# A bottom-up approach to MEDLINE indexing recommendations

**Antonio Jimeno-Yepes, PhD[1], Bartłomiej Wilkowski, MS[2], James G. Mork, MS[1], Elizabeth Van Lenten, PhD[1], Dina Demner Fushman, MD, PhD[1], Alan R. Aronson, PhD[1]**
**[1] National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894, USA**
**[2] Technical University of Denmark, DTU Informatics, Richard Petersens Plads, B321, DK-2800, Kongens Lyngby, Denmark**

## Abstract

*MEDLINE indexing performed by the US National Library of Medicine staff describes the essence of a biomedical publication in about 14 Medical Subject Headings (MeSH). Since 2002, this task is assisted by the Medical Text Indexer (MTI) program. We present a bottom-up approach to MEDLINE indexing in which the abstract is searched for indicators for a specific MeSH recommendation in a two-step process. Supervised machine learning combined with triage rules improves sensitivity of recommendations while keeping the number of recommended terms relatively small. Improvement in recommendations observed in this work warrants further exploration of this approach to MTI recommendations on a larger set of MeSH headings.*

## Introduction

The NLM indexing process involves analysis of journal articles for subject matter and subsequent assignment of appropriate subject headings, drawn from MeSH, the NLM controlled vocabulary. Maintaining the quality of MEDLINE® indexing is made difficult by the demand of the ever increasing size of the biomedical literature on a relatively small group of highly qualified indexing contractors and staff at the US National Library of Medicine (NLM). We hope that the situation can be eased through improvements to the recommendations made by NLM's indexing tool, the Medical Text Indexer (MTI) [1,2].

MTI is a support tool for assisting indexers as they add MeSH® indexing to MEDLINE citations. MTI has two main components: MetaMap and the PubMed® Related Citations (PRC) algorithm. MetaMap performs an analysis of the citations and annotates them with Unified Medical Language System (UMLS)® concepts. Then, the mapping from UMLS to MeSH follows the *Restrict-to-MeSH* [3] approach which is based primarily on the semantic relationships among UMLS concepts. The PRC [4] algorithm is a modified k-NN algorithm which relies on document similarity to assign MeSH headings. This method intends to increase the recall of MetaMap by proposing indexing candidates for MeSH headings which are not explicitly present in the citation but have a similar context.

There are 25,588 descriptors or main headings (MHs) in 2010 MeSH from which MTI recommends about 25 terms per article, on average. Based on the results for 142,262 citations processed by MTI between November 23, 2009 and February 8, 2010, the number of MHs we need to review for possible improvements in MTI recommendations is significantly smaller than 25,588. Figure 1 illustrates the breakout of the different MHs that can be removed. There are 12,350 MHs in the "B" (*Organisms*) and "D" (*Chemicals and Drugs*) MeSH trees (recommended automatically if the terms are found in the title), 1,854 MHs for which MTI recommendations using the current top-down approach, which exploits domain knowledge, are fairly accurate (precision over 60%), 251 MHs that are used for cataloging and other purposes but not for indexing, 3,609 MHs occurring less than 500 times in MEDLINE, and 832 MHs that are too general for indexing journal articles (e.g. *Accidents*, these MHs are identified by annotations like "im gen only" or "gen; prefer specific"). The remaining 6,692 (26.15% of the 2010 MeSH) need improved recommendations.

Our previous attempts to improve the quality of recommendations [5,6] indicate that the current, top-down method might be approaching the upper bound on its performance, and other methods need to be explored to improve recommendations for the remaining 26% of the headings.

