# "Hybrid topics" -- Facilitating the interpretation of topics through the addition of MeSH descriptors to bags of words

**Zhiguo Yu[a], Thang Nguyen[b], Ferdinand Dhombres[c], Todd Johnson[a], Olivier Bodenreider[c]**

*[a] The University of Texas of Biomedical Informatics at Houston, Houston, Texas, USA,*
*[b] Department of Computer Science, University of Maryland, College Park, Maryland, USA*
*[c] U.S. National Library of Medicine, National Institute of Health, Bethesda, Maryland, USA*

**Abstract**

*Extracting and understanding information, themes and relationships from large collections of documents is an important task for biomedical researchers. Latent Dirichlet Allocation (LDA) is an unsupervised topic modeling technique using the "bag-of-words" assumption that has been applied extensively to unveil hidden thematic information within large sets of documents. In this paper, we added MeSH descriptors to the 'bag-of-words' assumption to generate 'hybrid topics', which are mixed vectors of words and descriptors. We evaluated this approach on the quality and interpretability of topics in both a general corpus and a specialized corpus. Our results demonstrated that the coherence of 'hybrid topics' is higher than that of regular bag-of-words topics in the specialized corpus. We also found that the proportion of topics that are not associated with MeSH descriptors is higher in the specialized corpus than in the general corpus.*

***Keywords:***

***Medical Subject Headings, Statistical Models, Statistical Data Interpretation***

## Introduction

***Motivation***. Knowledge discovery is a fundamental and important activity in biomedical research. Given that unstructured data is increasing exponentially, extracting and understanding information, themes, and relationships from large collections of documents is increasingly important for biomedical researchers. PubMed currently comprises more than 26 million articles from the biomedical literature and has been widely used to help researchers keep up with the state of the art in their research domains and explore other unfamiliar research areas. To help users better retrieve relevant information and manage this tremendous volume of biomedical literature, the US National Library of Medicine (NLM) has developed the Medical Subject Headings (MeSH) controlled vocabulary for indexing MEDLINE articles. MeSH has been used to improve PubMed query results [1; 2]. However, users are still often overloaded by a tremendous number of relevant articles from their PubMed queries [3]. Hence, biomedical researchers need an efficient and convenient way to discover knowledge from large sets of documents.

***MeSH indexing vs. topics***. Latent Dirichlet Allocation (LDA) [4], is a popular topic modeling method that aims to extract the semantic themes (topics) automatically from a corpus of documents. These topics describe the thematic composition of documents and can thus capture the semantic similarity between them. In contrast, MeSH descriptors are manually created and maintained by domain experts to cover all important themes. Topics extracted from a subset of documents are specific to that subset [5]. Thus they may identify corpus-specific themes that may not be covered in MeSH. Such themes may uncover a specific set of concepts for a particular domain or sub-domain.

***Related work***. Considerable research has examined the application of topic models to MeSH descriptors. Labeled Latent Dirichlet Allocation (labeled LDA) [6] is a supervised topic model developed to uncover latent topics that correlate with user tags in labeled corpora. In other words, each tag will be represented as a topic. Zhu et al. have used labeled LDA in an attempt to automatically assign MeSH descriptors to new publications (not yet indexed with MeSH descriptors) [7]. Elsewhere, Newman et al. presented a resampled author model that combines both general LDA and the author-topic model (in this case MeSH descriptors were treated as the "authors"). The resampled author model provided an alternative and complementary view of the relationships between MeSH descriptors [8]. All these investigations used topic models to interpret MeSH descriptors. However, they cannot be used to identify themes that are not covered in MeSH. In 2014, a graph-sparse LDA model [9] was developed to generate more interpretable topics by leveraging relationships expressed by controlled vocabulary structure (e.g. MeSH). In this model, a few concept-words from the controlled vocabulary can be identified to represent generated topics. Though MeSH was shown to work well in this model to help summarize biomedical articles, there is still no way for this model to identify themes that may not exist in the MeSH vocabulary.

***Specific contribution***. The specific contribution of this work is to introduce an alternative LDA approach by migrating its original 'bag-of-words' assumption to a 'bag-of-MeSH&words' approach. By enriching each document with its indexed MeSH descriptors, 'hybrid topics' (mixed vectors of words and MeSH descriptors) can be generated by LDA.

***Objectives***. In this paper, we investigate whether the addition of labels (e.g. MeSH descriptors) to bags of words can improve the quality and facilitate the interpretation of LDA-generated topics. More specifically, to assess the quality and interpretability of topics, we test two hypotheses using one large general biomedical corpus and one smaller specialized biomedical corpus.

