Exhaustive Indexing of PubMed Records with Medical Subject Headings

Modest von Korff
Idorsia Pharmaceuticals Ltd.
Hegenheimermattweg 91, 4123 Allschwil, Switzerland
modest.korff@idorsia.com

Abstract

With fourteen million publication records the PubMed database is one of the largest repositories in medical science. Analyzing this database to relate biological targets to diseases is an important task in pharmaceutical research. We developed a software tool, MeSHTreeIndexer, for indexing the PubMed medical literature with disease terms. The disease terms were taken from the Medical Subject Heading (MeSH) Terms compiled by the National Institutes of Health (NIH) of the US. In a first semi-automatic step we identified about 5'900 terms as disease related. The MeSH terms contain so-called entry points that are synonymously used for the terms. We created an inverted index for these 5'900 MeSH terms and their 58'000 entry points. From the PubMed database fourteen million publication records were stored in Lucene. These publication records were tagged by the inverted MeSH term index. In this contribution we demonstrate that our approach provided a significant higher enrichment in MeSH terms than the indexing of the PubMed records by the NIH themselves. Manual control proved that our enrichment is meaningful.

Keywords: Text mining, MeSH, PubMed, Indexing

1. Introduction

The starting point for drug discovery is to find a new biological target to cure a disease. As always in drug discovery, first step is analysing the related medical literature. Therefore, searching the medical literature for diseases is a common task. Searching the life science literature for diseases belongs in the category of biomedical named entity recognition. With the increasing amount of information the interest in indexing biomedical information becomes of more and more interest. The largest repository for medical literature is provided by the National Institutes of Health (NIH) of the US (NIH, 2022b). In January 2022 the PubMed database contained 33 million records for biomedical literature compiled from MEDLINE, life science journals, and online books. For searching and indexing the medical records in PubMed the NIH developed a thesaurus, the Medical Subject Headings (NIH, 2022; Lipscomb, 2000). These MeSH terms are organized in a tree with 16 main branches. Branch A contains terms from anatomy, branch B lists organisms, branch C is dedicated to diseases, branch D lists chemicals and drugs, branch E structures analytical diagnostic and therapeutic techniques, and branch F organizes terms from psychiatry and psychology. The following branches contain terms from phenomena and processes (G), disciplines and occupations (H), anthropology (I), technology (J), humanities (K), information science (L), named groups (M), health care (N), publication characteristics (V), and geographicals (Z). For medical literature other indexing systems exist as well. Widely used is SNOMED, a collection of clinical terminology to represent patient data for clinical purposes (Ruch et al., 2008). Health insurances use a disease classification system 'International Classification of Diseases (ICD)', version 11 (World Health Organization, 2016). However, neither SNOMED nor ICD were intended to capture content of scientific literature. The MeSH terms were derived from the scientific literature in life sciences. The NIH index semi-automatically the PUBMED records with MeSH terms. Manual annotation processes are labour-intensive. It is a fact that the indexing is delayed and incomplete because of the amount of publications in life sciences and limited resources (Hadfield, 2020; Irwin and Rackham, 2017).

2. Related work

Because of the high importance for research in life science, automatic indexing systems for MeSH terms were developed. They can be classified into three categories: 1) pattern matching, 2) text classification, 3) learning-to-rank. From all software tools to be named, MetaMap (Aronson and Lang, 2010) was developed first by the US National Library of Medicine. MetaMap applies pattern matching to the unified medical language system UMLS. UMLS are not MeSH terms, but closely related. Indexing MeSH terms is PubTator, a web based indexing system also developed by the US National Library of Medicine (Wei et al., 2013). PubTator uses DNorm (Leaman et al., 2013) to tag PubMed articles with MeSH disease terms. DNorm is based on a pairwise learning-to-rank algorithm. Learning-to-rank algorithms make use of identified nearest neighbour documents to retrieve the most relevant MeSH terms (Huang et al., 2011). A more recent approach combines several machine learning techniques (Mao and Lu, 2017). Convolutional neural networks are used in (Gargiulo et al., 2019; Dai et al., 2020). MeSHLabeler uses a combination of Medical Text Indexer, pattern matching, and indexing rules (Liu

Proceedings of the CMLC-10 Workshop (@LREC 2022), pages 8–15
Marseille, 20 June 2022

et al., 2015). For all MeSH indexing must be considered that it is a complex task. Even between human indexers only 48.2% consistency was reported for main heading assignment (Funk and Reid, 1983). This unsatisfying consistency is easily explained by the aim of the indexing approaches. All here mentioned approaches aimed to index the literature with the most important concepts. But what are the most important concepts? This is often hard to recognize from the publication alone. Because, after a manuscript was published the relevance of its content depends on the context of the reader. The same publication has different meanings for two scientists studying different subjects.