The motivation for this work comes from an approach suggested by indexers who use certain *indicators* in the articles that lead to assignment of specific indexing terms which might complement MTI annotation. In our work, we intend to improve MTI's MEDLINE MeSH recommendations by targeting ones where MTI's performance is poor. We describe

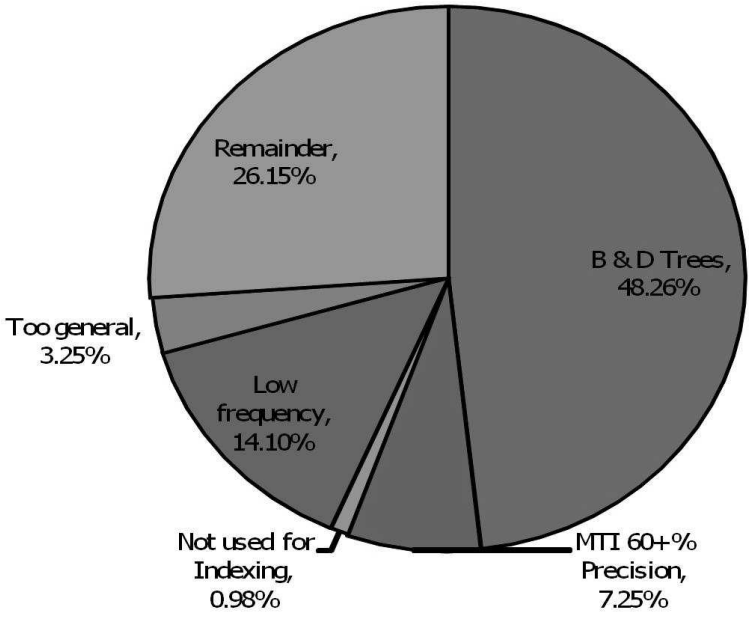


Figure 1: MTI accuracy primarily needs to be improved for 26% of MeSH suggestions

a semi-automatic procedure we followed to identify triage rules and a filtering step which are designed to emulate the approach used by the indexers. We start our exploration of this bottom-up approach using rules developed by a domain expert for recommending the Carbohydrate Sequence heading. Then, we further evaluate the proposed approach on two MeSH headings from the latest MTI test set. The results encourage the exploration of this method with other MeSH headings.

**Related work**

Publication of the OHSUMED collection[7] containing all MEDLINE citations into 270 medical journals over a five-year period (1987-1991) including MeSH indexing, provided for a large body of data that enabled us to view MH assignment as a classification problem. The scope of the collection determines the subset of MeSH that can be explored. For example, Lewis et al.[8] and Ruiz and Srinivasan[9] used 49 categories related to heart diseases with at least 75 training documents, and Yetisgen-Yildiz and Pratt[10] expanded the number of headings to 634 disease categories. Poulter[11] provides an overview of these and other studies of classification methods applied to MEDLINE and MeSH subsets.

The two-step approach to document triage and filtering was implied in the definition of the Text REtrieval Conference (TREC) Genomics track 2004 and 2005 categorization tasks, in which the main task was to consider each document for routing for further expert review or not, and in the subtasks the documents were annotated with specific terms[12,13].

A growing body of work approaches retrieval of MEDLINE citations as a classification task. For example, MScanner classifies all MEDLINE citations as relevant to a set of positive examples submitted by a user or not[11], and Kastrin et al.[14] determine the likelihood of MEDLINE citations topical relevance to genetics research. The large body of related work provides valuable insights with respect to classification of MEDLINE citations and feature selection methods. Applicability of these methods and suitability of the features for our specific task of improving indexing suggestions needs to be explored further. We find that most of the methods fit either into pattern matching methods which are based on a reference terminology (like UMLS or MeSH) and machine learning approaches which learn a model from examples of previously indexed citations.

Among the pattern matching methods we find the first component of MTI, as shown above, and an information retrieval approach by Ruch[15]; in his system the categories are the documents and the query is the text to be indexed. Pattern matching considers only the inner structure of the terms but not the terms with which they co-occur. This means that if a document is related to a MeSH heading but does not appear in the reference source, it will not be suggested. Machine learning based on previously indexed citations might help to overcome this problem.

This problem has been approached in several ways from a machine learning point of view. Machine learning methods tend to be ineffective with a large number of categories; MeSH contains more than 25k. Small scale studies with machine learning approaches already exist[16,10]. But the presence of a large number of categories has forced machine learning approaches to be combined with information retrieval methods designed to reduce the search space. For instance, PRC and a k-NN approach by Trieschnigg et al.[17] look for similar citations in MEDLINE and predict MeSH headings by a voting mechanism on the top-scoring citations. Experience with MTI shows that k-NN methods produce high recall but low precision indexing.

