Mining MEDLINE for problems associated with vitamin D

Dina Demner-Fushman, MD, PhD, James G. Mork, MSc, Alan R. Aronson, PhD,
Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine, National Institutes of Health, DHHS, Bethesda, MD

Abstract

This paper presents a two-step approach to generating comprehensive abstractive overviews for biomedical topics. It starts with a sensitivity-maximizing search of MEDLINE/PubMed and MeSH-based filtering of the results that are then processed using NLP methods to extract relations between entities of interest. We evaluate this approach in a case study based on the IOM report on the role of vitamin D in human health. The report defines disorders that serve as health indicators for the role of vitamin D. We evaluate the abstractive overviews generated using MeSH indexing and the extracted relations using the disorders listed in the IOM report as reference standard. We conclude that MeSH-based aggregation and filtering of the results is a useful and easy step in the generation of abstractive overviews. Although our relation extraction achieved 83.6% recall and 92.8% precision, only half of the disorders of interest participated in these relations.

Introduction

Establishing the state of the current knowledge on the topic of interest is an essential and time-consuming step in biomedical research and authoring of guidelines and systematic reviews. This step traditionally starts with sensitivity-maximizing searches for topically relevant publications followed by a thorough manual review of the abstracts and selection of the full text publications for an in-depth review. The sensitivity-maximizing searches for MEDLINE® usually retrieve broad ranges of potentially relevant citations through combining fairly long lists of search terms in Boolean OR queries[1]. Whereas search engines find potentially relevant publications, creating an overview of retrieved articles largely remains a manual task. Efforts to reduce the burden of a full review of the results of high sensitivity searchers follow two major directions: reducing the workload using statistical methods and creating overviews of the retrieval results[2,3]. Approaches to automatic generation of the overviews in the form of abstractive and multi-document summarization have been explored in the past. These approaches included automatic Unified Medical Language System® (UMLS®)-based hierarchical agglomerative clustering with semantic distance as the link function, in which potential treatments for a given disease were extracted from the abstracts and assigned each to its own cluster, which were then iteratively merged into larger clusters whose interventions shared a common UMLS hypernym[4]. Each cluster was represented by the highest ranking sentence containing "bottom-line" advice extracted from the highest ranking abstract in the cluster. In an evaluation of a similar final representation of the overviews generated using a summarizer in the semantic abstraction paradigm and SemRep, the knowledge based summarizer modestly outperformed a baseline which extracted all frequently occurring drugs in retrieved documents[5]. Clustering of extracted named entities and sentences containing the entities was also used to summarize information about gene studies in microarray experiments[6,7] and breast cancer[8].

In addition to titles and abstracts, MEDLINE provides abstractive summaries of the articles in the form of MeSH (Medical Subject Headings) indexing. MeSH indexing is widely used in the analysis of biomedical literature. For example, it was used in the original literature-based discovery experiments[9] and later in MeSHmap, an application for interactive exploration of PubMed® search results[10]. In MeSHmap, the summary of the retrieval results is provided in the form of agglomerated separate lists of headings and subheadings along with the frequencies of their occurrences in the result set. We extend the MeSHmap approach by automatically utilizing the relational information encoded in subheadings and focusing on the main ("starred") headings. We further explore if the relations inferred based on the subheadings are supported by extraction of relations explicitly stated in the titles and abstracts. To that end, we use a previously developed approach to extraction of clinical events[11]. The goal of this study is to evaluate a two-step approach to generating comprehensive abstractive overviews for topics of interest. Our proposed approach first agglomerates and filters retrieval results using the inferred relations between MeSH terms and then extracts relations between entities of interest from the text of MEDLINE citations comprising the reduced document set. To test our approach, we focus on disorders associated with vitamin D, taking disorders listed in the 2010 Institute of Medicine (IOM) report[12] as the reference set.

