

Analyzing NIH Funding Patterns over Time with Statistical Text Analysis

Jihyun Park
Department of Computer Science
University of California, Irvine
Irvine, CA 92697-3435

Margaret Blume-Kohout
New Mexico Consortium
Los Alamos, NM 87544

Ralf Krestel
Web Science Research Group
Hasso-Plattner-Institut
14482 Potsdam, Germany

Eric Nalisnick and Padhraic Smyth

Department of Computer Science
University of California, Irvine
Irvine, CA 92697-3435

Abstract

In the past few years various government funding organizations such as the U.S. National Institutes of Health and the U.S. National Science Foundation have provided access to large publicly-available online databases documenting the grants that they have funded over the past few decades. These databases provide an excellent opportunity for the application of statistical text analysis techniques to infer useful quantitative information about how funding patterns have changed over time. In this paper we analyze data from the National Cancer Institute (part of National Institutes of Health) and show how text classification techniques provide a useful starting point for analyzing how funding for cancer research has evolved over the past 20 years in the United States.

1 Introduction

The U.S. National Institutes of Health (NIH) invests over \$30 billion each year in scientific research and development, with the mission of promoting advances and innovations that will improve population health and reduce the burden of disease. As a result of this investment there is significant interest in quantifying and understanding public outcomes from the NIH funding process (Lane and Bertuzzi 2011; Talley, Hortin, and Bottomly 2011), for example to help determine policies for equitable and fair practices for allocation of funding resources across different diseases (Hughes 2013; Hanna 2015).

In this paper we use statistical text mining techniques to analyze how NIH funding for cancer-related research has changed over the past two decades. As the basis for our analysis we use data extracted from the NIH RePORTER online system for a 20-year period from FY1994 through FY2013. From this data we use NIH's Research Categorization and Disease Classification (RCDC) labels, project titles and abstracts, as well as annual funding amounts, to create a quantitative picture of how NIH funding for cancer has evolved since 1994.

Copyright © 2016, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

The work we describe in this paper has two main parts. First, we use text classification techniques (labeled topic modeling and logistic regression) to infer RCDC labels for projects from 1994 to 2013, inferring the probability that each grant is associated with each label. In the second part of our work we then utilize the probabilities generated by text classification to partition each grant's total funding, per year, across categories, and use the resulting information to analyze how NIH funding is changing at the category level over time.

In terms of related work, NIH grant abstracts have been analyzed using unsupervised topic modeling methods (Talley et al. 2011; Mimno et al. 2011) but without leveraging RCDC category information. In the public policy sphere, RCDC categories have been used to make broad inferences about directions in funding policy and to connect this information to external data. For example, Sampat, Buterbaugh, and Perl (2013) used RCDC categories to investigate associations between NIH funding and deaths from diseases. However, to the best of our knowledge, the work we describe here is the first to systematically RCDC labels to extrapolate funding patterns for specific research categories over a two-decade period.

2 National Cancer Institute Grant Data

We focused in this study on NIH projects that were funded through the National Cancer Institute¹ (NCI) in the years 1994 to 2013. The upper plot in Figure 1 illustrates how the NCI's budget has grown over this time period, from slightly above \$1 billion per year in the early 1990's to around \$3 billion in recent years. The 4 year-period from 1999 to 2002 corresponds to a doubling of the overall NIH budget during that time. The peaks in years 2009 and 2010 (and to a certain extent in 2011) could be viewed as somewhat anomalous since they correspond to an infusion of new funds to the NIH via the American Recovery and Reinvestment Act (ARRA). The lower plot in Figure 1 shows the number of projects