Other machine learning algorithms have been evaluated which rely on a more complex representation of the citations which do not only rely on unigrams or bigrams, e.g., learning based on ILP (Inductive Logic Programming)[18].

**Methods**

In the first part of this section, we present the work performed by a domain expert building triage rules to recommend the Carbohydrate Sequence heading. Then, we present the bottom-up approach proposed in this paper which is composed of two methods. In the first method, triage rules related to the MeSH term under study are identified in a semi-automatic fashion. In the second method, a false positive filter is applied based on statistical learning algorithms.

**Triage rules for recommending the Carbohydrate Sequence heading**

The abstracts of the scientific publications are reviewed to identify strings potentially containing carbohydrate sequences. The following carbohydrate names are used to identify candidate strings: *GlcpNAc, GalpNAc, GalNAc, GlcNAc, Neu5Ac, NeuAc, GalpA, GlcpA, Galp, Glcp, Rhap, NANA, Man, Fuc, Gal, Glc*. Based on empirical results, the first five names are converted to lower case, and for the rest of the list case information is preserved. When one of the carbohydrate names (starting with the longest) is found, the extent of the continuous string of text (with no blanks) enclosing the name is identified. The string containing the name is searched for the remaining carbohydrates if it is longer than 4 characters. The occurrences of carbohydrates in the string are marked as found and counted. If at least three carbohydrates names (not necessarily unique) occur within the string and at least one of the names is longer than 3 characters (or the string contains digits or parentheses in addition to 3-character long names), the string is considered a Carbohydrate Sequence and MTI recommends the heading. The 3 character carbohydrates are too commonly found in text to be allowed on their own without the support of digits or parentheses. Figure 2 illustrates the rules applied to an excerpt from an abstract (PMID 1368642). First, the longer carbohydrate *GlcNAc* was found. The extent of the continuous string (marked by arrows) was identified next, and then carbohydrates *Man*, and *Fuc* were identified in the string. Due to this combination of three carbohydrates, MTI recommended the Carbohydrate Sequence heading.

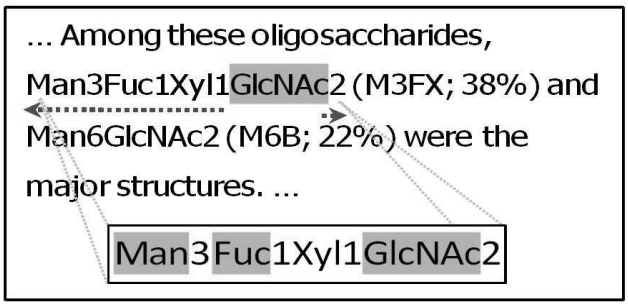


Figure 2: Excerpt detailing rules for identifying sequence

**Semi-automatic learning of triage rules**

Triage rules are learned in two steps: feature selection and rule learning, as shown in Figure 3. In the feature selection step, we select the most salient terms from a set of training citations. In the rule learning step, we build models which target the MeSH Heading giving preference to high recall, at the cost of precision in many cases. The rules produced in this step will provide an upper bound on recall.

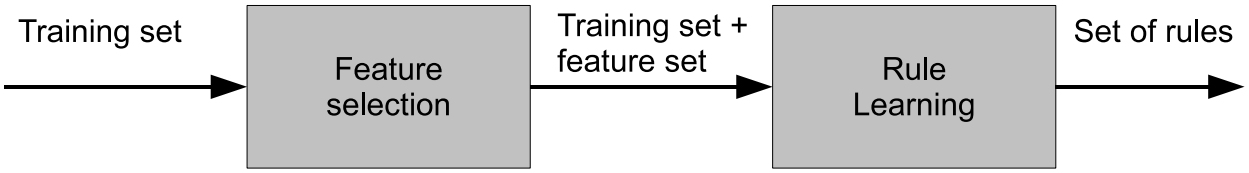


Figure 3: Triage rule learning

An example of triage rule derived from the Carbohydrate Sequence set is the following:

```
If any of the terms carbohydrate, polysaccharide, oligosaccharides combined
   with structure are found in citations in journals known to have Carbohydrate
Then assign Carbohydrate Sequence
```

**Feature selection step**

Feature selection is done before running the rule learning algorithm. Specifically, we used Mallet's implementation of Latent Dirichlet Allocation (LDA)[19] and Non-negative matrix factorization (NMF)[20]. From the derived model, we selected the top-n terms having a higher probability of belonging to a given topic. Examples of distilled terms for Carbohydrate Sequence are listed in Figure 5.