Background

For many adults (including the authors of this study), a routine checkup has revealed a vitamin D deficiency followed by the doctor's recommendation to take a vitamin D supplement. It is not surprising therefore that the benefits and dangers of the long term use of vitamin D supplements are of broad interest to the public. Given the growing interest in vitamin D, the U.S. and Canadian governments asked the IOM to conduct a review of current scientific evidence pertaining to roles of calcium and vitamin D in human health and to update the Dietary Reference Intakes (DRI) for vitamin D and calcium. The IOM established an ad hoc consensus committee of 14 scientists who conducted a review of existing data, and developed a report[12] that serves as the reference standard for our study. The committee has identified a comprehensive set of potential indicators for establishing DRIs shown in Table 1. The committee concluded that with the exception of measures related to bone health, the potential indicators were not supported by evidence. The outcomes related to other indicators were often conflicting and could not be linked to vitamin D intake. The report describes the role of vitamin D related to falls and physical performance, cardiovascular disease, autoimmune disorders, and immune functioning as hypotheses of emerging interest. Our study uses this information to a) establish if information about the indicators could have been summarized for the committee using MeSH headings, b) if the relations extracted from the abstracts suggest the conclusions about the bone health outcomes and the other indicators, and if c) our methods are able to identify diseases presented as associated with high levels of vitamin D in the report (see Table 2). The side effects of the high doses of vitamin D are of particular concern to those who take vitamin D supplements.

Table 1. Potential indicators of health outcomes for nutrient adequacy for calcium and vitamin D and their mapping to MeSH.

IOM Report	MeSH	MeSH Tree	Specificity
Total cancer	Neoplasms	C04	General
Prostate cancer	Prostatic Neoplasms	C04, C12	
Colorectal cancer/adenoma	Colorectal Neoplasms	C04, C06	
Breast cancer	Breast Neoplasms	C04, C17	
Pancreatic cancer	Pancreatic Neoplasms	C04, C06, C19	
Immune function clinical outcomes			
Autoimmune disease	Autoimmune Diseases	C20	
Diabetes (type 1)	Diabetes Mellitus, Type 1	C18, C19, C20	
Crohn's disease / Inflammatory bowel disease	Crohn Disease	C06	
Multiple sclerosis	Multiple sclerosis	C10, C20	
Rheumatoid arthritis	Arthritis, Rheumatoid	C05, C17, C20	
Infectious diseases	Communicable Diseases	C01	
Preeclampsia	Pre-Eclampsia	C13	
pregnancy outcomes	Pregnancy Outcome	F01, G08	General
All-cause mortality	Disease*/mortality		General
Cardiovascular diseases	Cardiovascular Abnormalities	C14, C16	General
Hypertension	Hypertension	C14	
Obesity/ metabolic syndrome	Obesity	C18, C23, C01, G07	
Growth	Growth Disorders	C23	General
Bone health clinical outcomes	Bone Diseases	C05	General
Fracture risk	Fractures, Bone	C26	General
Rickets	Rickets	C05, C18	
osteomalacia	Osteomalacia	C05, C18	

Table 2. Disorders mentioned in association with high levels of Vitamin D and their mapping to MeSH.

IOM Report	MeSH	MeSH Tree	Specificity
hypercalcemia	Hypercalcemia	C18	
Hypercalciuria/ Calcium phosphate crystals in urine	Hypercalciuria	C23	
retarded growth	Growth Disorders Bone Diseases, Developmental	C23 C05	General
All cause mortality	Disease*/mortality		General
cancer	Neoplasms	C04	General
cardiovascular risk/damage	Cardiovascular Abnormalities	C14, C16	General
falls	Accidental Falls	N06	General
fractures	Fractures, Bone	C26	General
soft tissue calcification/ ligamentous calcification	Calcinosis	C18	General
renal damage	Kidney Diseases	C12, C13	General
kidney stones	Kidney Calculi	C12, C13, C23	
nephrocalcinosis	Nephrocalcinosis	C12, C13, C18	
renal failure	Renal Insufficiency	C12, C13	
Renal colic	Renal Colic	C23	
polyuria	Polyuria	C12, C13	
Albuminuria	Albuminuria	C12, C13, C23	
fibrotic changes in vascular tissue/ arterial calcification	Vascular calcification	C18	
corneal calcification	Corneal Diseases	C11	General
Conjunctivitis	Conjunctivitis	C11	
Edema	Edema	C23	General
Anorexia	Anorexia	C23	
weight loss	Weight Loss	C23	
Asthenia/ weakness/ Fatigue/ tiredness/muscular weakness	Asthenia	C23	
Anemia	Anemia	C15	
Vomiting	Vomiting	C23	
Nausea	Nausea	C23	
Fever	Fever	C23	
Chills	Chills	C23	
Lethargy	Lethargy	C10	
Extreme thirst/ polydipsia	Polydipsia	C23	
Dehydration	Dehydration	C18,C23	
Leg pain	Pain	C10	general
back pain	Back Pain	C10, C23	
Constipation	Constipation	C23	
Colic	Colic	C23	
Diarrhea	Diarrhea	C23	
persistent hypertension	Hypertension	C14	
Coma	Coma	C10, C23	
Headache	Headache	C10, C23	
Confusion	Confusion	C10, C23, F01	
memory loss/ forgetfulness	Memory Disorders	C10, C23, F01	
psychotic symptoms	Mental Disorders	F03	

