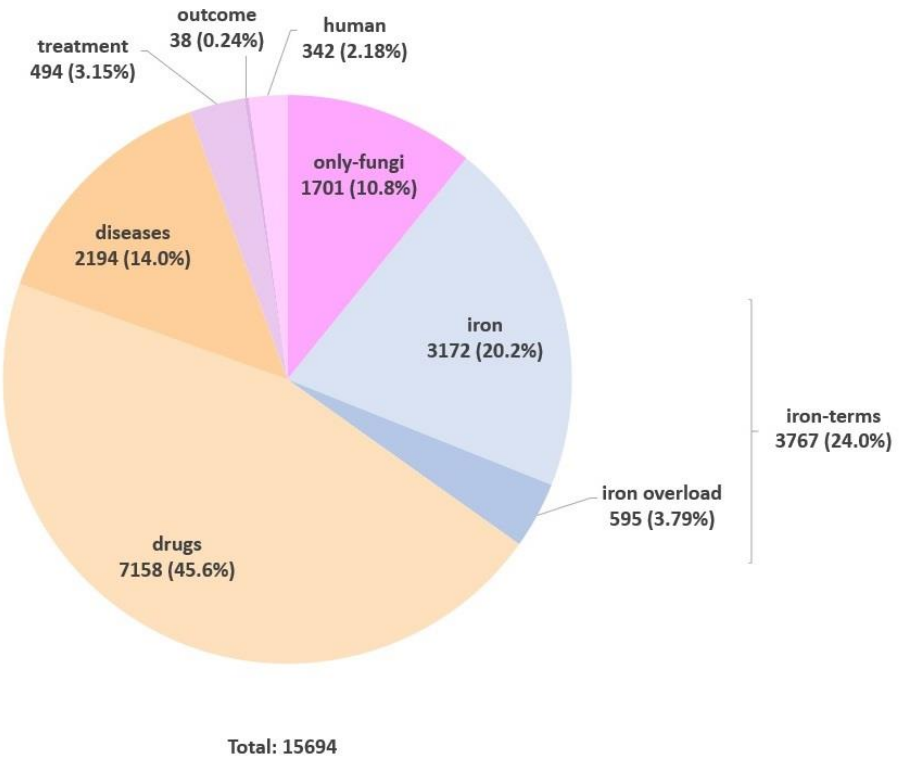


Figure 5 – Number of annotations of *Candida*, fungus and iron terms, for both corpora, when annotated with different dictionaries for organisms, in the first step of NER.



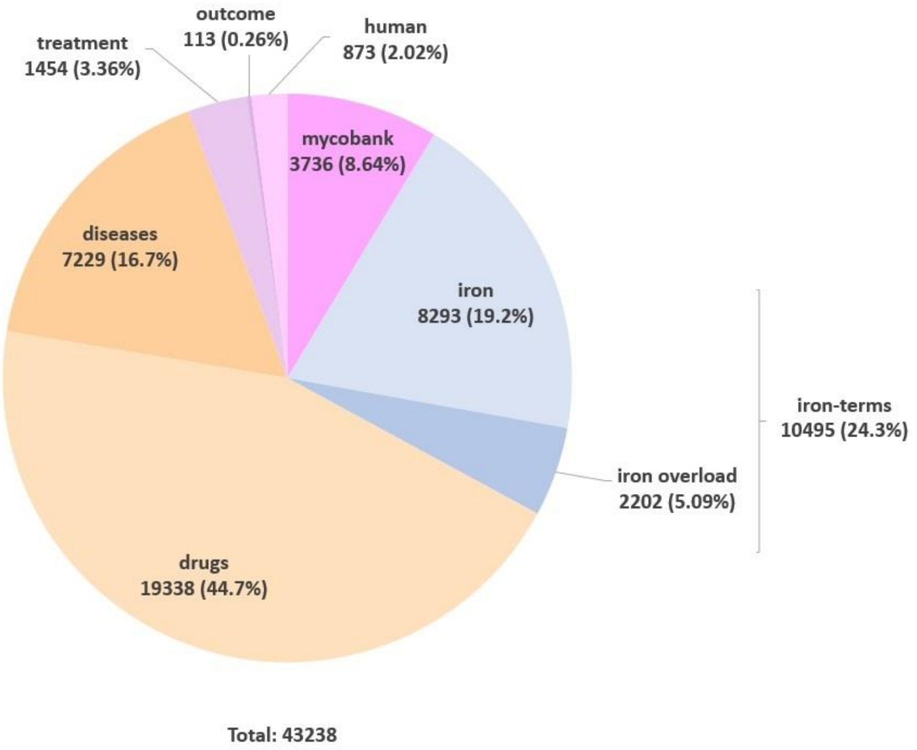


Figure 6 – Number of annotations obtained with all the lexical resources used in the second step of NER, for both corpora.

In addition to those, the words “bacteria”, “bacterium”, “algae” were annotated by the “mycobank” dictionary even though they do not represent any fungal organism, because this dictionary contains taxa that were considered fungal organisms in the past but currently are not. This explains the wide difference between the number of publications obtained with both dictionaries used to annotate fungi entities in the iron overload corpora, observed in the previous step.

In order to overcome this ambiguity problem, these words were defined as stopwords. However, after repeating this NER step using stopwords, those words were still being annotated, which proved to be a limitation of @Note. Therefore, those terms were manually filtered only after the RE process.

4.2. Extraction of Relevant Associations

After the RE step, 21018 relations were extracted from the corpus annotated with the “only-fungi” dictionary, and 48613 relations were extracted from the corpus annotated with the “mycobank” dictionary. Due to the ambiguity issue of both dictionaries for fungi, all publications not containing any annotations regarding actual fungal organisms were discarded. After that filtration, only 95 documents had annotations for fungus and iron, when annotated with the “only-fungi” dictionary, and 99 when annotated with the “mycobank” dictionary. Relationships were then filtered regarding the remaining lexical resources: those containing entities not related to either iron, fungus, disease, drug, treatment, outcome or human were not considered. At the end, there were a total of 5285 and 5754 relationships when annotated with the “only-fungi” and the “mycobank” dictionaries, respectively.

As stated previously, both “only-fungi” and “mycobank” dictionaries contain ambiguous terms, although this issue seems to be more prevalent in the “mycobank” dictionary. Nevertheless, “mycobank” is more complete, therefore it is able to annotate taxa that are not annotated with the “only-fungi” dictionary. For instance, one publication mentions the fungus *Pneumocystis jiroveci*, which is fully annotated with the “mycobank” dictionary, but only partially annotated for *Pneumocystis* with the “only-fungi” dictionary. Also, in another publication, *Memnoniella echinate* is annotated with the “mycobank” dictionary, but is not annotated with the “only-fungi” dictionary.

In addition to the previously mentioned limitations of both dictionaries for fungal organisms, other limitations were observed. For instance, *Pythium insidiosum* and *Dictyostelium discoideum* are not fungi, but both are annotated by "mycobank" dictionary because they have been categorized as fungi in the past.

In some cases, an abbreviation for a given fungus species may represent more than one different species of fungi within the dictionary. For example, in one publication, *Candida humicola* was not annotated with the "only-fungi" dictionary because it does not contain this species name. However, this term was annotated when abbreviated to *C. humicola*, although it does not refer to *Candida humicola* species but to *Corneriella humicola*. The "mycobank" dictionary, on the other hand, annotated this term correctly. This issue could be solved with tools for abbreviation annotation (@Note has an option for abbreviation). However, these tools only work if the term appears in full before it appears abbreviated in the same publication, which is not always the case. Additionally, using these tools can lead to ambiguity issues, which decreases the precision in annotation.

Regarding the "drugs" dictionary and the "diseases" ontology, both contain organism names. For instance, *Corynebacterium diphtheriae* is annotated as a drug, and *Vibrio cholerae* is annotated as a disease. This can be explained by the fact that many organisms are used as drugs, and many are the cause of diseases. However, in this context, this may be a limitation, since it contributes to the ambiguity problem of using dictionaries. For instance, the word "yeast" is annotated by the "drugs" dictionary, although in no case this term referred to yeast as a drug in any of the publications.