3. Our work

Our goal was to index exhaustively a corpus of 14 million PubMed records with disease related MeSH terms. In contrary to all other approaches mentioned in the section above, we aimed to index every occurrence of a term. The major concept of a publication did not matter for us. The indexed corpus was intended to be analyzed for co-occurrences of index tags. A ranking of concepts was not intended. For this reason we decided to use a non-machine-learning approach for indexing. After analyzing the structure of the MeSH tree we realized that the information in the MeSH tree together with the entry terms would be sufficient for our needs. A node in the MeSH tree is labeled by a descriptor, e.g. diabetes mellitus, type 1. Additionally, so-called entry terms are given. These terms are synonyms, alternate forms, and other closely related terms that are generally used interchangeably with the descriptor term. For diabetes mellitus, type 1, 27 entry terms are given. These are terms which are alternatively used in life science literature. The alternative terms may differ only in a hyphen, order of words, complete different synonyms, or they are abbreviations. On average there are about ten entry points for each disease MeSH term. These entry points, in the following named as entry terms, represent the major forms of writing for a disease term in the life science literature. Our idea was to search for these entry points in the corpus of 14 million PubMed records. In the following section it is shown how we overcame many of the issues for text matching in medical literature, as it was discussed by Díaz and López in (Díaz and López, 2015).

4. Methods

4.1. Medical Subject Headings

For drug discovery purposes of interest are disease related terms. As described above, the MeSH tree contains 16 main branches. Two of these branches were used for disease indexing. The disease branch C and branch F, with terms from psychiatry and psychology. From these two branches unspecific expressions were removed. This included terms used as common words. These excluded expressions form a so-called stoplists. Stoplists for indexing disease MeSH terms were introduced by (Swanson et al., 2006). We loosely orientated our disease term collection at Swanson's stoplist. Very general disease terms were removed. Mainly entries from the psychology branch were removed, terms like affect, behavior, and aptitude. Additionally we took a list of the most common English words (Kaufman, 2017) to further exclude disease terms that are frequently used. Although, some common words are also important disease terms. So, a whitelist with needed disease terms was created. This whitelist contains disease terms like arthritis, measles and cholera. MeSH nodes are not unique in the MeSH tree. Because of the structure of the tree the same node can occur at more than one position. For example 'Gaucher Disease' is in branch 'Central Nervous System Diseases', 'Genetic Diseases, Inborn', 'Metabolic Diseases', and in other branches. The number of non-unique MeSH nodes sum up to 13'969. Finally, the disease MeSH tree contains 5'904 unique disease-related nodes. These nodes contain 63'072 entry points. In Table 1 a histogram is shown that represents the distribution of the entry points on the unique MeSH nodes.

Min bin	1	5	10	20	50
Max bin	5	10	20	50	1000
Counts	2161	1520	1357	771	94

Table 1: Histogram of the number of entry points in the disease MeSH terms

The MeSH terms and the entry terms were normalized for indexing. It is important to notice for our algorithm that the MeSH entry phrases contain the original terms in different orders, as they occur in the publications. Also alternatives with different punctuation marks, spelling variants and acronyms are given as entry terms. The 32 entry terms for the MeSH descriptor 'Diabetes Mellitus, Type 2' are given as example. The normalization procedure is described below in more detail for the PubMed records.

- Adult-Onset Diabetes Mellitus
- Diabetes Mellitus, Adult Onset
- Diabetes Mellitus, Adult-Onset
- Diabetes Mellitus, Ketosis Resistant
- Diabetes Mellitus, Ketosis-Resistant
- Diabetes Mellitus, Maturity Onset
- Diabetes Mellitus, Maturity-Onset
- Diabetes Mellitus, Non Insulin Dependent
- Diabetes Mellitus, Non-Insulin-Dependent
- Diabetes Mellitus, Noninsulin Dependent
- Diabetes Mellitus, Noninsulin-Dependent
- Diabetes Mellitus, Slow Onset
- Diabetes Mellitus, Slow-Onset
- Diabetes Mellitus, Stable

- Diabetes Mellitus, Type 2
- Diabetes Mellitus, Type II
- Diabetes, Maturity-Onset
- Diabetes, Type 2
- Ketosis-Resistant Diabetes Mellitus
- MODY
- Maturity Onset Diabetes
- Maturity Onset Diabetes Mellitus
- Maturity-Onset Diabetes
- Maturity-Onset Diabetes Mellitus
- NIDDM
- Non-Insulin-Dependent Diabetes Mellitus
- Noninsulin Dependent Diabetes Mellitus
- Noninsulin-Dependent Diabetes Mellitus
- Slow-Onset Diabetes Mellitus
- Stable Diabetes Mellitus
- Type 2 Diabetes
- Type 2 Diabetes Mellitus

These entry terms are text phrases frequently occurring in medical literature. After being normalized and stemmed, the number of terms reduced to 23.