1) The coherence (used as a surrogate for quality) of 'hybrid topics' is expected to be higher than that of regular bag-of-words topics
2) The proportion of topics that are not associated with some MeSH descriptor, which reflects limited interpretability, is expected to be higher in a specialized corpus than in a general corpus.

## Background

### Medical Subject Headings (MeSH)

The Medical Subject Headings (MeSH) controlled vocabulary has been developed by NLM to help manage, index, and search MEDLINE articles. There are 27,883 descriptors in the 2016 MeSH, with over 87,000 entry terms that assist in finding the most appropriate MeSH descriptor (for example, 'Vitamin C' is an entry term to 'Ascorbic Acid'). In the 2016 MeSH, 82 qualifiers can be attached to MeSH descriptors to describe a particular aspect of a concept, such as 'adverse effects', 'diagnosis', etc. Each year, the MeSH specialists revise and update the MeSH vocabulary to cover emerging research areas and improve indexing consistency and efficiency. MeSH specialists are responsible for areas of the health sciences in which they have knowledge and expertise. MEDLINE indexers make suggestions for new descriptors to MeSH specialists during their indexing processes. Besides, MeSH specialists also collect new terms as they appear in the scientific literature or emerging areas of research. After defining these terms within the context of the existing vocabulary, MeSH specialists recommend their addition to MeSH. During each MEDLINE year-end processing (YEP) activities, changes made to MeSH are applied to MEDLINE (retrospectively) for conformance with the current version of MeSH.

### Latent Dirichlet Allocation (LDA)

LDA is a generative model that assumes that each document is generated from a mixture of topics and that each topic corresponds to a distribution over all words in the corpus. Informally, the 'generative story' for LDA is as follows. First, a document is generated by drawing a mixture of topics that the document is about. To generate each word in this document, one draws a topic from this distribution and subsequently selects a word from the distribution over the vocabulary of the whole corpus corresponding to this topic. The LDA algorithm uses this generative model to uncover the latent topics contained within a given a corpus. Specifically, it estimates the parameters that define document topic mixtures and the conditional probabilities of each word given each topic. Parameter estimation is usually done via sampling approaches.

The number of topics produced by LDA must be prespecified. Determining the 'right' number of topics for different datasets remains a challenge. When the number of topics increases, redundant and nonsense topics may be generated. Running LDA with a small number of topics will generate more general themes. In this paper, we used a topic coherence measure to determine the optimal number of topics for our dataset [10]. Details are described in the next section.

### Quality of topics

Recently, O'Callaghan et al. [10] reviewed a number of topic coherence studies using various corpora and proposed a general measure based on distributional semantics, *TC-W2V*. It evaluates the relatedness of a set of top terms describing a topic, based on the similarity of their representations in a *word2vec* distributional semantic space. Specifically, the coherence of a topic *n* represented by its top *t* ranked terms is given by the mean pairwise cosine similarity between all relevant term vectors in the *word2vec* space:

$$coh(t_n) = \frac{1}{\binom{t}{2}} \sum_{j=2}^{t} \sum_{i=1}^{j-1} \cos(wv_i, wv_j)$$

An overall score for a topic model *M* consisting of *k* topics is calculated by averaging the individual topic coherence scores:

$$coh(M) = \frac{1}{k} \sum_{n=1}^{k} coh(t_n)$$

In this investigation, we use topic coherence not only to help determine the optimal number of topics, but also, more generally, the quality of topics.

### Relations between MeSH and topics

Topics are generated based on the 'bag-of-words' assumption, which ignores word order. Each topic is represented as a list of ranked words, which is used to provide the user a sense of what this topic is about. Each document is displayed as a list of weighted topics, which represents different aspects of this document. Since the tokens within each topic are ranked according to the conditional probabilities *P(w|t)* learned when training the model, where w is a word and t is a topic, the top few words of each topic provide insights into the subject of the topic. However, the interpretation of the topics (i.e., lists of words) is left as an exercise for the user.

As mentioned earlier, MeSH is developed to cover all important themes and each article in MEDLINE is indexed with a few relevant MeSH descriptors assigned by the MEDLINE indexing staff for retrieval purposes. To capture the potential relationships between MeSH and topics, we simply added the MeSH descriptors assigned to each article to the 'bag-of-words' for this document, creating a 'bag-of-MeSH&word' instead. Under this 'bag-of-MeSH&word" assumption, 'hybrid topics' are generated and each topic is represented as a list of tokens, i.e., a mixed list of ranked words and MeSH descriptors. The presence of MeSH descriptors among the top tokens for a given topic is expected to facilitate the interpretation of topics. More specifically, if a MeSH descriptor appears among the top *m* (for some *m*) tokens of a topic, we will assume this MeSH descriptor is highly associated with this topic.