¹<http://www.cancer.gov/>

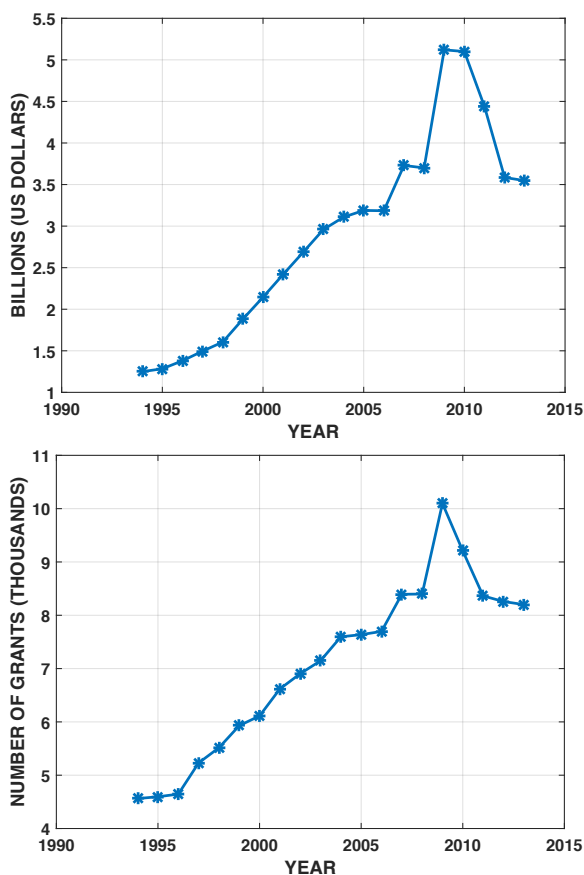


Figure 1: Plots over time of (a) the total amount of funding per year for NCI awards (upper plot), and (b) the number of projects funded per year by NCI.

funded each year by NCI, ranging between 4,500 and 10,000 per year. There is a close correlation between the number of grants per year in the lower plot and the total NCI funding per year in the upper plot.

NIH provides online public access to information on projects it funds². Each project has entries for each year it was funded, with information such as project id, project title, short summary of the project proposal (abstract), application type, support year, total cost, spending categories³. The spending categories are based on the Research, Condition, and Disease Categorization (RCDC), consisting of over 200 categories of disease, condition, or research area—they are described in more detail below.

2.1 RCDC Categories

The RCDC system⁴ consists of a set of categories developed by NIH to characterize funded research projects in a standardized manner. Examples of RCDC categories in-

²<http://exporter.nih.gov/>

³More details can be found at <http://exporter.nih.gov/about.aspx>

⁴<http://report.nih.gov/rcdc/>

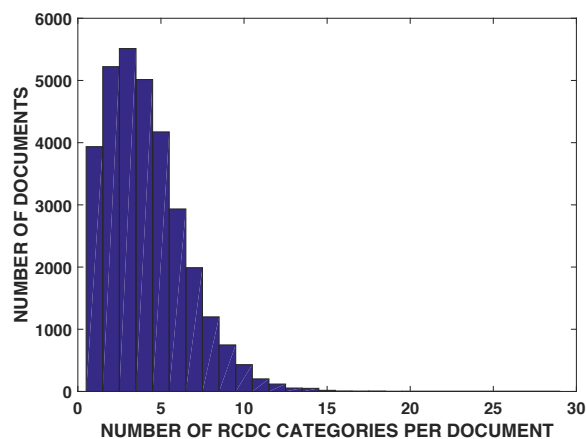


Figure 2: Histogram of the number of RCDC categories per document for the 31,628 labeled documents in the NCI data set.

clude disease categories such as *Brain Cancer* and *Liver Disease*, broad research areas such as *Biotechnology* and *Gene Therapy*, as well as general categories associated with research projects such as *Clinical Trials* and *Networking and Information Technology R&D*. RCDC categories are assigned to a project by mapping terms from a project’s title/abstract/specific aims, into a vector representation based on an RCDC thesaurus. The thesaurus is derived from well-known medical and biomedical ontologies such as Medical Subject Headings (MeSH) and the Unified Medical Language System (UMLS) metathesaurus. A system developed by NIH is then used to map term vectors to RCDC categories (Baker, Bhandari, and Thotakura 2009). The RCDC categories are not mutually exclusive, and thus, each vector (or project) can be mapped to multiple labels in the manner of a multi-label classification process.

The goal of our work is to attribute funding to specific RCDC categories over time, to derive (for example) an estimate of how much total funding was allocated by NCI to each RCDC category per year. NCI projects between 2008 and 2011 have been assigned RCDC category labels already by NIH, but projects from other years in our data do not have labels. Thus, the first part of our work involves constructing a multi-label classifier that is trained on the 2008-2011 documents (that have RCDC labels), and then using the resulting classifier to infer category labels for the projects that were funded prior to 2008 or after 2011.