Methods

To answer our questions about the coverage of the IOM indicators by MeSH indexing we generated high sensitivity PubMed searches that resulted in a set of 44,961 MEDLINE citations and analyzed the aggregated MeSH indexing as described below. We then used MeSH indexing to identify a subset of citations that could potentially answer our questions about toxicity of the high doses of vitamin D and analyzed these citations using the Stanford dependency parser[13] and extracting causal, treatment and association relations between vitamin D and disorders.

High sensitivity search for publications concerning vitamin D

Using the search terms found in the IOM *Table E-1 Sample Search History for Literature Published after AHRQ-Tufts*, we identified a list of MeSH terms to search via PubMed/Entrez and a list of trigger words to search the 2013 MEDLINE Baseline.

We were able to identify the following MeSH terms from the IOM list of search terms: *Ergocalciferols*, *25-Hydroxyvitamin D 2*, *Cholecalciferol*, *Calcifediol*, *Vitamin D*. We then expanded this list to include *Calcitriol* (physiologically active form of vitamin D), *Dihydrotachysterol* (product of vitamin D2), and *Hydroxycholecalciferols* (previous indexing for *Calcifediol*) providing us with the PubMed query found in Figure 1. In the query, we use "[mh:noexp]" to make sure PubMed/Entrez does not expand the query beyond our targeted list of MeSH Headings.

```
"Ergocalciferols" [mh:noexp] OR "25-Hydroxyvitamin D 2" [mh noexp] OR "Cholecalciferol"
[mh:noexp] OR "Calcifediol" [mh:noexp] OR "Vitamin D" [mh:noexp] OR "Calcitriol"
[mh:noexp] OR "Dihydrotachysterol" [mh noexp] OR "Hydroxycholecalciferols" [mh noexp]
```

Figure 1: PubMed Query 1

We also noticed that the MeSH Heading *Dehydrocholesterols* was used to index citations discussing *Cholecalciferols* in 1966, so we created another PubMed query found in Figure 2. Since *Dehydrocholesterols* was used to index multiple terms over the years, we also limited the search to citations where the precursor *7-dehydrocholesterol* was found somewhere in the title or abstract. The citations from this query were added to our dataset.

```
Dehydrocholesterols [mh:noexp] AND "7-dehydrocholesterol" [tiab]
```

Figure 2: PubMed Query 2

We noticed that we had coverage of vitamins D1 through 4, but, not vitamin D5, so to add some coverage of vitamin D5, we used the precursor *7-dehydrositosterol* (required in the title or abstract) and the MeSH Heading *Sitosterols* (Heading Mapped To for *7-dehydrositosterol*) in the PubMed query found in Figure 3. The citations from this query were also added to our dataset.

```
Sitosterols [mh:noexp] AND "7-dehydrositosterol" [tiab]
```

Figure 3: PubMed Query 3

We filtered out citations from our dataset that were not indexed yet, were duplicate PMIDs from the three queries, had no indexing applied, or were identified as PubMed-not-MEDLINE leaving us with 36,078 citations.

To further expand our coverage of potentially relevant citations, we used the list of trigger words developed from the IOM table, the MeSH terms, and the associated Entry Terms from each of the MeSH descriptors. We performed a string search through the 21,508,439 citations in the 2013 MEDLINE Baseline using this list of trigger words. The search identified 8,883 additional citations for a total of 44,961 citations.