We consider three types of association patterns between topics and MeSH descriptors.

1. One topic has no MeSH descriptor in its top *m* tokens (*1-0 mapping*);
2. One topic has a single MeSH descriptor in its top *m* tokens (*1-1 mapping*);
3. One topic has multiple MeSH descriptors in the top *m* tokens (*1-many mapping*).

Examples of topics for each association pattern are shown in Table 1, along with the top 10 tokens for each topic.

*Table 1– Topics generated based on 'bag-of-MeSH&word'. (Asterisks indicate MeSH descriptors)*

| **Topic1** | **Topic2** | **Topic3** |
|---|---|---|
| model | *brain | motor |
| predict | cortex | visual |
| value | region | *movement |
| prediction | functional | *face |
| analysis | cortical | right |
| predictive | activity | response |
| regression | neural | *hand |
| datum | network | processing |
| estimate | change | object |
| predictor | area | stimuli |
| ***1-0 mapping*** | ***1-1 mapping*** | ***1-many mapping*** |

## Methods

### Data preparation

One large general corpus and one small specialized corpus are used in this investigation. The general corpus consists of 200k articles randomly selected from all PubMed articles published in 2013. The specialized corpus consists of 2472 articles from the journal *Prenatal Diagnosis*, which focuses on fetal medicine.

***General corpus***. There are about 1.2 million articles in PubMed for the year 2013. We randomly selected 200k articles (titles and abstracts) from these. This represents an appropriate amount of data given our computing resources.

To reduce the sparsity of document-to-words distribution, we performed Part of Speech tagging on the dataset and merged several categories, including *NN* and *NNS* (e.g., *patient* and *patients*); *VB*, *VBD*, *VBG*, and *VBN* (e.g., *eat*, *ate*, *eaten*, and *eating*); and *JJ*, *JJR*, and *JJS* (e.g., *good*, *better*, and *best*).

We also removed PubMed stop-words and infrequent words (with a frequency lower than 50). A total of 21,922 unique words remained. Similarly, for MeSH descriptors, we treated specific frequently used descriptors known as check tags (e.g., human, male, female, etc.) as stop words, and ignored infrequent descriptors (with a frequency lower than 5). A total of 13,853 MeSH descriptors remained.

***Specialized corpus***. We applied a similar preprocessing to the specialized corpus, but with different cutoff values due to its smaller size. After setting a cutoff frequency of 5 for words, we obtained 3623 unique words. With a cutoff frequency of 1 for MeSH descriptors, we obtained 919 MeSH descriptors.

### Experiment #1

We investigated whether the addition of MeSH descriptors to bags of words increases the quality of topics. As a surrogate for the quality of topics, we use topic coherence [11].

In practice, to determine whether our 'hybrid topic' approach (i.e., 'bag-of-MeSH&words') outperforms the original LDA 'bag-of-words' approach (baseline), we generated LDA models under both these assumptions for a various number of topics on the two datasets.

For the general corpus, the number of topics tested ranges from 50 to 600. For the specific corpus, we tested from 4 to 100. For each number of topics, we calculated topic coherence for both the baseline and the hybrid topic approach.

More specifically, for the large general corpus, we used the indexed PubMed articles (titles and abstracts) published in 2013 as our background corpus when building the *word2vector* space for the original LDA with 'bag-of-words' assumption. To build the *word2vector* space containing both MeSH descriptors and words, we simply appended the MeSH descriptors for an article to the end of the document. In this way, we could get a mixed *word2vector* space of MeSH descriptors and words. In our experiment, we tested two different positions of MeSH descriptors in the citation (front and end) and obtained similar topic coherence results. Following [10], we used the same *word2vec* setting and the number of top terms per topic ($t$=10).

For the small specialized corpus, we used the full-text of these articles as the background corpus when building the *word2vec* space. To build the mixed *word2vector* space of MeSH descriptors and words for this background corpus, we added MeSH descriptors to the end of each full article. In the *word2vec* setting for this dataset, we set vector size to 200, cutoff frequency to 3, and window size to 20.

To compare the topic coherence measures obtained within each corpus at different numbers of topics for the baseline and the hybrid topics, we used a paired t-test.