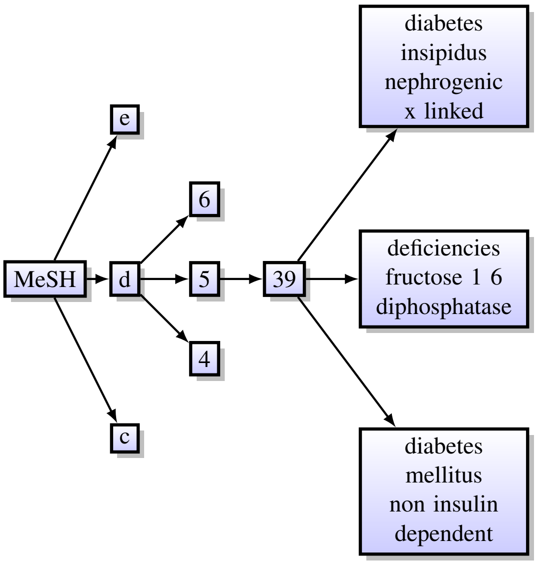
- adult onset diabetes mellitus
- diabetes maturity onset
- diabetes mellitus adult onset
- diabetes mellitus ketosis resistant
- diabetes mellitus maturity onset
- diabetes mellitus non insulin dependent
- diabetes mellitus noninsulin dependent
- diabetes mellitus slow onset
- diabetes mellitus stable
- diabetes mellitus type 2
- diabetes mellitus type ii
- diabetes type 2
- ketosis resistant diabetes mellitus
- maturity onset diabetes
- maturity onset diabetes mellitus
- mody
- niddm
- non insulin dependent diabetes mellitus
- noninsulin dependent diabetes mellitus
- slow onset diabetes mellitus
- stable diabetes mellitus
- type 2 diabetes
- type 2 diabetes mellitus

To index a publication with the MeSH descriptor 'Diabetes Mellitus, Type 2' one of the 23 entry terms must be found in the text record. Therefore, the MeSH entry terms overcome many of the problems that were described by Díaz and López in (Díaz and López, 2015).

4.2. The MeSH term index tree

Apache Lucene was used to store the PubMed records, MeSH terms and the entry terms. Lucene is a widely used open-source database and text search engine (Białecki et al., 2012). Already, many string matching algorithms exist to search text indices. The best performing algorithms rely on preprocessing of a dictionary. So, a powerful key-word matching algorithm was developed by Aho and Corasick (Aho and Corasick, 1975). Their algorithm constructs a finite state pattern matching machine from the keywords. The keywords are processed once and the text is matched against the processed keywords. Another dictionary oriented approach used a modified Levenshtein distance for high-throughput spelling correction (Schulz and Mihov, 2002). Lucene contained the MeSH term index and is capable of text matching. So, in our first approach we tested the fuzzy search for multiple word terms in Lucene for indexing. Indexing performance was a major criteria for our algorithm. However, the indexing performance of Lucene was too low for the corpus of 14 million text records. It was also difficult to fine tune the fuzzy string matching. The matching criteria tended either be too coarse or too fine for a successful comparison. This was due to the structure of the MeSH terms, containing abbreviations and numbers. After several days of proprietary experiments with Lucene we decided to implement our own algorithm for performance reasons, and for better control of the matching terms. Our approach is similar to (Aho and Corasick, 1975), we also decided to implement a pre-processed tree-structure. With the difference, that the design of our algorithm was much simpler. It was tailored for our purpose to match medical subject headings. Medical subject headings are standardized phrases used in medical literature. They are collected in the MeSH terms and the entry terms. To prevent mismatches, matching a term needs to be exact. Only small typos or spelling variants may be accepted for a match. MeSH terms have significant meaning in medical literature. Typos at the beginning of a MeSH term phrase are not common. For this reason we decided that typos in the first character of a MeSH term result in a no-match. From the normalized entry terms a tree-based index was created and implemented as a list of lists. The first level of the tree represents the starting character of the normalized entry term. Numbers from 1 to 9 and lowercase letters a-z occur. The second list represents the number of tokens in the normalized entry term. As for the starting letter the number of tokens needs an exact match. A third list indexes the string length of the entry term. In the last list, the leaf node in the MeSH term index tree, the categorized normalized terms, are stored. As example: The MeSH descriptor 'Diabetes Mellitus, Type 2' contains the entry term 'Diabetes Mellitus, Non Insulin Dependent'. After being normalized the entry term becomes 'diabetes mellitus non insulin dependent'. First character of the normalized term is ‘d’, which equals index 39 in tree level one. The normalized entry term contains five tokens, index five in the second layer of the index tree. Finally, a length of 39 characters for the normalized entry term results in index 39 in tree level three. This node has three children: ‘diabetes mellitus non insulin dependent’, ‘deficiencies fructose 1 6 diphosphatase’, and ‘diabetes insipidus nephrogenic x linked’. The normalized expressions occur without punctuation, hyphens and always in lowercase letters 1.