Some limitations regarding @Note were also observed. The sentence splitter does not always split sentences correctly, as sometimes it splits sentences in half, other times does not split at all and, as a result, two sentences are joint as one. This may happen due to the presence of abbreviations with dots, which hampers the sentence splitting process. For instance, in one publication, there is a sentence that contains the word "fig." and it is split at that word. In another publication, two sentences are joint together, and C from the term "hepatitis C" from the first sentence is joint with the word "Serum" from the next sentence, and as a result, *C. serum*, which is a fungal organism, is annotated. Another reason is the formatting of some documents, which does not always allow for a correct sentence splitting, nor for a correct annotation of bio-entities.

Additionally, in some publications, some sentences containing bio-entities of interest are not annotated as relationships. For instance, in one of the publications, the sentence "Mucormycosis is caused by fungi." is not annotated as a relationship, although the terms "Mucormycosis", annotated by the “diseases” ontology, and the word “fungi”, annotated by the “only-fungi” dictionary, co-occur in that same sentence. In a similar way, in another publication, the term *Aspergillus* is not included in any relationship, even though they co-occur with other annotated bio-entities in the sentence “Aspergillus disease after lung transplantation includes airway anastomotic infections, severe asthma, and invasive pulmonary aspergillosis, an infection of the lower respiratory tract.".

There are other limitations regarding the annotation of bio-entities that are caused due to the way publications are written. For instance, misspelled taxa names in publications are never annotated, as well as species names that are incorrectly abbreviated, as they do not match any term in the dictionary (ex., *Cr. Neoformans* instead of *C.* neoformans; *C pseudotropicalis* instead of *C. pseudotropicalis*). In a similar way, some authors use abbreviations that are not in the dictionaries. For example, the abbreviation for pulmonary hemorrhage, PH, is not annotated because it is not present in the “diseases” ontology. As a result of this kind of limitations, many bio-entities of interest will not be annotated and, consequently, some eventually interesting relationships might be missing.

Finally, there are some interesting terms that are missing in the lexical resources used. For instance, the term “invasive fungal infection” is not present in the “diseases” ontology, nor the word “fungal” in any of the dictionaries for fungi, which is a limiting factor for the analysis in the present work.

In conclusion, the same relevant publications were obtained whether using the “only-fungi” dictionary or the “mycobank” dictionary. Overall, the “only-fungi” dictionary annotated fungus entities with more precision than the “mycobank” dictionary since it has shown to be less ambiguous. However, since the number of relevant relationships extracted when using the “mycobank” dictionary was slightly higher than when the “only-fungi” dictionary was used, from this point on, the results hereby presented will be the ones obtained regarding the “mycobank” dictionary.

4.3. Curated Dataset

Publications with relevant sentences were analysed regarding their relevance. After analysing all the publications, it was observed that 5 of them are written in languages other than English: two are written in Spanish, one in German, one in Hungarian and one in Japanese. As a result, only the abstracts, which are in English, can be analysed in these cases.

There were 38 publications of the 99 which were not related to fungal infections:

- 11 publications were about the study of iron-related diseases where fungi (*Saccharomyces cerevisiae*) were used as model organisms;
- 6 were about exposure to moulds causing acute pulmonary hemorrhage in infants;
- 5 publications contained the term fungi but did not focus on fungal infections. Of those, three were about bacterial infections, one was about HIV infection, and another was about the effect of iron supplements in host defence from pathogenic microorganisms;
- Two were about fungal organisms that cause diseases in plants;
- The remaining focused on, among others, effects of feeding cattle a diet contaminated with fungi, hepatocellular carcinoma caused by human consumption of food contaminated with carcinogenic fungi, liver injury caused by ingestion of mushrooms, iron-chelating activity of mushrooms and iron tolerance/toxicity in fungi.

The remaining 61 publications were related to fungal infections. Of those:

- 20 were reviews, 13 of them were about mucormycosis, two were about *Aspergillus*, one was about *Candida*, and 6 were about fungi in general. In 11, iron overload was not the main focus, 3 focused on iron chelators, and one reviewed methods and agents for strengthening host's iron-withholding defence;
- 11 were studies in animal models, 4 of them using *Candida*, 3 using *Pneumocystis carinii*, 2 using *Cryptococcus neoformans*, one using *Rhizopus oryzae*, and one using *Aspergillus fumigatus*;
- 5 studies in vitro, two using *Cryptococcus neoformans*, one using *Candida albicans*, one using *Candida glabrata*, and one using *Penicillium marneffei*;
- 25 clinical cases, 8 of which iron overload is not the case of infection (they only contain the term as one of many risk factors for fungal infections), and one case is related to iron overload but there is no fungal infection. All the remaining cases were considered relevant.

Ultimately, only 15 publications were considered relevant [121-135]. All of them are either case reports or studies based on case reports. Of those, 8 are about mucormycosis/zygomycosis (4 are caused by *Rhizopus* [122, 129, 131, 134], one is caused by *Cunninghamella bertholletiae* [133]), two are cutaneous infections [123, 130], two about invasive mould infections [125, 135], one about *Candida* [128], one about *Trichosporon asahii* fungemia [121], one about *Trichophyton rubrum* dermatophytosis [130], and the remaining two about fungal infections in general [126, 132] (Table 3, Annex).

It is important to note that two of the 15 publications that were considered relevant did not contain any extracted associations that could be considered relevant, due to some of the limitations described earlier regarding the extraction of relevant associations. Nevertheless, they were considered relevant since both of them describe indeed cases of fungal infections in a clinical context of iron overload.

Table 5 summarizes important information regarding the relevant publications obtained, including number of relevant relationships and interesting annotations in each publication. Relevant associations between iron and fungus, iron and disease, and fungus and drug, extracted from those publications, are summarized in tables 6, 7 and 8, respectively.

Table 5 – Overview of the relevant publications obtained.

PubMed ID	Text available	Number of relevant relationships	Interesting annotations
20434128	Abstract	2	*Trichosporon asahii*; nosocomial fungemia; secondary hemochromatosis
28348771	Abstract	1	*Rhizopus*; transfusion-dependent beta thalassemia; iron overload; deferiprone
7805414	Abstract	1	Mucormycosis; iron overload
10723242	Abstract	3	mucormycosis; aplastic anemia; myelodysplastic syndrome; iron overload; deferoxamine
18781877	Full text	5	Invasive mould infections; hematopoietic stem cell transplantation; *Aspergillus*; iron overload
21331523	Abstract	2	Bloodstream infections; iron overload

17852457	Abstract	4	zygomycosis; iron overload; deferoxamine
2640481	Abstract	4	*Candida*; thalassemia major
10394647	Abstract	3	Mucormycosis; aplastic anemia; neutropenia; *Rhizopus*; hemochromatosis, desferrioxamine; amphotericin B
20092423	Abstract	3	dermatophytosis; *Trichophyton rubrum*; hereditary hemochromatosis; cirrhosis: iron overload
3662280	Abstract	0	*Rhizopus*; deferoxamine; hemodialysis
16741903	Abstract	2	iron overload; liver transplant; fungal infection; *Candida*; *Aspergillus*; *Cryptococcus*; *Saccharomyces*
3060947	Abstract	1	mucormycosis; *Cunninghamella bertholletiae*; iron overload; deferoxamine
15078434	Abstract	1	myelodysplastic syndrome; iron overload; pulmonary *Rhizopus oryzae* infection; itraconazole
25082161	Abstract	0	invasive mould infections; *Aspergillus*; mucorales; allogeneic hematopoietic stem cell transplantation; ferritin

Table 6 – Overview of the iron-fungus relevant associations.