3 Predicting RCDC Categories

3.1 Document Preprocessing

We downloaded from the NIH ExPORTER system⁵ all records for projects funded by the NCI over a 20-year period from 1994 to 2013. After removing project records with missing abstracts, etc., this resulted in a data set of 149,901 projects. We represented each project record using

⁵<http://exporter.nih.gov/>

a standard bag of words representation as follows. The titles and abstracts were tokenized and a set of 300 common stopwords removed. In addition to unigrams we also used n-grams in our vocabulary, corresponding to noun phrases extracted using the UIUC Chunker (Punyakanok and Roth 2001), with examples such as “radiation biology”, “research project”, and “flow cytometric analysis.” Terms that appeared in less than 15 documents were discarded, resulting in a vocabulary of 29,713 terms in total and about 50 terms on average per document.

Documents in the 4 years from 2008 to 2011 had RCDC categories associated with them and the other 16 years did not. The number of documents with labels (in the 4-year window) was 31,628 and the number without was 118,273. Figure 2 shows a histogram of the number of RCDC categories attached to each document in the labeled set. The RCDC categories present in the NCI project awards are primarily of relevance to cancer, and thus, there is a significant fraction of the categories that occur relatively infrequently in the NCI documents. Removing codes that occurred in less than 50 documents, and merging a smaller set of redundant codes, resulted in a set of 88 RCDC codes that we used in our experiments.

3.2 Learning a Labeled Topic Model for RCDC Categories

In order to analyze funding patterns across the full 20-year period we developed a document classification approach by training on the 31,628 documents with RCDC labels and then estimating RCDC category probabilities for all 149,901 documents. Our classification method consisted of two components.

The first component uses a variant of LDA (latent Dirichlet allocation, or topic modeling), called labeled-LDA (L-LDA), to learn a labeled topic model from the 31,628 labeled documents. L-LDA is an extension of the more widely-known unsupervised LDA approach, and has been found to be broadly useful for multi-label classification problems (e.g., see (Ramage et al. 2009; Rubin et al. 2012; Gaut et al. 2016)). L-LDA operates by using topics in one-to-one correspondence with labels (where by ‘labels’ we are referring here to RCDC categories). In the collapsed Gibbs sampling version of topic modeling, which is what we use in the work reported here, the sampling algorithm learns the topics by iteratively sampling assignments of individual word tokens to topics. In the unsupervised version of LDA any topic can be assigned to any word-token. In labeled-LDA the word tokens within a document can only be assigned to labels (topics) associated with that document. This provides the sampling algorithm with a significant amount of additional “semi-supervised” information compared to the fully unsupervised case, and the sampling algorithm tends to converge quickly.

We have found it useful in our work to include in the model a certain number (B) of additional “background topics”, that the sampling algorithm can use for any word token — one way to think about the background topics is that they can be associated with any documents in the corpus. These topics are useful because they can account for words in doc-

uments that are not necessarily associated with any of the known labels, e.g., for the NIH documents these topics tend to learn sets of words associated with general research topics that would be considered too general and broad to be defined as RCDC categories. In our experiments in this paper we used $B = 10$ background topics with $L = 88$ labeled topics when fitting our models. For the results in this paper we ran 70 collapsed Gibbs sampling iterations for training the model, and 20 iterations for making predictions with documents without labels. Our results appeared to be relatively insensitive to these specific choices of parameters and algorithm settings.

The L-LDA model requires the specification of hyperparameters for (a) the Dirichlet prior parameters for the word-topic distributions (the β 's, one β_w value per term w in the vocabulary), and (b) the Dirichlet prior parameters for the document-topic distributions (the α_c 's, one per topic). We followed the usual convention of setting all the β 's equal to each other (for a symmetric prior) with each $\beta_w = 0.01$. For the α_c values, we set the priors on a document-by-document basis as follows. In a particular document d in the training data, α_c^d is either proportional to the frequency with which the label c occurs across documents in the training data (if label c is attached to the document) or $\alpha_c^d = 0$ if the document does not have this label. The sum of the non-zero α_c 's is set to 5 for each document. The background topics are set to be equally likely across all training documents with $\sum_b^B \alpha_b = 1$. These same settings were used for both training with labeled documents and prediction with unlabeled documents. For prediction, when we have no labels for a document d , we use the “proportional” α_c 's as described above, with none set to 0, and $\sum_{c=1}^{88} \alpha_c = 5$.

We found that the performance of the model was not particularly sensitive to the β values, but could be somewhat sensitive to how the α 's were set, consistent with prior findings on the effect of priors in LDA (Wallach, Mimno, and McCallum 2009). We found some evidence for example that larger α values could give better test performance—but these models had poorer calibration, so we did not use the larger α 's in the results shown here. How to set the priors in labeled-LDA in a systematic way, to optimize a metric such as precision, is beyond the scope of this paper, but is an interesting avenue for future work.