Figure 1: Part of the MeSH term index tree. Level one: starting letter of the MeSH term, level two: number of tokens, level three: number of characters, level four: MeSH term.


diabetes insipidus nephrogenic x linked
e
6
MeSH
d
5
39
deficiencies fructose 1 6 diphosphatase
4
c
diabetes mellitus non insulin dependent

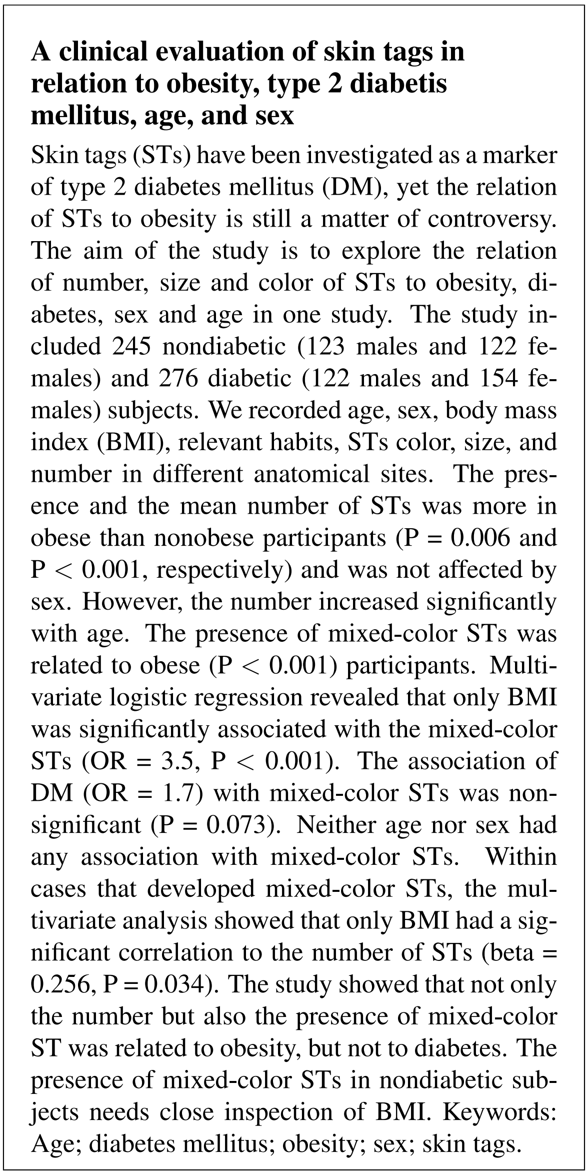
4.3. Searching for MeSH terms in PubMed records

PubMed records contain a lot of information, but they do not contain the full publication text. Publication title, abstract, and the PubMed identifier (PMID) were used for our algorithm. An example is given in Figure 2.

Searching a PubMed record for normalized MeSH terms starts by splitting the text into sentences, Figure 3. Searching is followed by stemming with the Apache OpenNLP library. The pre-processed phrases are tokenized. Every non-literal, Greek letters, written out Greek letters, and numbers directly attached to words are tokenized. Uppercase letters are converted into lowercase. Except, an uppercase letter is followed by another uppercase or by a number. There is no extra treatment for floating point numbers. Stop words are removed.