Bio-entities	**Sentence**
hemochromatosis, *T. asahii*	"We report a 53-year-old nongranulocytopenic female with secondary hemochromatosis, who developed nosocomial fungemia caused by *T. asahii*." [121]
T. asahii, hemochromatosis	"This case suggests that clinicians should be aware that *T. asahii* fungemia can develop in nongranulocytopenic patients with secondary hemochromatosis." [121]

Mucor, iron overload	“Primary cutaneous mucormycosis with a Mucor species: is iron overload a factor?” [123]
iron overload, zygomycosis	“Most of the evidence for iron overload impacting on the risk of IMI comes from studies of zygomycosis.” [125]
zygomycosis, iron overload	“Maertens *et al*. [12] reported 5 cases of zygomycosis in allogeneic HSCT recipients, and iron overload was present in all 5 cases.” [125]
candida, iron overload	“With respect to phagocytes, the capacity to ingest candida is preserved while the candidacidal activity and the generation of toxic oxygen metabolites during the respiratory burst are diminished, and are inversely proportional with age and serum ferritin concentration, that is, older in age and higher in iron overload, more profound are the phagocyte dysfunctions.” [128]
hemochromatosis, *Rhizopus*	“The second patient did not survive his severe aplastic anemia (with neutropenia) and hemochromatosis (treated with desferrioxamine), complicated with a systemic *Rhizopus* infection, despite treatment with amphotericin B and granulocyte-colony-stimulating factors.” [129]
dermatophytosis, hemochromatosis	“Disseminated dermatophytosis in a patient with hereditary hemochromatosis and hepatic cirrhosis: case report and review of the literature.” [130]
mucormycosis, iron	“Recent reports of mucormycosis in dialysis patients receiving deferoxamine for iron or aluminum overload have raised the possibility that deferoxamine therapy is a risk factor for mucormycosis.” [133]
C. bertholletiae, iron overload	“A case of *C. bertholletiae* infection in a patient receiving deferoxamine for iron overload unrelated to hemodialysis was investigated in detail, and possible explanations for this patient's infection were assessed.” [133]
iron overload, *Rhizopus oryzae*	“We report a non-neutropenic patient with myelodysplastic syndrome and iron overload receiving cytotoxic therapy who presented with pulmonary *Rhizopus oryzae* infection.” [134]

Table 7 – Overview of the iron-disease relevant associations.

Bio-entities	**Sentence**
iron overload, aplastic anemia	"Deferoxamine has been also used in the treatment of iron overload patients with aplastic anemia." [124]
myeloma, iron	"In a study involving 365 patients with myeloma who underwent autologous HSCT, bone marrow iron level was an independent risk factor for the development of severe infection." [125]
iron overload, acute myeloid leukemia	"We retrospectively studied the association between iron overload and bloodstream infections (BSI) in the 100-day period following allogeneic hematopoietic stem cell transplantation (allo-HSCT) for acute myeloid leukemia or myelodysplastic syndromes." [126]
iron overload, liver disease	"It is also the second case report of a Pearson patient suffering from severe iron overload and liver disease that responded to therapy with deferoxamine." [127]
thalassemia, iron overload	"It is accepted that the immune alterations in patients with thalassemia major (TM) are secondary to the continuous transfusion-related antigenic stimulation together with iron overload." [128]
aplastic anemia, hemochromatosis; neutropenia, hemochromatosis	"The second patient did not survive his severe aplastic anemia (with neutropenia) and hemochromatosis (treated with desferrioxamine), complicated with a systemic *Rhizopus* infection, despite treatment with amphotericin B and granulocyte-colony-stimulating factors." [129]
myelodysplastic syndrome, iron overload	"We report a non-neutropenic patient with myelodysplastic syndrome and iron overload receiving cytotoxic therapy who presented with pulmonary *Rhizopus oryzae* infection." [134]

Table 8 – Overview of the fungus-drug relevant associations.

Bio-entities	**Sentence**
deferiprone, *Rhizopus*	"Although deferiprone, a newer iron chelator agent, has antifungal properties in vivo, this case illustrates that angioinvasive *Rhizopus* infections can occur in patients treated with deferiprone." [122]
mucormycosis, deferoxamine	"Cases of mucormycosis occurring in dialysis patients receiving deferoxamine have recently appeared in the literature." [124]
mucormycosis, deferoxamine	"There may be a relationship between mucormycosis and deferoxamine in patients with aplastic anemia." [124]
deferoxamine, *Pneumocystis jiroveci* deferoxamine, zygomycosis	"After an initial response to deferoxamine she presented with cutaneous zygomycosis and died after metabolic derangement and *Pneumocystis jiroveci* pneumonia." [127]
desferrioxamine, *Rhizopus*; *Rhizopus*, amphotericin B	"The second patient did not survive his severe aplastic anemia (with neutropenia) and hemochromatosis (treated with desferrioxamine), complicated with a systemic *Rhizopus* infection, despite treatment with amphotericin B and granulocyte-colony-stimulating factors." [129]
deferoxamine, rhizopus	"Four hemodialysis patients receiving deferoxamine for metal overload had fatal rhinocerebral rhizopus infections." [131]
Rhizopus, deferoxamine	"Fatal *Rhizopus* infections in hemodialysis patients receiving deferoxamine." [131]
mucormycosis, deferoxamine	"Recent reports of mucormycosis in dialysis patients receiving deferoxamine for iron or aluminum overload have raised the possibility that deferoxamine therapy is a risk factor for mucormycosis." [133]
C. bertholletiae, deferoxamine	"A case of *C. bertholletiae* infection in a patient receiving deferoxamine for iron overload unrelated to hemodialysis was investigated in detail, and possible explanations for this patient's infection were assessed." [133]
itraconazole, zygomycosis	"This patient was cured through the use of itraconazole alone and the literature on the utility of azole antifungals for zygomycosis is reviewed." [134]

Rhizopus oryzae, itraconazole	"Complete resolution of pulmonary *Rhizopus oryzae* infection with itraconazole treatment: more evidence of the utility of azoles for zygomycosis." [134]

Upon analysing all the publications obtained and their respective relevant associations, it was possible to observe that the most mentioned fungal infections are mucormycosis and invasive mould infections. As such, organisms from the Mucorales order, such as *Rhizopus*, and moulds, such as *Aspergillus*, were the most referred.

Antifungals mentioned in the publications include itraconazole, voriconazole, amphotericin B and caspofungin. Itraconazole has shown to be effective in the treatment of a pulmonary *Rhizopus oryzae* infection [134]. Amphotericin B was not effective in a case of a systemic *Rhizopus* infection [129], but it has shown to be effective together with voriconazole and caspofungin, in another case of a *Rhizopus* infection [122].

High iron levels seem to be an indicator of susceptibility for fungal infections. Other factors or diseases related to iron overload that can also increase the risk for fungal infections include anemia, blood transfusions, hemodialysis, liver transplants and hematopoietic stem cell transplants. Usually, iron chelators are used as adjuvants in treating fungal infections, since they reduce iron levels in patients with iron overload. The most cited iron chelator in the resulting publications was deferoxamine (or desferrioxamine), followed by deferiprone. However, in many cases, iron chelators end up having the opposite effect, which made patients even more susceptible to fungal infections, since many organisms have developed strategies to take up chelated iron from iron chelators. In 6 out of the 8 publications about mucormycosis, patients who received iron chelating treatments for iron overload had shown complications with mucormycosis [122, 124, 127, 129, 131, 133]. In three cases [124, 129, 131], the patients did not survive, despite two of them having received antifungal treatment [124, 129]. Three publications suggested a possible link between deferoxamine therapy and the emergence of mucormycosis [124, 131, 133]. Deferiprone, a newer iron chelator agent which has shown to have antifungal properties *in vivo*, has also been associated to mucormycosis infections. A case of a patient who started on deferiprone therapy and subsequently presented with an angioinvasive *Rhizopus* infection suggests that this kind of infections may occur in patients treated with deferiprone [122].

Although associations of iron levels with both severity of infection and treatment efficacy were not extracted, studies in patients who had undergone hematopoietic stem cell transplantation had shown that iron overload is a biological risk factor for fungal infections after hematopoietic stem cell transplantation [125], and pretransplantation serum ferritin is a strong predictor of BSIs, including IFIs, within a 100-day period after the cell transplant [126, 135]. The same is observed in liver transplantation, where hepatic iron overload is strongly associated with posttransplantation IFIs [132].