### Experiment #2

To assess whether the proportion of 'hybrid topics' that are not associated with some MeSH descriptor, which reflects limited interpretability, is higher in a specialized corpus than in a general corpus, we first have to determine the optimal number of topics in each corpus for this assessment.

Choosing the number of topics k is a key parameter selection decision in topic modeling. Too few topics will produce results that overly broad, while too many will lead to many small, highly similar topics. One general strategy proposed in the literature has been to compare the topic coherence of topic models with different values of k. An appropriate value for *k* can then be identified by examining a plot of the mean *TC-W2V* coherence scores for a fixed range and selecting a value corresponding to the maximum coherence score. Since we only expected the MeSH descriptors to help interpret topics rather than for introducing new topics, we just used LDA's original 'bag-of-words' assumption to determine the optimal number of topics for each test corpus.

Having determined the optimal number of topics for each corpus, we examine the 'hybrid topics' obtained for this number of topics and count which ones are not associated with MeSH descriptors, i.e., which ones do not contain at least one MeSH descriptor among their top-20 tokens.

We use the chi-square statistics to compare the distribution of topics of 2 patterns between the 'hybrid topics' and the baseline.

## Results

### Experiment #1

Figures 1 and 2 display the difference in topic coherence between our 'bag-of-MeSH&words' assumption (hybrid topics) and LDA's original 'bag-of-words' assumption (baseline) for the general and specialized corpora respectively.

For the general corpus, we computed topic coherence for 10 different numbers of topics for both the baseline and our hybrid topics. As shown in Figure 1, topic coherence scores are very close between the baseline and hybrid topics. The coherence is slightly better with hybrid topics after 100 topics, but slightly lower for 50 and 100 topics.

For the specialized corpus, however, we can see a clear improvement on the coherence of topics in favor of hybrid topics compared to the baseline. As shown in Figure 2, topic coherence scores are systematically higher for hybrid topics across all numbers of topics.

Though hybrid topics are over the baseline after 100 topics on the general corpus, the paired t-test is not significant (p=0.1624). We cannot properly assess the difference between the two approaches on this general corpus. With the specialized corpus, however, the paired t-test is highly significant (p=6.8e-25), demonstrating that the quality of the hybrid topics is better than that of the baseline topics.

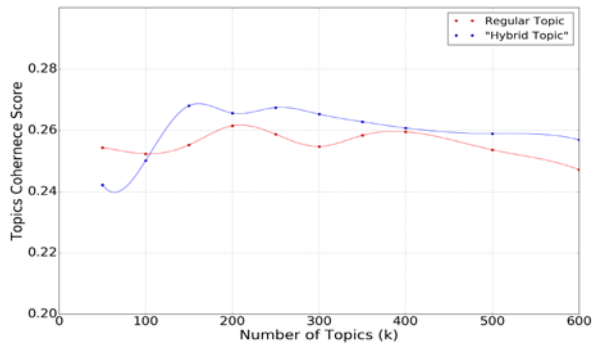


*Figure 1- Comparison of mean TC-W2V topic coherence scores for different numbers of topics k, generated from the general corpus*

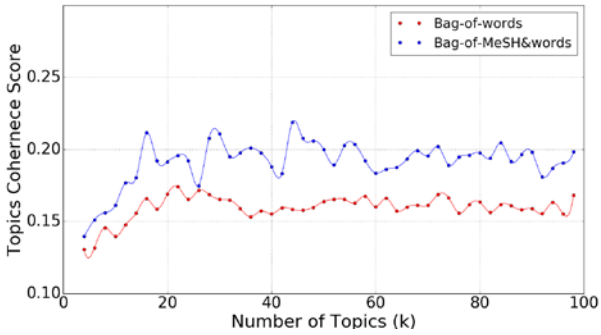


*Figure 2- Comparison of mean TC-W2V topic coherence scores for different numbers of topics k, generated from the specialized corpus*

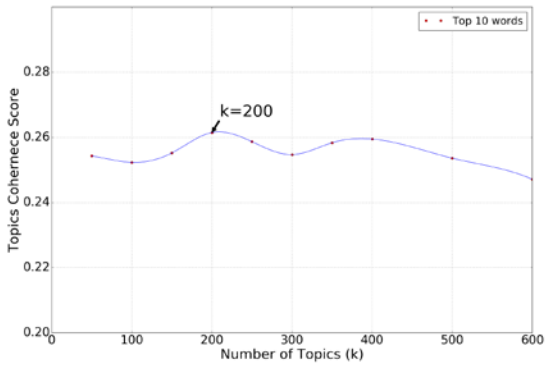


*Figure 3- Plot of mean TC-W2V topic coherence scores for different numbers of topics k, generated from the general corpus.*