Publication title and abstract were used for MeSH term matching. Every phrase is parsed by the MeSH term index tree. Parsing is done phrase-wise with a sliding window of increasing size for the tokens in the phrase. Parsing starts with the first character of the first token in the sentence. The node in level one that corresponds to the first character is the starting point for further parsing. Level two of the index tree corresponds to the number of tokens in the phrase. The number of tokens for the start phrase is 1. Level three of the index node corresponds to the number of characters in the phrase to analyze. All terms in the level three index nodes from minus three characters up to plus three characters are compared with the phrase to analyze. The comparison is a two step process. A first string match checks for misleading similarities. Misleading similarities are calculated from word pairs that differ by one or two characters but have a complete different meaning, e.g. injection and infection. If the word

Figure 2: PubMed record for PMID 21965846. Part of the record, used for indexing. The typo in the title ‘diabetis’ instead of ‘diabetes’ is from the original publication.



A clinical evaluation of skin tags in relation to obesity, type 2 diabetis mellitus, age, and sex

Skin tags (STs) have been investigated as a marker of type 2 diabetes mellitus (DM), yet the relation of STs to obesity is still a matter of controversy. The aim of the study is to explore the relation of number, size and color of STs to obesity, diabetes, sex and age in one study. The study included 245 nondiabetic (123 males and 122 females) and 276 diabetic (122 males and 154 females) subjects. We recorded age, sex, body mass index (BMI), relevant habits, STs color, size, and number in different anatomical sites. The presence and the mean number of STs was more in obese than nonobese participants (P = 0.006 and P $<$ 0.001, respectively) and was not affected by sex. However, the number increased significantly with age. The presence of mixed-color STs was related to obese (P $<$ 0.001) participants. Multivariate logistic regression revealed that only BMI was significantly associated with the mixed-color STs (OR = 3.5, P $<$ 0.001). The association of DM (OR = 1.7) with mixed-color STs was nonsignificant (P = 0.073). Neither age nor sex had any association with mixed-color STs. Within cases that developed mixed-color STs, the multivariate analysis showed that only BMI had a significant correlation to the number of STs (beta = 0.256, P = 0.034). The study showed that not only the number but also the presence of mixed-color ST was related to obesity, but not to diabetes. The presence of mixed-color STs in nondiabetic subjects needs close inspection of BMI. Keywords: Age; diabetes mellitus; obesity; sex; skin tags.

Figure 3: Normalized and stemmed PubMed record for PMID 21965846. A period indicates the end of a phrase detected by the stemming algorithm.

> clinical evaluation skin tags relation obesity type 2 diabetis mellitus age sex. skin tags STs investigated marker type 2 diabetes mellitus DM relation STs obesity matter controversy. aim study explore relation number size color STs obesity diabetes sex age study. study included 245 nondiabetic 123 males 122 females 276 diabetic 122 males 154 females subjects. recorded age sex body mass index BMI relevant habits STs color size number different anatomical sites. presence mean number STs obese nonobese participants P 0 006 P 0 001 respectively affected sex. number increased significantly age. presence mixed color STs related obese P 0 001 participants. multivariate logistic regression revealed BMI significantly associated mixed color STs OR 3 5 P 0 001. association DM OR 1 7 mixed color STs nonsignificant P 0 073. age sex association mixed color STs. cases developed mixed color STs multivariate analysis showed BMI significant correlation number STs beta 0 256 P 0 034. study showed number presence mixed color ST related obesity diabetes. presence mixed color STs nondiabetic subjects needs close inspection BMI. keywords age diabetes mellitus obesity sex skin tags.

pair passes this test the similarity is calculated by the Damerau-Levenshtein algorithm. For phrase comparison with more than one token the comparison is done token by token. If the token pair similarity is below 0.75 the phrase is dissimilar. This threshold allows a small change in a word, i.e. the change of a single letter. This takes into account the morphological or orthographic variations of scientific writing. If the comparison matches the threshold, the average from all token pair comparisons in the phrase are calculated. If the average similarity is equal or above 0.85 the PubMed record is tagged with the MeSH descriptor corresponding to the matching phrase. MeSH entry terms are not necessarily unique, one matching phrase may result in two tags. An example is given with the string ”Background: Skin tags (STs) have been investigated as a marker of type 2 diabetes mellitus (DM), yet the relation of STs to obesity is still a matter of controversy”, PMID 21965846. The string is parsed after normalization. The following items demonstrate how the string is parsed with the sliding token window.

- ‘background‘ → no match
- ‘background skin‘ → no match
- ‘background skin sts‘ → no match
- ...proceed up to maximum term length
- ‘skin‘ → no match
- ‘skin sts‘ → no match
- ...
- ‘type‘ → no match
- ‘type 2‘ → no match
- ‘type 2 diabetes mellitus‘ → match
- ‘diabetes‘ → no match
- ‘diabetes mellitus‘ → match

The sentence is tagged with two MeSH descriptors ‘Diabetes Mellitus’ and ‘Diabetes Mellitus, Type 2’. With this procedure all 14 million PubMed records were indexed with the matching MeSH term descriptors.