Moreover, it was possible to indirectly infer that high iron is associated with more severe fungal infections, since some iron chelating treatments have shown to increase the emergence of fungal infections in some cases, as opposed to what would be expected. It was also possible to conclude that treatment efficacy is highly affected by patients' iron levels, since in many cases patients with iron overload or who were being treated with iron chelators, and who had invasive fungal infections, did not survive, even though they were receiving antifungal therapy.

5. Conclusions and Future Work

The BioTM process described in the present thesis enabled the creation of a dataset of relevant biomedical publications containing interesting associations between fungal infections, drugs and associated diseases in a clinical context of iron overload. Although initially the main goal of this work was to obtain interesting publications related specifically to *Candida* and iron overload, that was not possible to achieve since there are very few publications on cases related to candidiasis and iron overload in PubMed.

In order to compare the performance of this work's BioTM process with the performance of a simple PubMed search, a PubMed search was done with the query *(Fungal OR Fungus OR Fungi OR Fungemia OR mycosis) AND (transferrin OR ferritin OR iron) AND (iron overload OR Fe overload OR iron excess OR Fe excess) AND infection AND (treatment OR therapy OR outcome OR mortality OR morbidity OR prognostic) NOT bacteria NOT review*, which returned 32 publications, 12 of which are relevant. When comparing those 12 relevant publications with the 15 relevant publications of the dataset obtained in this work, only a single publication [125] was found to be common between the two datasets. On one hand, the PubMed search is effective in returning relevant publications that mention the term "fungal infection", as it includes the keywords "fungal" and "infection" in the query. As pointed out before, none of the dictionaries used in this work annotates these words, which limits to a certain extent the number and relevance of the publications obtained. However, those words were not considered for the initial search in this work because they are quite generic and, as such, their inclusion would lead to a high number of uninteresting publications. On the other hand, the PubMed search clearly failed to return publications that do not mention the terms "fungal infection" but that may mention fungal taxa instead. This is why the BioTM process described in the present work proves to be more effective than a simple search on PubMed, especially when the goal is to create a corpus on a given theme that encompasses a large number of publications and/or involves a large number of keywords, such as a group of genes, proteins, organisms, etc., which makes this task impossible to be done manually.

Although the annotation process using dictionaries may take time (for instance, the NER process of the *Candida* corpus, which was the largest corpus, when annotated with the all-organisms dictionary, which was also the largest dictionary, took approximately two weeks), the manual curation of a dataset obtained from this BioTM process is less time-consuming than manually curating a dataset obtained from a PubMed search, since the NER and RE processes help filtering publications of interest. If dictionaries are optimized and they do not contain ambiguous words, the manual curation step may be even less time-consuming.

One disadvantage of this process is the lack of context in some cases, where extracted relationships do not give enough information to conclude about the relevance of the whole publication. Overall, documents with relevant relationships were considered relevant, as in most of the cases, the relationship itself is enough for the evaluation of its relevance. However, sentences immediately before and after are usually needed to infer the overall context of a given sentence, since there were a few cases of relevant relationships extracted from publications that were not relevant. This is the main reason why manual curation is an important step of this process.

Another disadvantage is the limitation of getting full-text publications from PubMed, since for most of them only the abstract is available, which seriously limits the text analysis, especially in the cases where the abstract does not clearly summarize the study and, consequently, is not possible to draw any conclusions about the publication. Additionally, half of the clinical reports obtained in this work date from more than 15 years ago.

In conclusion, this method is useful and effective, although in the future it could be improved regarding several points, in order to return a higher number of publications on iron overload and fungal infections:

- NER process could be done in a case-sensitive way, especially when annotating taxa names, in order to improve the accuracy of annotation of taxa bio-entities;
- dictionaries could be improved or additional ones could be used. For instance, the words “fungal” and “infection” could be included, and the lookup table for human terms could be improved and include terms for clinical context, such as “patient”, in order to optimize the search for clinical cases;
- the first step of NER could be done using dictionaries containing more synonyms (more ambiguous), and the publications obtained could be filtered afterwards using more precise dictionaries (less ambiguous);
- different methods for NER and RE could be used, for instance, ML-based;
- clinical databases such as PubMed Clinical Queries [136] or in ClinicalTrials.gov database [137] could be searched, to obtain a higher number of clinical studies.

In addition, further searches with different keywords could also be done, to explore which genes/proteins are involved in the effect of iron excess in exacerbating fungal infections.

References

1. Beaute, J., Hollo, V., Kodmon, C., Snacken, R., Nicoll, A., van der Werf, M. J., ... & Spiteri, G. (2014). Annual Epidemiological Report, 2013. Reporting on 2011 surveillance data and 2012 epidemic intelligence data. Annual Epidemiological Report, 2013. Reporting on 2011 surveillance data and 2012 epidemic intelligence data.
2. Wisplinghoff, H., Bischoff, T., Tallent, S. M., Seifert, H., Wenzel, R. P., & Edmond, M. B. (2004). Nosocomial bloodstream infections in US hospitals: analysis of 24,179 cases from a prospective nationwide surveillance study. *Clinical infectious diseases*, 39(3), 309-317.
3. Soares, M. P., & Weiss, G. (2015). The Iron age of host–microbe interactions. *EMBO reports*, 16(11), 1482-1500.
4. Bullen, J. J., Rogers, H. J., Spalding, P. B., & Ward, C. G. (2006). Natural resistance, iron and infection: a challenge for clinical medicine. *Journal of medical microbiology*, 55(3), 251-258.
5. Shatkay, H., & Craven, M. (2012). *Mining the biomedical literature*. MIT Press.
6. Clark, A., Fox, C., & Lappin, S. (Eds.). (2013). The handbook of computational linguistics and natural language processing. John Wiley & Sons.
7. Rodriguez-Esteban, R. (2009). Biomedical text mining and its applications. *PLoS computational biology*, 5(12), e1000597.
8. Miner, G., Elder IV, J., Fast, A., Hill, T., Nisbet, R., & Delen, D. (2012). *Practical text mining and statistical analysis for non-structured text data applications*. Academic Press.
9. Hoffmann, R., & Valencia, A. (2004). A gene network for navigating the literature. *Nature genetics*, 36(7), 664-664.
10. Quan, C., Wang, M., & Ren, F. (2014). An unsupervised text mining method for relation extraction from biomedical literature. *PloS one*, 9(7), e102039.
11. Pawar, S., Palshikar, G. K., & Bhattacharyya, P. (2017). Relation extraction: A survey. *arXiv preprint arXiv:1712.05191*.
12. Vandeputte, P., Ferrari, S., & Coste, A. T. (2011). Antifungal resistance and new strategies to control fungal infections. *International journal of microbiology*, 2012.
13. Kullberg, B. J., & Arendrup, M. C. (2015). Invasive candidiasis. *New England Journal of Medicine*, 373(15), 1445-1456.
14. Paiva, J. A., Pereira, J. M., Tabah, A., Mikstacki, A., de Carvalho, F. B., Koulenti, D., ... & Antonelli, M. (2016). Characteristics and risk factors for 28-day mortality of hospital acquired fungemias in ICUs: data from the EUROBACT study. *Critical Care*, 20(1), 53.
15. Fridkin, S. K. (2005) Candidemia is costly—plain and simple, *Clinical infectious diseases*, 41(9), 1240-1241.
16. Arnold, H. M., Micek, S. T., Shorr, A. F., Zilberberg, M. D., Labelle, A. J., Kothari, S., & Kollef, M. H. (2010). Hospital resource utilization and costs of inappropriate treatment of candidemia. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, 30(4), 361-368.
17. Perlin, D. S. (2011). Current perspectives on echinocandin class drugs. *Future microbiology*,

6(4), 441-457.

18. Klis, F. M., Groot, P. D., & Hellingwerf, K. (2001). Molecular organization of the cell wall of *Candida albicans*. *Medical mycology*, 39(1), 1-8.

19. Caroline, L., Rosner, F., & Kozinn, P. J. (1969). Elevated serum iron, low unbound transferrin and candidiasis in acute leukemia. *Blood*, 34(4), 441-451.