MeSH-based abstractive summarization

Once we created the dataset of relevant vitamin D citations, we could then identify articles of interest using the MeSH indexing. We tracked all of the cases where one or more of our ten vitamin D related MeSH terms was determined by the indexer to be a main topic of an article (indicated by a star "*") and where one or more Disease related MeSH terms from the *C* MeSH tree (that contains almost all disorders in Tables 1 and 2) were also identified as a main topic of the article. We kept track of this co-occurrence of MeSH terms and also tracked the MeSH qualifiers (subheadings) assigned to the MeSH descriptors. This data set consisting of 12,271 citations allowed us to summarize the results and try to identify possible patterns in the data.

Extraction of associations between diseases and vitamin D

From the above summarized data, we picked *Ergocalciferols* as an appropriate vitamin D MeSH term (MH) for a more detailed analysis of side effects. We created a subset of our dataset where *Ergocalciferols* was a main topic in the citation and where any Disease was also a main topic in the citation. We further filtered the citations by requiring that either *Ergocalciferols* be assigned the qualifier *adverse effects*, *poisoning*, or *toxicity*; or one of the starred Disease terms was assigned the *adverse effects*, *poisoning*, *toxicity*, or *complications* qualifier. This created a dataset with 211 citations that we could review for any causal, treatment or association relationships between the *Vitamin D* MH and the *Disease* MH. These citations were also used to automatically extract relations between vitamin D and diseases from the title and abstract. To extract the relations, we used the Stanford parser to identify causal, treatment, and association dependencies between all surface representations of vitamin D and all concepts in the Semantic group Disorders identified using MetaMap[14].

PMID- 4162109
TI - Effect of condensed phosphates on vitamin D-induced aortic calcification in rats.

MH - Aortic Diseases/*pathology
MH - Calcinosis/*chemically induced
MH - Ergocalciferols/*toxicity

Figure 4: PMID: 4162109 with title and partial MeSH Indexing

The example in Figure 4 shows the title and some of the MeSH indexing for one of the 211 citations in this dataset. Because *Ergocalciferols* and *Aortic Diseases* are identified as main topics in the article and the *Ergocalciferols* MH is combined with the qualifier *toxicity*, the citation was automatically included in our dataset of 211 citations for further evaluation.

Evaluation

We manually intersected disorders listed in Tables 1 and 2 with the starred diseases extracted from the MeSH indexing and with the relations extracted from the text. For the latter, the relation type had to be correct for the extracted disease to count as true positive. For example, relation (*causes: vitamin D, aortic calcification*) had to be extracted from the title in Figure 4. In our manual analysis of 211 citations, we extracted all causal, treatment and association relations involving vitamin D and explicitly stated in one sentence. If either the vitamin D term or the disease term was represented by an anaphor (see Figure 5), the relation was included in the gold standard, although our current extraction tool cannot resolve anaphora or coreference.

PMID- 21786580
AB - Calcitriol is important in nephroprotective strategy in chronic disease of the kidneys (CDK). However, **its** long-term use often results in hypercalciemia with metastatic calcification.
Relation – Calcitriol CAUSES hypercalciemia

Figure 5: PMID: 21786580, a snippet of the abstract showing the anaphor *its* referring to calcitriol and the relation included in the reference standard

Results

Using the main (starred) MeSH descriptors, we were able to identify all 22 indicators listed in Table 1 with frequencies shown in Table 3. The 211 citations dataset contained 10 of 22 potential indicators. The relations

extracted from the abstracts also contained 10 potential indicators, nine of which are the same as extracted using MeSH indexing.

Table 3. Potential indicators of health outcomes for nutrient adequacy for calcium and vitamin D. Frequency of occurrence in MeSH descriptors (third column) and in relations extracted from the text of 211 citations (fourth column).

IOM Report	MeSH	Star MeSH Frequency		Frequency in text
		Starred	211	
Total cancer	Neoplasms	2,706	20	2 (ovarian, prostatic)
Prostate cancer	Prostatic Neoplasms	306	4	1
Colorectal cancer/adenoma	Colorectal Neoplasms	110	0	
Breast cancer	Breast Neoplasms	323	0	
Pancreatic cancer	Pancreatic Neoplasms	38	0	
Immune function clinical outcomes				
Autoimmune disease	Autoimmune Diseases	62	0	
Diabetes (type 1)	Diabetes Mellitus, Type 1	102	0	2
Crohn's disease / Inflammatory bowel disease	Crohn Disease	41	0	
Multiple sclerosis	Multiple Sclerosis	110	1	
Rheumatoid arthritis	Arthritis, Rheumatoid	79	0	
Infectious diseases	Communicable Diseases	6	0	
Preeclampsia	Pre-Eclampsia	2		
pregnancy outcomes	Pregnancy Outcome	3	0	
All-cause mortality	Disease*/mortality	274	14	
Cardiovascular diseases	Cardiovascular Abnormalities	718	18	7
Hypertension	Hypertension	142	1	1
Obesity/ metabolic syndrome	Obesity	132	0	
Growth	Growth Disorders	18	0	1
Bone health clinical outcomes	Bone Diseases	285	8	12
Fracture risk	Fractures, Bone	414	6	1
Rickets	Rickets	669	12	4
osteomalacia	Osteomalacia	284	9	7