4.4. Implementation

The PubMed records were retrieved from the MEDLINE database and stored in Lucene (Białecki et al., 2012). The normalized PubMed records were also stored in Lucene. The MeSH term indexer was implemented in Java 11. The NIH MeSH tree was taken from the file mtrees2022.bin (NIH, 2022a). This file was serialized and stored. Entry terms were read on the fly from desc2022.xml. The disease MeSH tree was compiled from the serialized MeSH tree file, the descriptor file and the hard-coded stoplists. Index tags were written to the normalized records in Lucene.

5. Results

5.1. Results MeSH term index

A detailed view into the structure of the MeSH term index tree is given in the following section. As mentioned above, level one of the MeSH term index tree corresponds to the starting characters of the MeSH entry terms. In total, 58’774 unique MeSH entry terms were indexed in the MeSH term index tree. In Table 2 the counts for every starting character for all unique MeSH entry terms are shown. Some MeSH entry terms start with a number between 1 and 9, but none starts with a zero. The other characters are well distributed over the alphabet.

Char	Counts	Char	Counts	Char	Counts
1	19	e	2356	p	5693
2	11	f	2153	q	49
3	6	g	1506	r	1769
4	29	h	3475	s	5643
5	8	i	2984	t	3076
6	2	j	241	u	545
7	2	k	408	v	1155
a	4898	l	2274	w	430
b	2193	m	3620	x	143
c	5447	n	2690	y	35
d	4583	o	1268	z	63

Table 2: Counts for each starting character of all MeSH terms

Level two of the MeSH term index tree represents the number of tokens for each MeSH entry term. A token in a normalized MeSH term can be a single character, number, or letter. The distribution for the token count in all MeSH entry terms is given in Table 3. Even the bin with the maximum number of tokens still contains 210 normalized MeSH entry terms. Here, the counts follow a broadly skewed distribution.

Num	Counts	Num	Counts	Num	Counts
1	2730	6	2625	11	210
2	2835	7	1995	12	315
3	3150	8	1470	13	105
4	3465	9	840	14	210
5	2835	10	630		

Table 3: Counts for number of tokens in a term

For level three of the MeSH term index tree the distribution is given as a graph in Figure 4. This level encodes the string length of the MeSH entry terms. Again, the distribution is broad, this time nearly Gaussian.

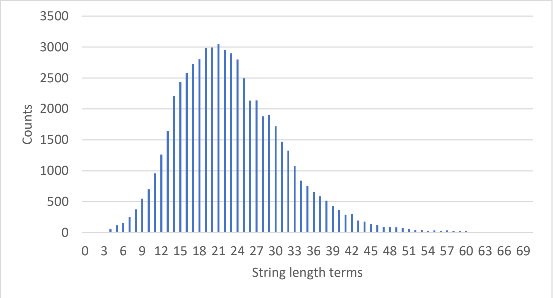


Figure 4: Length distribution of MeSH terms

In level four of the MeSH term index tree it is revealed that the divide and conquer strategy of the index tree was successful. The histogram in Table 4 shows the distribution for the number of MeSH entry terms in the leaf nodes of the index tree. In the first bin the list size is 1 for 1’002 leaf nodes, in the second bin the list size is between 2 and 5 for 871 leaf nodes. So, the majority of leaf nodes contains small lists.

Min bin	1	2	5	10	20	50
Max bin	2	5	10	20	50	150
Counts	1’002	871	512	392	379	310

Table 4: Histogram of the number of terms in a list

5.2. Results tagging fourteen million PubMed records

The corpus to index contained fourteen million (14’138’576) PubMed records. These were records which were previously found by querying PubMed for gene names. Only twelve million records contained a summary, additionally to the title. The number of tokens indexed by Lucene summed up to 2.8 billion tokens. MeSH term indexing for all records took around three days with a performance of around 4’000 records per minute. A file was compiled which listed all disease MeSH terms together with the PubMed identifiers where this disease term was found (*diseaseList*). A sample of 10’000 indexed PubMed records was analyzed for detail. The sample was drawn from *diseaseList* by random sampling technique with a limit of ten records per disease. Resulting, the 10’000 records represented 1’056 diseases from *diseaseList*. All records contained at least one disease MeSH term from the MeSH term index tree. In total, 61’413 MeSH terms were found by the MeSH term index tree. For 857 records no MeSH term was retrieved from PubMed. A sum of 25’982 MeSH terms was retrieved from PubMed. The MeSH term index tree did not tag 3’387 of these MeSH terms. The results were summarized in Table 5. The result file for the 10’000 records is available on request from the author.