20. Ramanan, N., & Wang, Y. (2000). A high-affinity iron permease essential for *Candida albicans* virulence. *Science*, 288(5468), 1062-1064.

21. Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv preprint arXiv:1707*.02919.

22. Feldman, R., & Sanger, J. (2007). The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge university press.

23. Fleuren, W. W., & Alkema, W. (2015). Application of text mining in the biomedical domain. *Methods*, 74, 97-106.

24. MEDLINE®: Description of the Database. nlm.nih.gov/bsd/medline.html

25. PubMed. ncbi.nlm.nih.gov/pubmed

26. Masic, I., & Milinovic, K. (2012). On-line biomedical databases–the best source for quick search of the scientific information in the Biomedicine. *Acta Informatica Medica*, 20(2), 72.

27. Falagas, M. E., Giannopoulou, K. P., Issaris, E. A., & Spanos, A. (2007). World databases of summaries of articles in the biomedical fields. *Archives of internal medicine*, 167(11), 1204-1206.

28. Frijters, R., Verhoeven, S., Alkema, W., van Schaik, R., & Polman, J. (2007). Literature-based compound profiling: application to toxicogenomics. *Pharmacogenomics*, 8(11), 1521-1534.

29. Fleuren, W. W., Toonen, E. J., Verhoeven, S., Frijters, R., Hulsen, T., Rullmann, T., ... & Alkema, W. (2013). Identification of new biomarker candidates for glucocorticoid induced insulin resistance using literature mining. *BioData mining*, 6(1), 2.

30. Fontelo, P., Liu, F., & Ackerman, M. (2005). ask MEDLINE: a free-text, natural language query tool for MEDLINE/PubMed. *BMC Medical Informatics and Decision Making*, 5(1), 5.

31. HelioBLAST. helioblast.heliotext.com

32. Fontaine, J. F., Barbosa-Silva, A., Schaefer, M., Huska, M. R., Muro, E. M., & Andrade-Navarro, M. A. (2009). MedlineRanker: flexible ranking of biomedical literature. *Nucleic acids research*, 37(suppl_2), W141-W146.

33. States, D. J., Ade, A. S., Wright, Z. C., Bookvich, A. V., & Athey, B. D. (2008). MiSearch adaptive pubMed search tool. *Bioinformatics*, 25(7), 974-976.

34. Thomas, P., Starlinger, J., Vowinkel, A., Arzt, S., & Leser, U. (2012). GeneView: a comprehensive semantic search engine for PubMed*. Nucleic acids research*, 40(W1), W585-W591.

35. Huang, K. C., Chiang, I. J., Xiao, F., Liao, C. C., Liu, C. C. H., & Wong, J. M. (2013). PICO element detection in medical text without metadata: Are first sentences enough? *Journal of biomedical informatics*, 46(5), 940-946.

36. Hokamp, K., & Wolfe, K. H. (2004). PubCrawler: keeping up comfortably with PubMed and

GenBank. *Nucleic acids research*, 32(suppl_2), W16-W19.

37. Plikus, M. V., Zhang, Z., & Chuong, C. M. (2006). PubFocus: semantic MEDLINE/PubMed citations analytics through integration of controlled biomedical dictionaries and ranking algorithm. *BMC bioinformatics*, 7(1), 424.

38. Becker, K. G., Hosack, D. A., Dennis, G., Lempicki, R. A., Bright, T. J., Cheadle, C., & Engel, J. (2003). PubMatrix: a tool for multiplex literature mining. *BMC bioinformatics*, 4(1), 61.

39. Douglas, S. M., Montelione, G. T., & Gerstein, M. (2005). PubNet: a flexible system for visualizing literature derived networks. *Genome biology*, 6(9), R80.

40. Lourenço, A., Carreira, R., Carneiro, S., Maia, P., Glez-Peña, D., Fdez-Riverola, F., ... & Rocha, M. (2009). @ Note: a workbench for biomedical text mining. *Journal of biomedical informatics*, 42(4), 710-720.

41. Cowie, J., & Wilks, Y. (2000). Information extraction. *Handbook of Natural Language Processing*, 56, 57.

42. Sarawagi, S. (2008). Information extraction. *Foundations and Trends® in Databases*, 1(3), 261-377.

43. Athenikos, S. J., & Han, H. (2010). Biomedical question answering: A survey. *Computer methods and programs in biomedicine*, 99(1), 1-24.

44. Trippe, E. D., Aguilar, J. B., Yan, Y. H., Nural, M. V., Brady, J. A., Assefi, M., ... & Kissinger, J. C. (2017). A vision for health informatics: Introducing the sked framework. an extensible architecture for scientific knowledge extraction from data. *arXiv preprint arXiv:1706.07992*.

45. Liekens, A. M., De Knijf, J., Daelemans, W., Goethals, B., De Rijk, P., & Del-Favero, J. (2011). BioGraph: unsupervised biomedical knowledge discovery via automated hypothesis generation. *Genome biology*, 12(6), R57.

46. Radev, D. R., Hovy, E., & McKeown, K. (2002). Introduction to the special issue on summarization. *Computational linguistics*, 28(4), 399-408.

47. Leser, U., & Hakenberg, J. (2005). What makes a gene name? Named entity recognition in the biomedical literature. *Briefings in bioinformatics*, 6(4), 357-369.

48. Kang, N., Singh, B., Bui, C., Afzal, Z., van Mulligen, E. M., & Kors, J. A. (2014). Knowledge-based extraction of adverse drug events from biomedical text. *BMC bioinformatics*, 15(1), 64.

49. Xu, R., & Wang, Q. (2014). Automatic construction of a large-scale and accurate drug-side-effect association knowledge base from biomedical literature. *Journal of biomedical informatics*, 51, 191-199.

50. Gurulingappa, H., Mateen-Rajpu, A., & Toldo, L. (2012). Extraction of potential adverse drug events from medical case reports. *Journal of biomedical semantics*, 3(1), 15.

51. Eltyeb, S., & Salim, N. (2014). Chemical named entities recognition: a review on approaches and applications. *Journal of cheminformatics*, 6(1), 17.

52. Naderi, N., Kappler, T., Baker, C. J., & Witte, R. (2011). OrganismTagger: detection, normalization and grounding of organism entities in biomedical documents. *Bioinformatics*, 27(19), 2721-2729.

53. Martinez, D., Pitson, G., MacKinlay, A., & Cavedon, L. (2014). Cross-hospital portability of information extraction of cancer staging information. *Artificial intelligence in medicine*, 62(1), 11-

21.

54. Nobata, C., Collier, N., & Tsujii, J. I. (1999). Automatic term identification and classification in biology texts. In *Proc. of the 5th NLPRS* (pp. 369-374).

55. Mitsumori, T., Fation, S., Murata, M., Doi, K., & Doi, H. (2005). Gene/protein name recognition based on support vector machine using dictionary as features. *BMC bioinformatics*, 6(1), S8.

56. Zhao, S. (2004). Named entity recognition in biomedical texts using an HMM model. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications* (pp. 84-87). Association for Computational Linguistics.

57. Corbett, P., & Copestake, A. (2008). Cascaded classifiers for confidence-based chemical named entity recognition. *BMC bioinformatics*, 9(11), S4.

58. Settles, B. (2004). Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications* (NLPBA/BioNLP) (pp. 107-110).

59. Brancotte, B., Biton, A., Bernard-Pierrot, I., Radvanyi, F., Reyal, F., & Cohen-Boulakia, S. (2011). Gene List significance at-a-glance with GeneValorization. *Bioinformatics*, 27(8), 1187-1189.

60. Smalheiser, N. R., Zhou, W., & Torvik, V. I. (2008). Anne O'Tate: A tool to support user-driven summarization, drill-down and browsing of PubMed search results. *Journal of biomedical discovery and collaboration*, 3(1), 2.

61. MEDIE. nactem.ac.uk/medie

62. PubReMiner. hgserver2.amc.nl/cgi-bin/miner/miner2.cgi

63. Pafilis, E., O'Donoghue, S. I., Jensen, L. J., Horn, H., Kuhn, M., Brown, N. P., & Schneider, R. (2009). Reflect: augmented browsing for the life scientist. *Nature biotechnology*, 27(6), 508.