**Rule induction step**

Once the citations are represented by the most salient terms, we prepared a set of decision trees, which allow selecting the terms which correlate with the set of positive citations. The positive citations are the false negatives and the set of negative citations contain the false positive of the MTI predictions. Figure 4 shows a sample tree produced for *Carbohydrate Sequence* feature set. In this sample, the tree has been turned into a set of rules with output POS (recommend the MH) or NEG (do not recommend the MH). Once a rule is matched, the outcome is defined by the element after the colon punctuation mark. The rules are composed of the occurrence of a term in the citation and a combination of operators. The & sign is the Boolean operator AND and the ! sign is the negation operator. The second rule recommends Carbohydrate Sequence if the citation does not contain the terms *oligosaccharides polysaccharide* and *carbohydrate* but contains the term *gal*.

**Filtering based on machine learning**

The rule induction step improves recall but adds false positives in the process. In order to remove false positives and thereby improve precision, we applied machine learning. This task is considered a text categorization task; it attempts to filter out false positives by deciding if a given citation should be indexed with the given MeSH heading or not. We encounter issues, some of which are common to text categorization:

1. Imbalance between the number of positive and negative instances where the negative class usually overwhelms the positive one. Some machine learning algorithms have difficulty with this imbalance. We tested several approaches to deal with this issue: to balance the datasets, and to use a method based on the optimization of a

```
!oligosaccharides&!polysaccharide&!carbohydrate&!gal:NEG
!oligosaccharides&!polysaccharide&!carbohydrate&gal:POS
!oligosaccharides&polysaccharide&!structure&!intensity:POS
!oligosaccharides&polysaccharide&!structure&intensity:NEG
!oligosaccharides&polysaccharide&structure&!text:POS
!oligosaccharides&polysaccharide&structure&text:NEG
oligosaccharides&!structures&!galactopyranoside&!xylosyl:POS
oligosaccharides&!structures&!galactopyranoside&xylosyl:NEG
oligosaccharides&!structures&galactopyranoside&!enzyme:NEG
oligosaccharides&!structures&galactopyranoside&enzyme:POS
oligosaccharides&structures&!relies&!released:POS
oligosaccharides&structures&!relies&released:POS
oligosaccharides&structures&relies:NEG
```

Figure 4: MALLET decision tree learned for Carbohydrate Sequence

multivariate measure instead of relying on accuracy. Joachims[21] proposed an adaptation of SVM to optimize measures like $F$-measure or the area under the ROC-curve instead of accuracy, being an alternative to balancing the positive and negative instances.

2. Even if a term is correctly identified with a citation, it might not be significant enough to be included in the indexing.

3. Inconsistencies in the annotations might appear due to:

   (a) Inconsistency in MeSH indexing[22]. In machine learning terms, this is class label noise. Several existing techniques could be considered to overcome this problem. One of them[23] consists of selecting only documents for which a low level of discrepancy exists among classifiers. Then, the model is learned only from instances with a high level of confidence.

   (b) Changes in indexing policy over time can introduce inconsistencies with previously-indexed citations. This can even apply to routine changes to the structure of MeSH. In the selection of our set we carefully avoided this issue by selecting terms which were already in MeSH during the indexing period. In addition, the span of time considered is small enough to avoid most of the indexing policy changes which might have occurred.

**Filtering experimental setup**

In this section, configurations for our experiments are specified. These configurations can be combined to build different models. Unigram and bigram representation from the title and abstract of the MEDLINE citations are used as features. The classifiers considered for the experiments are:

1. Traditional classifiers (SVM, Naïve Bayes, decision trees and k-NN).

2. Ensemble of classifiers, which can reduce the variance of decision trees or can consider complementary views of the problem by different learning algorithms (Boosting, bagging, voting, ...).