Our manual analysis of the 211 citations dataset revealed that 40 citations had no abstracts, and their titles contained no explicitly stated relations, another 39 abstracts contained relations that could not be extracted without inference and understanding of complex linguistic phenomena. For example, readers could potentially infer that *paricalcitol* has caused *hypercalcemia* in some patients looking at the snippet in Figure 6, but the two statements are too complex for the existing bioNLP tools: Extracting this relation would require associating the increase in serum calcium levels with hypercalcemia and resolving ellipses to associate hypercalcemia and patients involved in the trial presented in the paper. The remaining 132 citations contained 171 relations that needed to be extracted (71 causal, 64 treatment, and 36 associated_with), of which our extraction method has correctly identified 143 (achieving 83.6% recall) , missed 28, and generated 11 false positive relations (achieving precision 92.8%).

PMID- 11576883
AB - … paricalcitol increased serum calcium levels and decreased PTH and bone alkaline phosphatase levels (all $P < 0.05$). However, hypercalcemia was infrequent.

Figure 6: PMID: 11576883, a snippet of the abstract

Most frequently extracted causal relations were with hypercalcemia, hyperphosphatemia, nephrocalcinosis, vascular calcification and arteriosclerosis, Most frequently extracted treatment relations were with secondary hyperparathyroidism, hypovitaminosis D, rickets, and osteomalacia. The remaining relations were primarily extracted once. Of the disorders associated with high levels of vitamin D (listed in Table 2), the following 21 were extracted correctly: hypercalcemia, hypercalciuria, growth retardation, cardiovascular damage, cancer, risk of falls, fractures, calcinosis, kidney damage, nephrocalcinosis, renal failure, albuminuria, vascular calcification, edema, weight loss, vomiting, nausea, fever, chills, constipation, hypertension. Whereas the following 20 were not extracted: kidney stones, renal colic, polyuria, corneal calcification, conjunctivitis, anorexia, asthenia, anemia, lethargy, polydipsia, dehydration, leg pain, back pain, colic, diarrhea, coma, headache, confusion, memory loss, and psychotic symptoms. Some of the additional correctly extracted disorders are close to the ones listed in the IOM report, for example, musculoskeletal pain, band keratopathy, and impaired mobility. Some (shown in Table 4) were not discussed in the IOM report.

MeSH indexing of the starred set covered all but three disorders associated with high levels of vitamin D. The missing disorders were: renal colic, chills, and polydipsia. The frequencies of the starred headings associated with high doses of vitamin D range from 1 to triple digits. The MeSH indexing in the 211 citations dataset contained 20 disorders associated with high doses of vitamin D, and missed 21.

Table 4. Side effects of vitamin D extracted from the text of MEDLINE citations in addition to side effects listed in the IOM report

Disorder	Example Sentences
Hearing loss/ deafness	Deafness due to hypervitaminosis D [PMID: 87663]
palpitations	Adverse events associated with paricalcitol use included, among others, chills, feeling unwell, fever, sepsis, palpitations, dry mouth, gastrointestinal bleeding, nausea, vomiting, edema, light-headedness, and pneumonia. [PMID: 10321413]
dry mouth	
gastrointestinal bleeding	
sepsis	
pneumonia	

Discussion

Agglomerating potential indicators of health outcomes associated with vitamin D using MeSH indexing has perfect recall of the indicators studied in the IOM report, but using these results will require reviewing more indicators, as can be seen in Table 5. Restricting the set of citations to those with vitamin D and at least one disease as main topic has reduced the set of publications from 44,961 to 12,271 without losing the perfect recall for the IOM indicators. The workload of reviewing all potential indicators could be reduced by establishing frequency thresholds; however, as can be seen in Table 3, higher thresholds will eliminate some of the indicators of interest, such as preeclampsia and growth disorders.