64. Rebholz-Schuhmann, D., Arregui, M., Gaudan, S., Kirsch, H., & Jimeno, A. (2007). Text processing through Web services: calling Whatizit. *Bioinformatics*, 24(2), 296-298.

65. Gerner, M., Nenadic, G., & Bergman, C. M. (2010). LINNAEUS: a species name identification system for biomedical literature. *BMC bioinformatics*, 11(1), 85.

66. Campos, D., Matos, S., & Oliveira, J. L. (2013). Neji: a tool for heterogeneous biomedical concept identification. *Proceedings of BioLINK SIG*, 20, 1178.

67. Hakenberg, J., Gerner, M., Haeussler, M., Solt, I., Plake, C., Schroeder, M., ... & Bergman, C. M. (2011). The GNAT library for local and remote gene mention normalization. *Bioinformatics*, 27(19), 2769-2771.

68. Settles, B. (2005). ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics*, 21(14), 3191-3192.

69. Fontaine, J. F., Priller, F., Barbosa-Silva, A., & Andrade-Navarro, M. A. (2011). Genie: literature-based gene prioritization at multi genomic scale. *Nucleic acids research*, 39(suppl_2), W455-W461.

70. Bravo, A., Cases, M., Queralt-Rosinach, N., Sanz, F., & Furlong, L. I. (2014). A knowledge-driven approach to extract disease-related biomarkers from the literature. *BioMed research international*, 2014.

71. Alako, B. T., Veldhoven, A., van Baal, S., Jelier, R., Verhoeven, S., Rullmann, T., ... & Jenster, G. (2005). CoPub Mapper: mining MEDLINE based on search term co-publication. *BMC*

bioinformatics, 6(1), 51.

72. Ananiadou, S., Pyysalo, S., Tsujii, J. I., & Kell, D. B. (2010). Event extraction for systems biology by text mining the literature. Trends in biotechnology, 28(7), 381-390.

73. Coppernoll-Blach, P. (2011). Quertle: the conceptual relationships alternative search engine for pubmed. Journal of the Medical Library Association: JMLA, 99(2), 176.

74. Chen, H., & Sharp, B. M. (2004). Content-rich biological network constructed by mining PubMed abstracts. BMC bioinformatics, 5(1), 147.

75. Coremine Medical™. coremine.com

76. Tsuruoka, Y., Miwa, M., Hamamoto, K., Tsujii, J. I., & Ananiadou, S. (2011). Discovering and visualizing indirect associations between biomedical concepts. Bioinformatics, 27(13), i111-i119.

77. Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., ... & Jensen, L. J. (2016). The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic acids research, 45(D1): D362-D368.

78. Raja, K., Subramani, S., & Natarajan, J. (2013). PPInterFinder—a mining tool for extracting causal relations on human proteins from literature. Database, 2013.

79. Qiao, N., Huang, Y., Naveed, H., Green, C. D., & Han, J. D. J. (2013). CoCiter: an efficient tool to infer gene function by assessing the significance of literature co-citation. PloS one, 8(9), e74074.

80. Barbosa-Silva, A., Soldatos, T. G., Magalhães, I. L., Pavlopoulos, G. A., Fontaine, J. F., Andrade-Navarro, M. A., ... & Ortega, J. M. (2010). LAITOR - literature assistant for identification of terms co-occurrences and relationships. BMC bioinformatics, 11(1), 70.

81. Bravo, À., Piñero, J., Queralt-Rosinach, N., Rautschka, M., & Furlong, L. I. (2015). Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. *BMC bioinformatics*, 16(1), 55.

82. Miwa, M., Thompson, P., McNaught, J., Kell, D. B., & Ananiadou, S. (2012). Extracting semantically enriched events from biomedical literature. *BMC bioinformatics*, 13(1), 108.

83. Donaldson, I., Martin, J., De Bruijn, B., Wolting, C., Lay, V., Tuekam, B., ... & Pawson, T. (2003). PreBIND and Textomy–mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC bioinformatics*, 4(1), 11.

84. Krallinger, M., Leitner, F., Rodriguez-Penagos, C., & Valencia, A. (2008). Overview of the protein-protein interaction annotation extraction task of BioCreative II. *Genome biology*, 9(S2), S4.

85. Jenssen, T. K., Lægreid, A., Komorowski, J., & Hovig, E. (2001). A literature network of human genes for high-throughput analysis of gene expression. *Nature genetics*, 28(1), 21-28.

86. Narayanaswamy, M., Ravikumar, K. E., & Vijay-Shanker, K. (2005). Beyond the clause: extraction of phosphorylation information from medline abstracts. *Bioinformatics*, 21(suppl_1), i319-i327.

87. Hoffmann, R., Krallinger, M., Andres, E., Tamames, J., Blaschke, C., & Valencia, A. (2005). Text mining for metabolic pathways, signaling cascades, and protein networks. *Science's STKE*, 2005(283), pe21-pe21.

88. Koike, A., Kobayashi, Y., & Takagi, T. (2003). Kinase pathway database: an integrated protein-kinase and NLP-based protein-interaction resource. *Genome Research*, 13(6a), 1231-1243.

89. Kuhn, M., von Mering, C., Campillos, M., Jensen, L. J., & Bork, P. (2007). STITCH: interaction networks of chemicals and proteins. *Nucleic acids research*, 36(suppl_1), D684-D688.

90. Chun, H. W., Tsuruoka, Y., Kim, J. D., Shiba, R., Nagata, N., Hishiki, T., & Tsujii, J. I. (2006). Extraction of gene-disease relations from Medline using domain dictionaries and machine learning. In *Biocomputing* (pp. 4-15).

91. Pospisil, P., Iyer, L. K., Adelstein, S. J., & Kassis, A. I. (2006). A combined approach to data mining of textual and structured data to identify cancer-related targets. *BMC bioinformatics*, 7(1), 354.

92. Li, X., Chen, H., Huang, Z., Su, H., & Martinez, J. D. (2007). Global mapping of gene/protein interactions in PubMed abstracts: A framework and an experiment with P53 interactions. *Journal of biomedical informatics*, 40(5), 453-464.

93. Xuan, W., Wang, P., Watson, S. J., & Meng, F. (2007). Medline search engine for finding genetic markers with biological significance. *Bioinformatics*, 23(18), 2477-2484.

94. Gachloo, M., Wang, Y., & Xia, J. (2019). A review of drug knowledge discovery using BioNLP and tensor or matrix decomposition. *Genomics & informatics*, 17(2).

95. Krallinger, M., Leitner, F., & Valencia, A. (2010). Analysis of biological processes and diseases using text mining approaches. In *Bioinformatics Methods in Clinical Research* (pp. 341-382). Humana Press.

96. Huang, C. C., & Lu, Z. (2016). Community challenges in biomedical text mining over 10 years: success, failure and the future. *Briefings in bioinformatics*, 17(1), 132-144.

97. Uzuner, Ö., Solti, I., & Cadag, E. (2010). Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5), 514-518.

98. Uzuner, Ö., South, B. R., Shen, S., & DuVall, S. L. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5), 552-556.

99. Sun, W., Rumshisky, A., & Uzuner, O. (2013). Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *Journal of the American Medical Informatics Association*, 20(5), 806-813.

100. Uzuner, Ö., Goldstein, I., Luo, Y., & Kohane, I. (2008). Identifying patient smoking status from medical discharge records. *Journal of the American Medical Informatics Association*, 15(1), 14-24.

101. Uzuner, Ö. (2009). Recognizing obesity and comorbidities in sparse data. *Journal of the American Medical Informatics Association*, 16(4), 561-570.

102. Stubbs, A., Kotfila, C., Xu, H., & Uzuner, Ö. (2015). Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task Track 2. *Journal of biomedical informatics*, 58, S67-S77.

103. Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., ... & Liu, H. (2018). Clinical information extraction applications: a literature review. *Journal of biomedical informatics*, 77, 34-49.

104. Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., & Chute, C. G. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5), 507-513.