3. SVM with multivariate measures[21].

**Evaluation**

The methodology presented above has been evaluated on two datasets. The first one is a screening of MEDLINE with the MeSH heading Carbohydrate Sequence. The second one is a subset of MEDLINE used routinely for testing changes to MTI's algorithm.

### Carbohydrate Sequence set

The Carbohydrate Sequence dataset is a subset of the 2010 MEDLINE baseline which contains 18,502,916 MEDLINE citations. The dataset consists of 2,307 citations found by Carbohydrate Sequence triage rules developed by a domain expert. The 2,307 citations that have the Carbohydrate Sequence heading (1,212 positive examples) and those that do not (1,095 negative examples) were further split into the training set containing 80% of the positive and negative examples and the test set containing the remaining 20% of the citations.

We consider citations with the Carbohydrate Sequence MH to be in the positive class because we are primarily interested in finding all citations that should be recommended for assigning the heading. The 80-20 split was performed using publication dates of the citations with the most recent citations in the test set. This split imitates the real-life situation in which a method is developed with an existing set of data and then applied to and tested on a future set.

A second set is developed for exploration of additional triage rules for the Carbohydrate Sequence heading. In this set we used 16,781 citations having the heading but not found by the current triage rules. We use a matching stratified sample of citations without the MH that were published in the journals containing at least one citation with the Carbohydrate Sequence MH as negative examples. These sets were also split into training and test subsets following the 80-20 rule described above.

### MTI test set

This set is a subset of the 2009 MEDLINE baseline as used by the MTI team for verifying changes to MTI. We selected candidate terms highly represented in MEDLINE but with poor recall performance by MTI. The list of selected terms is found in Table 1.

| MeSH Heading | Unique ID | Tree Number |
|---|---|---|
| Acute Disease | D000208 | C23.550.291.125 |
| Gene Expression | D015870 | G05.355.310 |

Table 1: Selected MeSH headings

This set is split into training and test sets based on the publication date field (DP field in PubMed). The citations from the first 8 months of 2009 are used for training and the final 4 months for test. In the training set there are 409,279 citations with a total of 343,504 citations with abstract and in the test set, there are 255,493 citations with a total of 214,064 citations with abstracts.

## Results

### Carbohydrate Sequence set

In the preliminary exploration of the triage rules presented in the methods section, we noticed that many candidate citations do not get the MH assigned, while the majority of the citations having the Carbohydrate Sequence heading are not categorized as candidates. These observations led to expansion of the approach in two directions as introduced in the methods section: (1) using supervised machine learning to improve precision of recommendations for candidate citations found by the original triage rules, and (2) generation of new rules to expand the candidate set with a subsequent application of supervised machine learning. The goal of the triage step is to reduce the number of irrelevant citations to be processed in the second step. The goal of the machine learning step is to improve precision without losing recall.

The original Carbohydrate Sequence triage rules reduce the size of the set to be considered for Carbohydrate Sequence to 0.012% of the full document set with 6.7% recall and 52.5% precision. In the subsequent machine learning step the Maximum Entropy classifier trained on 1, 2, 3, and 4 token sequences and cutoff confidence level set above 0.2 reduced the number of wrong recommendations (precision of 53.6%) with almost all correct recommendations (90% of them).

The topics built using the positive examples (citations having the MH but no sequence strings) contained many key phrases pertaining to analysis methods (for example: *NMR spectroscopy*, *mass spectrometry*, *methylation analysis*), model organism and chemical names, which we found to be too general to pertain only to studies of carbohydrate sequences. However, topic analysis also provided pertinent terms, which we combined with the rules generated by the MALLET Decision Tree classifier shown in Figure 4. The new triage rules select citations for further consideration combining the common segments of the positive rules in Figure 4 and the topic terms found by LDA as follows:

1. Rule 1: If any of the terms *carbohydrate*, *polysaccharide*, *oligosaccharides* combined with structure are found in citations in journals known to have Carbohydrate
2. Rule 2: If two or more of the 26 terms listed for Carbohydrate Sequence (in Figure 5) are found in citations in journals known to have Carbohydrate
3. Rule 3: If rules 1 and 2 apply to the citation

carbohydrate(s), disaccharide(s), Fuc, Gal, GalNAc, galacturonic acid, GlcNAc, glucopyranosyl, glucuronic acid, glycan(s), glycosidic linkages, glycosylation, hyaluronic acid, iduronic acid, lipopolysaccharide(s), LPS, Man, NeuAc, oligosaccharide(s), polysaccharide(s), sialyl Lewis, sialic acid, Smith degradation, sugar chains, triterpenoid saponins

Figure 5: Terms selected from topics built based on the positive training examples for Carbohydrate Sequence recommendation

The new rules were evaluated using the second set of citations that have the Carbohydrate Sequence MH and further reduce the set of candidate citations. Results are available in Table 2.

| Rule | True Positives | False Positives | Precision | Recall | $F$-measure | $F_2$-measure |
|---|---|---|---|---|---|---|
| Rule 1 | 3108 | 6528 | 0.4761 | 0.1733 | 0.2541 | 0.1986 |
| Rule 2 | 8144 | 40768 | 0.1998 | 0.4541 | 0.2775 | 0.3620 |
| Rule 1 & 2 | 2391 | 3043 | 0.7857 | 0.1333 | 0.2280 | 0.1599 |
| CH & structure | 1234 | 3883 | 0.3178 | 0.0688 | 0.1131 | 0.0816 |
| Poly & structure | 1292 | 1712 | 0.7547 | 0.0720 | 0.1315 | 0.0880 |
| Olig & structure | 1233 | 1566 | 0.7874 | 0.0688 | 0.1265 | 0.0841 |

Table 2: Rules and precision/recall values for the training and test set

Compared to the results using the initial triage rules, we obtain different precision recall levels which in most of the cases are better than the original triage rules developed manually by a domain expert.

**MTI test set**

**Triage rule learning**

Table 3 shows the results of the recall analysis for the MTI set. As described in the methods section, decision trees were built from the selected features, and common sections were manually analyzed. After careful analysis of the feature sets and common trees, rules where manually selected for each MeSH heading in order to obtain high coverage of the MeSH headings with reasonable precision. In Table 3 the rules and the performance measures are shown. The recall values are significantly increased compared to baseline MTI performance, but this was negatively compensated by a noticeable decrease in precision values.

**Filter analysis**

In Table 4, we find the results from the precision analysis for each of the MeSH headings. In this table, we show only the result produced by the best performing learning algorithm. For each MeSH heading we show the MTI result, the

| | Rule | True Positives | False Positives | Precision | Recall |
|---|---|---|---|---|---|
| Acute Disease | 'acut' | 1387 | 10409 | 0.1176 | 0.8562 |
| Gene Expression | ('protein' & 'express') \| ('gene' & 'express') \| ('cell' & 'express') | 1722 | 24978 | 0.0645 | 0.8165 |

Table 3: Rules and precision/recall values for the training and test set

MTI result with filtering, the recall analysis result, and the recall result with filtering. Overall, we find that filtering improves the precision but at a high recall cost. $F$-measure and $F_2$-measure results vary according to the method.

| Acute Disease | True Positives | False Positives | Precision | Recall | $F$-measure | $F_2$-measure |
|---|---|---|---|---|---|---|
| MTI | 256 | 705 | **0.2664** | 0.1580 | 0.1984 | 0.1720 |
| Filtering | 226 | 303 | **0.4272** | 0.1395 | 0.2103 | 0.1612 |
| Triage rules | 1387 | 10409 | 0.1176 | **0.8562** | 0.2068 | 0.3795 |
| Triage + filtering | 1071 | 4447 | 0.1941 | 0.6611 | **0.3001** | **0.4463** |
| Gene Expression | True Positives | False Positives | Precision | Recall | $F$-measure | $F_2$-measure |
| MTI | 572 | 2349 | **0.1958** | 0.2712 | **0.2274** | 0.2518 |
| Filtering | 293 | 816 | **0.2642** | 0.1389 | 0.1896 | 0.1805 |
| Triage rules | 1722 | 24978 | 0.0645 | **0.8165** | 0.1195 | 0.2450 |
| Recall + filtering | 1101 | 8639 | 0.1130 | 0.5220 | 0.1858 | **0.3029** |