MeSH terms	Found	Not found	Overlap
PubMed	25’982	38’818	22’595
Index tree	61’413	3’387	

Table 5: Counts for MeSH terms found in 10’000 PubMed records. Index tree for ‘MeSH term index tree’

These results revealed a high discrepancy between the indexing by the NIH and the MeSH term index tree. So, we took a close look to single records. After sorting the 10’000 sample records by disease terms the first disease term was ‘Abdomen, Acute’. The first record had the title ‘Patient factors influencing the effect of surgeon-performed ultrasound on the acute abdomen’, PubMed Id 21290005, from year 2010 in Critical Ultrasound Journal. No PubMed MeSH terms were given. The MeSH term index tree indexed the record with the disease MeSH terms ‘Abdomen, Acute’, ‘Abdominal Pain’, ‘Appendicitis’, and ‘Peritonitis’. As it can be taken from the record summary in InfoBox 1 all terms occur in the text.

An example with overlap between the two indexing methods and where the NIH indexing exclusively tagged a MeSH term is record PID 18294294. The PubMed MeSH terms ‘Leukemia, Lymphoid’, and ‘Recurrence’ were not tagged by the MeSH term index tree. Both methods found ‘Lymphoma, Extranodal NK-T-Cells’ in the text. The MeSH term index tree tagged exclusively the record with ‘Dis-

PURPOSE: To evaluate the effect of surgeon-performed ultrasound on acute abdomen in specific patient subgroups regarding the diagnostic accuracy and further management. METHODS: Eight hundred patients attending the emergency department at Stockholm South General Hospital, Sweden, for abdominal pain, were randomized to either receive or not receive surgeon-performed ultrasound as a complement to routine management. Patients were divided into subgroups based on patient characteristics. [...] Timing of surgery was evaluated for patients with peritonitis. [...] Decreased need for further examinations and/or fewer admissions were seen in all groups except in patients with a preliminary diagnosis of appendicitis. [...]

InfoBox 1: Part of summary for PubMed record with Id 21290005

ease Resistance', 'Glycogen Storage Disease Type VI', 'Leukemia', 'Leukemia, large Granular Lymphocytic', 'Lymphoma', 'Lymphoproliferative Disorders', 'Neoplasms', 'Neutropenia', 'Precursor T-Cell Lymphoblastic Leukemia-Lymphoma', and 'Sepsis'. Obviously, 'Leukemia, Lymphoid' is a meta term, given from the NIH index crew. And the NIH index crew skipped the leaf node tags 'Leukemia, large Granular Lymphocytic' and 'Precursor T-Cell Lymphoblastic Leukemia-Lymphoma'. Also not considered by the NIH were the meta tags 'Leukemia', 'Lymphoma', and 'Neoplasms'. The NIH summarized these tags with the meta tags. And our MeSH term index tree missed these meta tags, because no lexicographic matching pattern was found in the PubMed record. The two terms 'Sepsis' and 'Neutropenia' found from the MeSH term index tree were not tagged by the NIH. Presumably, because the tags are related to only one patient out of six. However, neutropenia, sepsis and chemotherapy have a causal relation (Ba et al., 2020). But if we are searching the PubMed MeSH terms for relations between neoplasms, sepsis, and neutropenia we will miss this publication.

6. Summary and conclusions

A new method to index PubMed records exhaustively with disease MeSH terms was developed and applied to a corpus of 14 million PubMed records. A random sample with 10'000 indexed records was analyzed in detail. In this sample 8.6% of the records were not indexed in PubMed. Why so many records were not indexed by the NIH, is under examination. Additionally, the MeSH term index tree found 2.4 times more MeSH terms than given in the PubMed records. This can partly explain that the NIH indexing aims to index with the most meaningful tags. And, the NIH indexing crew summarizes concepts with MeSH terms that are closer to the root of the MeSH term index tree. These summarizing MeSH terms explain the ten percent of the MeSH terms in PubMed that were not tagged by the MeSH term index tree. The NIH indexing includes a ranking of concepts, our approach is unbiased. Together, the two systems provide a large basis for information extraction from PubMed. Thus, in future work, the combination of the two index systems will be tested as input for machine learning systems to find new relations between diseases. The MeSH term index tree is highly precise, it only accepts records that match entry MeSH terms that were defined by the NIH. Our approach is unbiased. It does not need any training records. Only a minimum set of parameters is needed. Rough mismatches were excluded by defining the list of false similar terms. False similar term pairs are lexicographically similar but possess a different meaning. Also, common words are excluded from matching. The few parameters and very general rules make our algorithm highly reliable. The four level MeSH term index tree is very well balanced, this results in a very high indexing performance. We are convinced that our algorithm supports the scientific community in indexing life science literature and plan to provide the source code as open source project on git hub.