Filtering of the retrieved set by specific subheadings indicative of side effects reduced the set to manageable size; however, it also demonstrated a traditional recall-precision tradeoff, reducing the number of found disorders to 20 out of 41 disorders associated with high doses of vitamin D in the MeSH indexing and to 21 disorders extracted from relations with vitamin D. The remaining 20 disorders were missed for several reasons: 1) due to errors in the extraction (for example, myopathy was missed in PMID: 11715587 due to coreference); 2) because they were not mentioned in the abstract (for example, headache); 3) because their relation with vitamin D was not stated explicitly (see Figure 6), or 4) because no causal relation between vitamin D and the disease was discussed in the abstract. For example, citation PMID: 22422535 discussed safety of vitamin D in patients with kidney stones and vitamin D deficiency. The disorders found and missing in the MeSH indexing are somewhat different compared to those extracted from the relations stated in the text: hypercalciuria, risk of falls, edema, fever, chills, and constipation were found in the text but not in the MeSH indexing; whereas, kidney stones, asthenia, anemia, pain, and back pain were found in the MeSH indexing but not in the text. These differences could potentially be used to increase recall by using the union of the results. Alternatively, the intersection of the results will generate a smaller set of diseases with higher confidence.

Despite our focus on the subheadings indicative of adverse reactions, the set of 211 citations contained almost equal amounts of causal and treatment relations involving vitamin D. A substantial number of association relations and

extraction of both causal and treatment relations for the same diseases (see Figure 7) reproduces the IOMs conclusions about controversial and inconclusive results for some disorders.

PMID- 18393917
AB - …It was demonstrated that paricalcitol prevents vascular calcification in experimental models of renal failure…

PMID- 4162109
TI - Effect of condensed phosphates on vitamin D-induced aortic calcification in rats.

Figure 7: Controversial conclusions about the role of vitamin D in vascular calcification

Table 5. Overall and distinct frequencies of MeSH terms pertaining to Disorders in 12,271 starred citations and 44,961 containing information about vitamin D

Disease MeSH Tree	MeSH Tree	Frequency		Distinct Disorders	
		Full set	Starred	Full set	Starred
Bacterial Infections and Mycoses	C01	774	234	138	64
Virus Diseases	C02	414	144	65	35
Parasitic Diseases	C03	56	16	26	10
Neoplasms	C04	7,963	2,706	331	180
Musculoskeletal Diseases	C05	14,984	3,296	183	89
Digestive System Diseases	C06	2,832	801	148	75
Stomatognathic Diseases	C07	505	82	100	35
Respiratory Tract Diseases	C08	1,020	356	102	49
Otorhinolaryngologic Diseases	C09	182	39	42	19
Nervous System Diseases	C10	2,599	617	273	119
Eye Diseases	C11	302	76	67	23
Male Urogenital Diseases	C12	7,907	2,375	118	73
Female Urogenital Diseases and Pregnancy Complications	C13	8,093	2,327	174	98
Cardiovascular Diseases	C14	2,574	712	144	68
Hemic and Lymphatic Diseases	C15	1,362	352	132	62
Congenital, Hereditary, and Neonatal Diseases and Abnormalities	C16	2,990	693	240	90
Skin and Connective Tissue Diseases	C17	3,681	1,526	188	116
Nutritional and Metabolic Diseases	C18	20,978	5,906	179	111
Endocrine System Diseases	C19	7,053	1,843	105	62
Immune System Diseases	C20	2,470	846	114	72
Disorders of Environmental Origin	C21	0	0	0	0
Animal Diseases	C22	1,295	298	39	21
Pathological Conditions, Signs and Symptoms	C23	8,090	840	352	147
Occupational Diseases	C24	22	5	9	3
Substance-Related Disorders	C25	413	84	39	21
Wounds and Injuries	C26	2,856	486	69	36

Although the recall and precision in extraction of relations containing diseases seem relatively high, these results are due to frequent occurrences of a small set of well-known relations such as (*treats: ergocalciferol, vitamin D insufficiency*) or (*causes: vitamin D insufficiency, rickets*). This approach demonstrates a typical trade-off between using associations and NLP approaches: we gain confidence due to the explicit statements, compared to relations inferred using MeSH indexing, but at a cost of losing almost half of the diseases of interest.