105. Aronson, A. R., & Lang, F. M. (2010). An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3), 229-236.

106. Friedman, C., Alderson, P. O., Austin, J. H., Cimino, J. J., & Johnson, S. B. (1994). A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association*, 1(2), 161-174.

107. Denny, J. C., Irani, P. R., Wehbe, F. H., Smithers, J. D., & Spickard III, A. (2003). The KnowledgeMap project: development of a concept-based medical school curriculum database. In *AMIA Annual Symposium Proceedings* (Vol. 2003, p. 195). American Medical Informatics Association.

108. Goryachev, S., Sordo, M., & Zeng, Q. T. (2006). A suite of natural language processing tools developed for the I2B2 project. In *AMIA Annual Symposium Proceedings* (Vol. 2006, p. 931). American Medical Informatics Association.

109. Xu, H., Stenner, S. P., Doan, S., Johnson, K. B., Waitman, L. R., & Denny, J. C. (2010). MedEx: a medication information extraction system for clinical narratives. *Journal of the American Medical Informatics Association*, 17(1), 19-24.

110. Sohn, S., Clark, C., Halgrim, S. R., Murphy, S. P., Chute, C. G., & Liu, H. (2014). MedXN: an open source medication extraction and normalization tool for clinical text. *Journal of the American Medical Informatics Association*, 21(5), 858-865.

111. Lin, Y. K., Chen, H., & Brown, R. A. (2013). MedTime: A temporal information extraction system for clinical narratives. *Journal of biomedical informatics*, 46, S20-S28.

112. Liu, H., Bielinski, S. J., Sohn, S., Murphy, S., Wagholikar, K. B., Jonnalagadda, S. R., ... & Chute, C. G. (2013). An information extraction framework for cohort identification using electronic health records. *AMIA Summits on Translational Science Proceedings*, 2013, 149.

113. NCBI Taxonomy Downloads. ftp.ncbi.nlm.nih.gov/pub/taxonomy/

114. NCBI Taxonomy. ncbi.nlm.nih.gov/taxonomy

115. Robert, V., Vu, D., Amor, A. B. H., van de Wiele, N., Brouwer, C., Jabas, B., ... & Chouchen, O. (2013). MycoBank gearing up for new horizons. *IMA fungus*, 4(2), 371-379.

116. NCBI MeSH. ncbi.nlm.nih.gov/mesh

117. Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., ... & Assempour, N. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic acids research*, 46(D1), D1074-D1082.

118. Disease Ontology GitHub Repository. hgithub.com/DiseaseOntology/HumanDiseaseOntology/tree/ master/src/ontology

119. Schriml, L. M., Mitraka, E., Munro, J., Tauber, B., Schor, M., Nickle, L., ... & Bisordi, K. (2019). Human Disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic acids research*, 47(D1), D955-D962.

120. Lourenço, A., Carreira, R., Carneiro, S., Maia, P., Glez-Peña, D., Fdez-Riverola, F., ... & Rocha, M. (2009). @Note: a workbench for biomedical text mining. *Journal of biomedical informatics*, 42(4), 710-720.

121. Shang, S. T., Yang, Y. S., & Peng, M. Y. (2010). Nosocomial *Trichosporon asahii* fungemia in a patient with secondary hemochromatosis: a rare case report. *Journal of Microbiology, Immunology and Infection*, 43(1), 77-80.

122. Mititelu, R., Bourassa-Blanchette, S., Sharma, K., & Roth, V. (2016). Angioinvasive mucormycosis

and paradoxical stroke: a case report. *JMM case reports*, 3(4).

123. MacDonald, M. L., Weiss, P. J., Deloach-Banta, L. J., & Comer, S. W. (1994). Primary cutaneous mucormycosis with a Mucor species: is iron overload a factor? *Cutis*, 54(4), 275.

124. Miyata, Y., Kajiguchi, T., Saito, M., & Takeyama, H. (2000). Development of arterial thrombus of Mucorales hyphae during deferoxamine therapy in a patient with aplastic anemia in transformation to myelodysplastic syndrome. *[Rinsho ketsueki] The Japanese journal of clinical hematology*, 41(2), 129.

125. Garcia-Vidal, C., Upton, A., Kirby, K. A., & Marr, K. A. (2008). Epidemiology of invasive mold infections in allogeneic stem cell transplant recipients: biological risk factors for infection according to time after transplantation. *Clinical Infectious Diseases*, 47(8), 1041-1050.

126. Tachibana, T., Tanaka, M., Takasaki, H., Numata, A., Ito, S., Watanabe, R., ... & Fujisawa, S. (2011). Pretransplant serum ferritin is associated with bloodstream infections within 100 days of allogeneic stem cell transplantation for myeloid malignancies. *International journal of hematology*, 93(3), 368-374.

127. Kefala-Agoropoulou, K., Roilides, E., Lazaridou, A., Karatza, E., Farmaki, E., Tsantali, H., ... & Tsiouris, J. (2007). Pearson syndrome in an infant heterozygous for C282Y allele of HFE gene. *Hematology*, 12(6), 549-553.

128. Sen, L., Goicoa, M. A., Nualart, P. J., Ballart, I. J., Palacios, F., Diez, R. A., & Estévez, M. E. (1989). Immunologic studies in thalassemia major. *Medicina*, 49(2), 131-134.

129. Steenweghen, S. V., Maertens, J., Boogaerts, M., Deneffe, G., Verbeken, E., & Nackaerts, K. (1999). Mucormycosis, a threatening opportunistic mycotic infection. *Acta Clinica Belgica*, 54(2), 99-102.

130. Marconi, V. C., Kradin, R., Marty, F. M., Hospenthal, D. R., & Kotton, C. N. (2010). Disseminated dermatophytosis in a patient with hereditary hemochromatosis and hepatic cirrhosis: case report and review of the literature. *Medical mycology*, 48(3), 518-527.

131. WINDUS, D. W., STOKES, T. J., JULIAN, B. A., & FENVES, A. Z. (1987). Fatal *Rhizopus* infections in hemodialysis patients receiving deferoxamine. *Annals of internal medicine*, 107(5), 678-680.

132. Alexander, J., Limaye, A. P., Ko, C. W., Bronner, M. P., & Kowdley, K. V. (2006). Association of hepatic iron overload with invasive fungal infection in liver transplant recipients. *Liver transplantation*, 12(12), 1799-1804.

133. Rex, J. H., Ginsberg, A. M., Fries, L. F., Pass, H. I., & Kwon-Chung, K. J. (1988). *Cunninghamella bertholletiae* infection associated with deferoxamine therapy. *Clinical Infectious Diseases*, 10(6), 1187-1194.

134. Eisen, D. P., & Robson, J. (2004). Complete resolution of pulmonary *Rhizopus oryzae* infection with itraconazole treatment: more evidence of the utility of azoles for zygomycosis. *Mycoses*, 47(3-4), 159-162.

135. Dadwal, S. S., Tegtmeier, B., Liu, X., Frankel, P., Ito, J., Forman, S. J., & Pullarkat, V. (2015). Impact of pretransplant serum ferritin level on risk of invasive mold infection after allogeneic hematopoietic stem cell transplantation. *European journal of haematology*, 94(3), 235-242.

136. PubMed Clinical Queries. ncbi.nlm.nih.gov/pubmed/clinical

137. ClinicalTrials.gov database. clinicaltrials.gov

Annex

Table 1 – Most frequent annotations, obtained with the “only-fungi”, “mycobank”, “iron-terms”, “drugs” and “diseases” lexical resources for both corpora, in the second step of NER.

	Annotated with “only-fungi” dictionary	**Annotated with “mycobank” dictionary**
only-fungi	*R. oryzae*, *C. albicans*, *A. fumigatus*, fungi, *Drosophila*, *Cryptococcus neoformans*, *Candida*, *S. chartarum*, *P. carinii*, *S. cerevisiae*	-
mycobank	-	bacteria, necrosis, *R. oryzae*, *C. albicans*, *A. fumigatus*, *C. glabrata*, *Plasmodium*, *Candida albicans*, fungi, *Drosophila*
iron-terms	iron overload, hemochromatosis, iron accumulation	iron overload, hemochromatosis, iron accumulation
drugs	as, al, Fig, Iron, DFO, yeast, mg, *Candida albicans*, copper, deferoxamine	as, al, Fig, Iron, mg, DFO, Fe^{2+}, com, As, yeast
diseases	disease, can, all, mucormycosis, asm, All, Mucormycosis, anemia, schistosomiasis, March	can, disease, all, All, tuberculosis, malaria, anemia, asm, mucormycosis, March

Table 2 – Examples of terms with different meanings, present in the different lexical resources used in the second step of NER.