7. Bibliographical References

Aho, A. V. and Corasick, M. J. (1975). Efficient string matching: an aid to bibliographic search. *Communications of the ACM*, 18(6):333–340.

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

Ba, Y., Shi, Y., Jiang, W., Feng, J., Cheng, Y., Xiao, L., Zhang, Q., Qiu, W., Xu, B., Xu, R., et al. (2020). Current management of chemotherapy-induced neutropenia in adults: key points and new challenges: committee of neoplastic supportive-care (cons), china anti-cancer association committee of clinical chemotherapy, china anti-cancer association. *Cancer Biology & Medicine*, 17(4):896.

Białecki, A., Muir, R., Ingersoll, G., and Imagination, L. (2012). Apache lucene 4. In *SIGIR 2012 workshop on open source information retrieval*, page 17.

Dai, S., You, R., Lu, Z., Huang, X., Mamitsuka, H., and Zhu, S. (2020). Fullmesh: improving large-scale mesh indexing with full text. *Bioinformatics*, 36(5):1533–1541.

Díaz, N. P. C. and López, M. J. M. (2015). An analysis of biomedical tokenization: problems and strategies. In *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis*, pages 40–49.

Funk, M. E. and Reid, C. A. (1983). Indexing consistency in medline. *Bulletin of the Medical Library Association*, 71(2):176.

Gargiulo, F., Silvestri, S., Ciampi, M., and De Pietro, G. (2019). Deep neural network for hierarchical extreme multi-label text classification. *Applied Soft Computing*, 79:125–138.

Hadfield, R. M. (2020). Delay and bias in pubmed medical subject heading (mesh) indexing of respiratory journals. *medRxiv*.

Huang, M., Névéol, A., and Lu, Z. (2011). Recommending mesh terms for annotating biomedical articles. *Journal of the American Medical Informatics Association*, 18(5):660–667.

Irwin, A. N. and Rackham, D. (2017). Comparison of the time-to-indexing in pubmed between biomedical journals according to impact factor, discipline, and focus. *Research in Social and Administrative Pharmacy*, 13(2):389–393.

Kaufman, J. (2017). Most common english words. `https://github.com/first20hours/google-10000-english`.

Leaman, R., Islamaj Doğan, R., and Lu, Z. (2013). Dnorm: disease name normalization with pairwise learning to rank. *Bioinformatics*, 29(22):2909–2917.

Lipscomb, C. E. (2000). Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265.

Liu, K., Peng, S., Wu, J., Zhai, C., Mamitsuka, H., and Zhu, S. (2015). Meshlabeler: improving the accuracy of large-scale mesh indexing by integrating diverse evidence. *Bioinformatics*, 31(12):i339–i347.

Mao, Y. and Lu, Z. (2017). Mesh now: automatic mesh indexing at pubmed scale via learning to rank. *Journal of biomedical semantics*, 8(1):1–9.

NIH. (2022a). Mesh tree. `https://nlmpubs.nlm.nih.gov/projects/mesh/MESH_FILES/meshtrees/`.

NIH. (2022b). Pubmed entry site. `https://pubmed.ncbi.nlm.nih.gov/`.

NIH. (2022). Medical subject headings. `https://www.nlm.nih.gov/bsd/disted/meshtutorial/introduction/02.html`.

Ruch, P., Gobeill, J., Lovis, C., and Geissbühler, A. (2008). Automatic medical encoding with snomed categories. In *BMC medical informatics and decision making*, volume 8, pages 1–8. BioMed Central.

Schulz, K. U. and Mihov, S. (2002). Fast string correction with levenshtein automata. *International Journal on Document Analysis and Recognition*, 5(1):67–85.

Swanson, D. R., Smalheiser, N. R., and Torvik, V. I. (2006). Ranking indirect connections in literature-based discovery: The role of medical subject headings. *Journal of the American society for information science and technology*, 57(11):1427–1439.

Wei, C.-H., Kao, H.-Y., and Lu, Z. (2013). Pubtator: a web-based text mining tool for assisting biocuration. *Nucleic acids research*, 41(W1):W518–W522.

World Health Organization. (2016). International classification of diseases. 2016. `https://www.who.int/standards/classifications/classification-of-diseases`.