Table 4: Results comparing the different analyses

**Discussion**

Focusing on the Carbohydrate Sequence set, the original triage rules based on identification of carbohydrate names are too specific to capture all candidates for recommending the Carbohydrate Sequence MH. To our surprise, the rules also capture a significant number of citations that should not be recommended for this MH. Our domain expert reviewed five citations that our best scoring classifier erroneously assigned to the positive class with highest confidence and concluded that those citations qualify for the MH. This indicates that the actual accuracy of the original rule might be higher than in our evaluation. In both cases, the filtering and the semi-automatically derived triage rules achieve results appropriate for the MEDLINE MeSH indexing.

For the MTI set, we show that semi-automatic generation of triage rules is potentially helpful to MEDLINE MeSH indexing. From the MTI analysis, we find that only in the case of *Acute Disease* did the recall and filtering analysis provide a result much higher compared to MTI. We can see that precision can be largely improved using machine learning based filters. On the other hand, recall decreases significantly in most of the cases. The $F$-measure is sometimes improved due to the increase in precision but the $F_2$-measure is lower in almost all scenarios due to the loss in recall. We plan to evaluate this approach with a larger set of MeSH headings occurring in the MTI set used in the experiments, and not only on MeSH headings with MTI's poor performance.

In addition, considering triage rules in the MTI set, precision is much lower compared to the original MTI prediction even if recall is much higher. After filtering, we find that precision is largely improved compared to the results from the recall analysis. The precision values are lower compared to the results obtained with MTI, so in many cases the final $F$-measure is lower. But for the $F_2$-measure *Acute Disease* and *Gene Expression* achieve better performance. Only in the case of *Acute Disease* we find that the $F$-measure and $F_2$-measure are improved in all the scenarios.

If we consider the machine learning algorithms in the filtering step, we find that low variance methods have better performance in comparison to low bias and high variance methods. Examples of low variance methods are SVM and AdaBoost. In many learning algorithms, balancing the dataset produced an improvement in the performance of the machine learning algorithms. Multivariate optimization achieved the highest performance with *Acute Disease* while AdaBoost achieved the highest improvement with *Gene Expression*.

## Conclusions and Future work

Our work confirms that the method presented in this paper produces improved recommendations of some MeSH headings compared to existing methods and to manual assessment in the case of Carbohydrate Sequence. We plan to explore the scalability of the proposed approach, applying it to other headings for which indexing recommendations need to be improved.

The results presented in this work have been produced using unigrams and bigrams from the title and abstract. Other features extracted from text, like the position of the MeSH heading in the document or normalization of the features based on MetaMap, could be combined to improve the current results. This might then require moving to full-text analysis.

It should be mentioned that our results are probably negatively affected by papers which a title but no abstract (around 17% of all papers in the dataset). The fact that NLM indexers use the full-text of an article to perform MEDLINE indexing further argues for an extension of the current study to the use of full-text, not just titles and abstracts. Further studies on full-text might be required, but only 15% of the PMIDs in our dataset could be matched to PMC identifiers. Specific feature selection and combination might be required to process the articles efficiently.

## Acknowledgments

This work was supported in part by the Intramural Research Program of the NIH, National Library of Medicine. This research was supported in part by an appointment to the NLM Research Participation Program. This program is administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the National Library of Medicine. The second author also gratefully acknowledges funding from the Lundbeckfonden through the Center for Integrated Molecular Brain Imaging (Cimbi.org), Otto Mønsteds Fond, Kaj og Hermilla Ostenfelds Fond, and the Ingeniør Alexandre Haynman og hustru Nina Haynmans Fond.