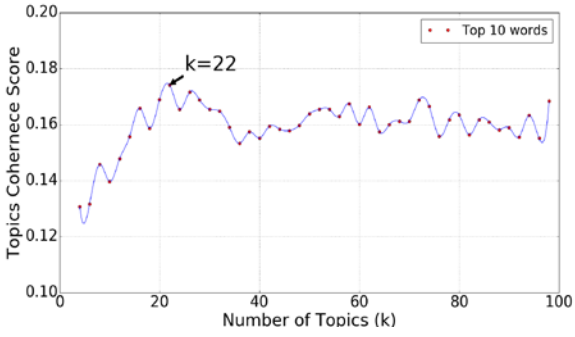


*Figure 4- Plot of mean TC-W2V topic coherence scores for different numbers of topics k, generated from the specialized corpus.*

**Experiment #2**

***Optimal number of topics***. For the large general corpus, we generated LDA models containing $k \in [50,600]$ topics and selected the value of *k* that maximized mean *TC-W2V* coherence. As shown in Figure 3, *k=200* is the first maximum and should therefore be selected as the optimal number of topics for this dataset.

For the small specialized corpus, we generated LDA models containing $k \in [4,100]$ topics. As shown in Figure 4, *k=22* is the first maximum and should therefore be selected as the optimal number of topics for this dataset.

***Proportion of topics not associated with a MeSH descriptor***. Table 2 displays the number of different patterns of association between topics and MeSH descriptors observed in the general and specialized corpus for their respective optimal number of topics.

As shown in Table 2, the proportion of topics not associated with a MeSH descriptor is higher in the specialized corpus (41%) than in the general corpus (3%). The chi-square statistics is 57.36 with a p-value of 3.502e-13, which shows that the distributions are significantly different in this two corpora.

*Table 2– # of Topics with 0,1,n MeSH descriptors (n>1)*

| Data Set | Optimal *K* | # of Topic with 0 MD | # of Topic with 1 MD | # of Topic with n MD |
|---|---|---|---|---|
| General Corpus | 200 | 6 (3%) | 16 (8%) | 178 (89%) |
| Spec. Corpus | 22 | 9 (41%) | 6 (27%) | 7 (32%) |

## Discussion

### Findings and significance

This investigation demonstrates that the addition of MeSH descriptors to the traditional bag-of-words approach to creating topic models ('hybrid topics') can improve the quality of the topics and facilitate their interpretation, but the impact is different on a general corpus and on a specialized corpus. The quality of the hybrid topics, assessed by their coherence, is better than that of the baseline topics in the specialized corpus, but it does not seem to be the case in the general corpus.

MeSH terms are created and maintained by MeSH specialists to cover all general themes in biomedicine. However, topics extracted from a subset of documents are often specific to these documents. For the general corpus, most of the topics captured by LDA are indeed general themes. Hence, this addition of MeSH descriptors to the bag-of-words approach did not contribute too much to the topic quality. This could be the reason that we did not see a significant improvement of the topic coherence score between regular topics and hybrid topics in the general corpus. In contrast, for the specialized corpus, adding MeSH descriptors can provide additional information for LDA to better differentiate between general and specific themes and to improve topics quality.

In terms of interpretability, however, the general corpus benefits from hybrid topics more than the specialized corpus does, because over 40% (9/22) of the hybrid topics remain unlabeled (i.e., not associated with any MeSH descriptors) in the specialized corpus, compared to 3% (6/200) in the general corpus.

**Applications to corpus exploration**

From the general corpus, we see that only 6 of the 200 topics (3%) contain 0 MeSH descriptors in their top 20 terms. For the specialized corpus, 9 of the 22 topics (40%) are generated with 0 MeSH descriptors in their top 20 terms. General themes from the MeSH vocabulary may not be able to cover in detail all aspects of a specialized corpus. In contrast, the topics generated by LDA from a corpus are specific to this corpus. It is therefore logical that more topics with no MeSH descriptors are generated from a specialized corpus than a general corpus. Hence LDA will be more useful for a specialized corpus on the task of exploring concepts that may not be covered by MeSH.