Conclusion

This paper presents a MeSH indexing-based approach to generation of abstractive overviews of broad topics in biomedical domain. We tested the feasibility of the approach in a case study based on the IOM report on current scientific evidence pertaining to roles of calcium and vitamin D in human health. To determine if MeSH based agglomerated summaries cover all potential health indicators studied in the IOM report, we first retrieved over

44,000 MEDLINE citations using a sensitivity oriented search, then reduced the set by filtering out citations that did not have vitamin D and at least one disease as main topics (indicated by starred MeSH terms), and finally focused on potential adverse events related to vitamin D by filtering out citations that did not have specific subheadings associated with the main headings of interest to our study. Filtering out non-starred citations proved to be useful in preserving the perfect recall of the sensitive search and reducing the set of document to slightly over 12,000 citations. The subheading-based filtering and extraction of the relations between the diseases and vitamin D lost almost half of the diseases associated with side effects of vitamin D. Future work will determine if the relations could be extracted from the full text of the articles or if a less restrictive filtering will provide abstracts with explicitly stated relations with more diseases. Overall, we can recommend broad searches focused on main headings as a useful and easy step in generation of abstractive overviews of the current state of knowledge for a topic of interest.

Data sets described in this paper are available at:

http://ii.nlm.nih.gov/TestCollections/index.shtml#2013_VITD

Acknowledgments

This work was supported by the intramural research program of the U. S. National Library of Medicine, National Institutes of Health.

References

1. Wilczynski NL, McKibbon KA, Walter SD, Garg AX, Haynes RB. MEDLINE clinical queries are robust when searching in recent publishing years. J Am Med Inform Assoc. 2013 Mar-Apr;20(2):363-8.
2. Cohen AM, Clive E. Adams CE, Davis JM, Yu CT, Yu PS, Meng W, Duggan L, McDonagh M, Smalheiser NR. Evidence-based medicine, the essential role of systematic reviews, and the need for automated text mining tools. IHI 2010: 376-380
3. Bekhuis T, Demner-Fushman D. Screening nonrandomized studies for medical systematic reviews: a comparative study of classifiers. Artif Intell Med. 2012 Jul;55(3):197-207.
4. Lin J, Demner-Fushman D. Semantic clustering of answers to clinical questions. AMIA Annu Symp Proc. 2007 Oct 11:458-62.
5. Fiszman M, Demner-Fushman D, Kilicoglu H, Rindflesch TC. Automatic summarization of MEDLINE citations for evidence-based medical treatment: a topic-oriented evaluation. J Biomed Inform. 2009 Oct;42(5):801-13.
6. Yang J, Cohen A, Hersh W. Evaluation of a gene information summarization system by users during the analysis process of microarray datasets. BMC Bioinformatics. 2009 Feb 5;10 Suppl 2:S5.
7. Jelier R, Schuemie MJ, Veldhoven A, Dorssers LC, Jenster G, Kors JA. Anni 2.0: a multipurpose text-mining tool for the life sciences. Genome Biol. 2008;9(6):R96.
8. Lin Y, Li W, Chen K, Liu Y. A document clustering and ranking system for exploring MEDLINE citations. J Am Med Inform Assoc. 2007 Sep-Oct;14(5):651-61.
9. Swanson DR. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspectives in Biology and Medicine. 1986. 30:7-18.
10. Srinivasan P. MeSHmap: a text mining tool for MEDLINE. Proc AMIA Symp. 2001:642-6.
11. Demner-Fushman D, Abhyankar S. Syntactic-Semantic Frames for Clinical Cohort Identification Queries. Lecture Notes in Computer Science. 2012;7348:100-112
12. Institute of Medicine. Dietary Reference Intakes for Calcium and Vitamin D. Washington, D.C.: National Academies Press, 2010.
13. de Marneffe M.-C, Manning C. Technical report. Stanford University; 2012. Stanford typed dependencies manual. http://nlp.stanford.edu/downloads/dependencies_manual.pdf Accessed March 10, 2013.
14. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc. 2010 May-Jun;17(3):229-36.