## References

1. A.R. Aronson, O. Bodenreider, H.F. Chang, S.M. Humphrey, JG Mork, SJ Nelson, TC Rindflesch, and WJ Wilbur. The NLM Indexing Initiative. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association, 2000.
2. A.R. Aronson, J.G. Mork, C.W. Gay, S.M. Humphrey, and W.J. Rogers. The NLM Indexing Initiative's Medical Text Indexer. In *Medinfo 2004: proceedings of the 11th World Conference on Medical Informatics,[San Francisco, september 7-11, 2004]*, page 268. OCSL Press, 2004.
3. O. Bodenreider, S.J. Nelson, W.T. Hole, and H.F. Chang. Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. In *Proceedings of the AMIA symposium*, page 815. American Medical Informatics Association, 1998.
4. J. Lin and W.J. Wilbur. PubMed related articles: a probabilistic topic-based model for content similarity. *BMC bioinformatics*, 8(1):423, 2007.
5. C.W. Gay, M. Kayaalp, and A.R. Aronson. Semi-automatic indexing of full text biomedical articles. In *AMIA Annual Symposium Proceedings*, volume 2005, page 271. American Medical Informatics Association, 2005.
6. A.R. Aronson, J.G. Mork, A. Névéol, S.E. Shooshan, and D. Demner-Fushman. Methodology for creating UMLS content views appropriate for biomedical natural language processing. In *AMIA Annual Symposium Proceedings*, volume 2008, page 21. American Medical Informatics Association, 2008.
7. W. Hersh, C. Buckley, TJ Leone, and D. Hickam. OHSUMED: an interactive retrieval evaluation and new large test collection for research. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 192–201. Springer-Verlag New York, Inc., 1994.
8. D.D. Lewis, R.E. Schapire, J.P. Callan, and R. Papka. Training algorithms for linear text classifiers. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 298–306. ACM, 1996.
9. M.E. Ruiz and P. Srinivasan. Hierarchical neural networks for text categorization (poster abstract). In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*,

pages 281–282. ACM, 1999.
10. M. Yetisgen-Yildiz and W. Pratt. The effect of feature representation on MEDLINE document classification. In *AMIA Annual Symposium Proceedings*, pages 849–853. American Medical Informatics Association, 2005.
11. G.L. Poulter, D.L. Rubin, R.B. Altman, and C. Seoighe. MScanner: a classifier for retrieving Medline citations. *BMC bioinformatics*, 9(1):108, 2008.
12. W. Hersh, R.T. Bhuptiraju, L. Ross, P. Johnson, A.M. Cohen, and D.F. Kraemer. TREC 2004 genomics track overview. In *Proceedings of the thirteenth Text REtrieval Conference*, 2004.
13. W. Hersh, A. Cohen, J. Yang, R.T. Bhupatiraju, P. Roberts, and M. Hearst. TREC 2005 genomics track overview. In *Proceedings of the fourteenth Text REtrieval conference*, 2005.
14. A. Kastrin, B. Peterlin, and D. Hristovski. Chi-square-based scoring function for categorization of MEDLINE citations. *Methods of Information in Medicine*, 48:10–3414, 2009.
15. P. Ruch. Automatic assignment of biomedical categories: toward a generic approach. *Bioinformatics*, 22(6):658, 2006.
16. Y. Aphinyanaphongs, I. Tsamardinos, A. Statnikov, D. Hardin, and C.F. Aliferis. Text categorization models for high-quality article retrieval in internal medicine. *Journal of the American Medical Informatics Association*, 12(2):207–216, 2005.
17. D. Trieschnigg, P. Pezik, V. Lee, F. De Jong, W. Kraaij, and D. Rebholz-Schuhmann. MeSH Up: effective MeSH text classification for improved document retrieval. *Bioinformatics*, 25(11):1412, 2009.
18. A. Névéol, S. Shooshan, and V. Claveau. Automatic inference of indexing rules for MEDLINE. *BMC bioinformatics*, 9(Suppl 11):S11, 2008.
19. D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.
20. D.D. Lee and H.S. Seung. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13, 2001.
21. T. Joachims. A support vector method for multivariate performance measures. In *Proceedings of the 22nd international conference on Machine learning*, pages 377–384. ACM, 2005.
22. M.E. Funk and C.A. Reid. Indexing consistency in MEDLINE. *Bulletin of the Medical Library Association*, 71(2):176, 1983.
23. X. Zhu, X. Wu, and Q. Chen. Eliminating class noise in large datasets. In *Proceedings of the 20th International Conference on Machine Learning*, 2010.