From the general corpus, we also see that 178 of the 194 topics associated with MeSH descriptors (92%) are generated with multiple MeSH descriptors. MeSH descriptors are characterized in 16 top-level categories, such as category A for anatomic terms, category B for organisms, C for diseases, etc. Of these 178 topics, 140 (79%) contain MeSH descriptors from different top-level MeSH categories. These topics are most likely interdisciplinary topics. For the specialized corpus, 7 of the 13 topics associated with MeSH descriptors (54%) are generated with multiple MeSH descriptors. Topics associated with multiple MeSH descriptors from different top-level MeSH categories could be used to explore the intersection of multiple domains. LDA clearly offers an advantage for discovering interdisciplinary topics.

**Limitations and future work**

One limitation of this work is that we ignored the MeSH qualifiers and only considered the MeSH descriptors when constructing our 'hybrid topics'. In the future, we will include the qualifiers to our 'hybrid topics' to test whether it improves the interpretation of topic models. We are also planning to run LDA with a larger number of topics.

## Conclusion

In this paper, we introduced an alternative LDA model by adding labels (here, MeSH descriptors) to the 'bag-of-words' assumption. With this setting, 'hybrid topics' can be generated to reveal relationships between topics and labels. In our evaluation, these 'hybrid topics' resulted in higher topic coherence scores compared the original LDA, but only on the specialized biomedical corpus. For the general corpus, we did not see a significant difference on topic quality between our 'hybrid topics' and regular topics. From our results, we can also conclude that LDA is more useful in the specialized corpus to explore concepts that may not be covered by the MeSH vocabulary and where topic models can capture aggregate concepts from different domains.

Topic models have a strong potential for analyzing the content of large text corpora. However, the deployment of topic models in the real world has been limited. Our goals in the future are to find more practical ways to apply topic models to help people better understand the massive amount of unstructured data available to us.

## Acknowledgements

This work was supported by the UTHealth Innovation for Cancer Prevention Research Training Program Predoctoral Fellowship (Cancer Prevention and Research Institute of Texas (CPRIT) grant # RP160015), the Intramural Research Program of the NIH, National Library of Medicine (NLM), and the NLM Medical informatics training program for graduate and medical students. The content is solely the responsibility of the authors and does not necessarily represent the official views of the CPRIT.

## References

[1] Z. Lu, W. Kim, and W.J. Wilbur, Evaluation of query expansion using MeSH in PubMed, *Information retrieval* **12** (2009), 69-80.
[2] R.R. Richter and T.M. Austin, Using MeSH (medical subject headings) to enhance PubMed search strategies for evidence-based practice in physical therapy, *Physical therapy* **92** (2012), 124-132.
[3] R.I. Dogan, G.C. Murray, A. Névéol, and Z. Lu, Understanding PubMed® user search behavior through log analysis, *Database* **2009** (2009), bap018.
[4] D.M. Blei, A.Y. Ng, and M.I. Jordan, Latent dirichlet allocation, *Journal of machine Learning research* **3** (2003), 993-1022.
[5] Z. Yu, T.R. Johnson, and R. Kavuluru, Phrase based topic modeling for semantic information processing in biomedicine, in: *Machine Learning and Applications (ICMLA), 2013 12th International Conference on*, IEEE, 2013, pp. 440-445.
[6] D. Ramage, D. Hall, R. Nallapati, and C.D. Manning, Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora, in: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, Association for Computational Linguistics, 2009, pp. 248-256.
[7] D. Zhu, D. Li, B. Carterette, and H. Liu, An Incremental Approach for MEDLINE MeSH Indexing, in: *BioASQ@ CLEF*, Citeseer, 2013.
[8] D. Newman, S. Karimi, and L. Cavedon, Using topic models to interpret MEDLINE's medical subject headings, in: *Australasian Joint Conference on Artificial Intelligence*, Springer, 2009, pp. 270-279.
[9] F. Doshi-Velez, B. Wallace, and R. Adams, Graph-sparse LDA: a topic model with structured sparsity, *arXiv preprint arXiv:1410.4510* (2014).
[10] D. O'Callaghan, D. Greene, J. Carthy, and P. Cunningham, An analysis of the coherence of descriptors in topic modeling, *Expert Systems with Applications* **42** (2015), 5645-5657.
[11] J.H. Lau, D. Newman, and T. Baldwin, Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality, in: *EACL*, 2014, pp. 530-539.

**Address for correspondence**
Olivier Bodenreider, MD, PhD
8600 Rockville Pike, 38A/ 09S904, Bethesda, MD 20894
Phone Number: (301) 827-4982
E-mail: olivier@nlm.nih.gov