Term	**Lexical resource**	**Meaning in the resource**	**Meaning in the text**
Drosophila	only-fungi, mycobank	Genus of fungus	Genus of a fly
melanogaster	only-fungi, mycobank	Genus of fungus	*Drosophila melanogaster* (model organism, fly)
P. aeruginosa	only-fungi, mycobank	*Pholiotina aeruginosa*	*Pseudomonas aeruginosa* (bacterium)
necrosis	mycobank	Genus of fungus	Disease
Plasmodium	mycobank	Genus of fungus	Genus of the malaria parasite
C. elegans	only-fungi, mycobank	*Callistosporium elegans*, *Chaetothyriothecium elegans*, *Corallomycetella elegans*, *Conocybe elegans*, *Canalisporium elegans*, *Cymatoderma elegans*, *Cyrenella elegans*, *Cylindrocladiella elegans*	*Caenorhabditis elegans* (model organism, nematode)
Xenopus	mycobank	Genus of fungus	Genus of a frog
omega	mycobank	Genus of fungus	Greek letter
all	diseases	Acute Lymphoblastic Leukemia	all
can	diseases	Crouzon syndrome-acanthosis nigricans syndrome	can
al	drugs	Aluminium	As in *et al*
as	drugs	Artesunate	as
Fig	drugs	Fig	Figure
mg	drugs	Glyceryl 1-oleate	Microgram
ng	drugs	Nitroglycerin	Nanogram

Table 3 – Brief summary of the relevant publications obtained.

Reference	**Summary**
[121]	A 53-year-old nongranulocytopenic female with secondary hemochromatosis developed nosocomial fungemia caused by *Trichosporon asahii*. This case suggests that clinicians should be aware that *T. asahii* fungemia can develop in nongranulocytopenic patients with secondary hemochromatosis.
[122]	A 33-year-old female with transfusion-dependent beta thalassemia was started on intravenous deferiprone therapy and subsequently *Rhizopus* species were present in her blood. This case illustrates that angioinvasive *Rhizopus* infections can occur in patients treated with deferiprone, a newer iron chelator agent that has antifungal properties *in vivo*.
[123]	A severely debilitated patient showed primary cutaneous mucormycosis with a *Mucor* species at a tape erosion site. Iron overload may be a risk factor for mucormycosis. The pathogenic nature and epidemiologic features of this unusual fungal infection are reviewed.
[124]	A 58-year-old woman with a diagnosis of aplastic anemia became dependent on red blood cell transfusions. Deferoxamine was administered for iron overload. The patient later developed pneumonia and pulmonary mycosis. Although an antifungal agent was administered, the patient experienced respiratory failure and eventually died. Deferoxamine has been used in the treatment of iron overload patients with aplastic anemia, and may be a risk factor for mucormycosis. There may be a relationship between mucormycosis and deferoxamine in patients with aplastic anemia.
[125]	Invasive mould infections are common in patients who have undergone hematopoietic stem cell transplantation. Clinical and biological risk factors for different types of invasive mould infections after allogeneic hematopoietic stem cell transplantation were studied. Of a total of 1248 patients, 13.1% received a diagnosis of probable or proven invasive mould infection. The majority of cases were caused by *Aspergillus* species (88%). Iron overload is an important biological risk factor.

[126]	The association between iron overload and bloodstream infections following allogeneic hematopoietic stem cell transplantation for acute myeloid leukemia or myelodysplastic syndromes was retrospectively studied in 114 adult patients who underwent transplantation. In conclusion, pretransplantation serum ferritin significantly predicts bloodstream infections within a 100-day period after the transplantation.
[127]	A female infant suffering from anemia since birth, with iron overload disproportionate to blood transfusions, also suffered from type I hemochromatosis. After an initial response to deferoxamine, she presented with cutaneous zygomycosis and died after metabolic derangement and *Pneumocystis jiroveci* pneumonia.
[128]	Continuous transfusion-related antigenic stimulation together with iron overload cause immune alterations in 10-year-old or younger patients with thalassemia major. The immune status of thalassemia major patients was evaluated, and alterations regarding white blood cells were found, such as altered B-cell function, dysfunction of T immunoregulatory cells and defective NK activity observed, which are independent of the patients' age and are attributed to blood transfusions. The capacity of phagocytes to ingest *Candida* is preserved, while the candidacidal activity and the generation of toxic oxygen metabolites during the respiratory burst are diminished, and are inversely proportional with age and serum ferritin concentration, meaning that the older the patient and the higher their iron overload, the more dysfunctional are the phagocytes.
[129]	A patient with severe aplastic anemia (with neutropenia) and hemochromatosis (treated with desferrioxamine), complicated with a systemic *Rhizopus infection*, did not survive despite treatment with amphotericin B and granulocyte-colony-stimulating factors.

[130]	A case of biopsy-proven, disseminated dermatophytosis caused by *Trichophyton rubrum* in a patient with hereditary hemochromatosis and hepatic cirrhosis is described. Over the course of the hospitalization, the dermatophytosis progressed to a more invasive form with widespread cutaneous dissemination. His risk factors for invasive fungal disease included cirrhosis and iron overload associated with hemochromatosis. Ultimately, he died from his underlying pneumonia, which prevented any conclusions to be taken regarding the efficacy of the antifungal therapy.
[131]	Four hemodialysis patients receiving deferoxamine for metal overload had fatal rhinocerebral *Rhizopus* infections. Serious fungal infections are not commonly seen in patients on dialysis, and none of these patients had the usual risk factors for *Rhizopus* infection. Deferoxamine is being used with increased frequency in dialysis patients for aluminum and iron overload states. It is proposed that there is a link between the deferoxamine therapy and this unusual infection. Deferoxamine may serve as a specific growth factor for *Rhizopus* species or may alter host immune function. Searching for fungal organisms in patients with unexplained illnesses receiving deferoxamine is suggested.
[132]	A cohort of 153 consecutive patients who underwent liver transplantation and who survived at least 7 days after transplantation was retrospectively studied. The association between various pretransplant patient characteristics, including hepatic explant iron and risk of invasive fungal infections, was analysed. During the first year after transplantation, 28 of 153 patients developed a total of 31 invasive fungal infections, of which 21 (68%) were caused by *Candida*, 7 (23%) by *Aspergillus*, 2 (6%) by *Cryptococcus*, and 1 (3%) by *Saccharomyces*. In conclusion, our study found that hepatic iron overload is an independent risk factor for posttransplantation invasive fungal infections in liver transplant recipients. Patients with iron overload might benefit from closer monitoring for invasive fungal infections and/or targeted antifungal prophylaxis. Moderate iron deprivation before liver transplantation could be an additional approach for reducing the risk of posttransplantation invasive fungal infections and improving the posttransplantation outcome in iron-loaded patients with cirrhosis.

[133]	A patient receiving deferoxamine for iron overload unrelated to hemodialysis developed a *Cunninghamella bertholletiae* infection. Possible explanations for the patient's infection were assessed.
[134]	A non-neutropenic patient with myelodysplastic syndrome and iron overload receiving cytotoxic therapy presented with pulmonary *Rhizopus oryzae* infection. This patient was cured through the use of itraconazole alone and the literature on the utility of azole antifungals for zygomycosis is reviewed.
[135]	Invasive mould infections are life-threatening complications of allogeneic hematopoietic stem cell transplantation. The association between elevated serum ferritin prior to hematopoietic stem cell transplantation and the increased risk of invasive mould infections was studied in a large cohort of patients who had undergone allogeneic hematopoietic stem cell transplantation.