3.3 Evaluation of the Topic Model

Table 1 shows the highest-probability words for several of the topics learned by the model, both for topics that are in one-to-one correspondence with RCDC categories, as well as for background topics. We can see for various disease topics (corresponding to known RCDC categories), such as *Brain Cancer*, *Breast Cancer*, and so on, that the words and n-grams that have high probability for that topic appear to be appropriate. In addition we show the high probability words for 3 of the 10 background topics, corresponding to general themes such as training/education, mouse models, and meetings/conferences. These are common themes in NIH research but are not explicitly identified as categories in the RCDC system. The labeled LDA model is able to ac-

Topic	Most Probable Terms
Brain Cancer	glioma, gbm, brain_tumor, malignant_glioma, glioblastoma, brain
Breast Cancer	breast_cancer, women, breast_cancer_cell, breast, breast_cancer_patient, brca1
Kidney Disease	rcc, kidney_cancer, renal_cell_carcinoma, vhl, renal_cancer
Hepatitis	hcv, hbv, liver_cancer, hepatitis_virus, hbv_infection, hbv_replication
Lung Cancer	lung_cancer, nsclc, lung, leading_cause, cancer_death, egfr
Mind and Body	life, quality, intervention, distress, women, exercise
Obesity	obesity, physical_activity, diet, association, change, bmi
Pediatric	children, childhood_cancer, parent, age, neuroblastoma, adolescent
Background1	program, trainee, university, training, candidate, field
Background7	model, mice, work, experiment, human, mouse_model
Background9	meeting, field, conference, researcher, area, collaboration

Table 1: The most probable terms inferred for the topics associated with RCDC codes as well as for background topics, obtained from training an L-LDA model on 31K documents labeled with RCDC categories. The 6 most probable n-grams are shown, i.e., that have the largest values $p(w|c)$ where w is a term in the vocabulary and c is a category.

count for words associated with these topics rather than being forced to assign them to the RCDC categories.

In addition to visual inspection we also conducted internal train-test evaluations by partitioning the 31,628 labeled documents into a randomly selected training set and test set with about 90% of the data in the training set and the other 10% in the test set. To make predictions on the test set we fit the L-LDA model to the training data and then ran collapsed Gibbs sampling on each test document, where the topic-word probabilities were treated as known and fixed, but the document-topic probabilities and the assignments of word tokens to topics were treated as unknown. After the sampler converges on each document d we take the sampling probabilities computed during the final iteration for each word token w_i in that document, $p(c|w_i, d)$, $c = 1, \dots, 88$, and compute the sum of these probabilities over all word tokens w_i (in document d) to get a score for each category in each document, i.e., $s(c|d) = \sum_i p(c|w_i, d)$. In effect this is the expected number of word tokens in document d that the model estimates will be associated with category c . We can normalize these numbers across the categories to obtain a conditional probability $p(c|d) = s(c|d) / \sum_k s(c = k|d)$, which can be interpreted as the probability (according to the model, given the word tokens observed in the document) that a randomly selected word token in this document will be assigned to topic c . These probabilities, $p(c|d)$ can then be used to rank documents for a given topic c (allowing computation of AUC (area-under-the-curve) scores, per topic, relative to

the known labels for the test documents), and can be thresholded to provide document-specific predictions.

Treating each label as a binary label we computed AUC and R-precision metrics on the test data using the L-LDA model’s $p(c|d)$ scores. We found that the model was able to achieve an average AUC (area-under-the-curve) score of 0.80 and an average R-precision number of 0.56, using weighted averaging across all 88 labels with weights proportional to the number of documents in the test data that contained each label. In diagnosing the predictions of the model we noticed that the predictions from L-LDA (the $p(c|d)$ scores) were not well-calibrated in the sense that they tended to spread probability mass among many labels for a given document, rather than concentrating the mass on a small subset of labels. To address the calibration issue, we added a logistic classifier component to our setup. We fit a binary logistic regression classifier⁶ for each category c on the training data, using the $p(c|d)$ scores from L-LDA as input features (so, 88 inputs plus an intercept) — this serves to calibrate the relatively poorly-calibrated L-LDA scores. On the test data documents we then used the 88 logistic models, with the L-LDA inferred scores $p(c|d)$ as inputs, to produce a probability $p_l(c|d)$ for each document for each category from the logistic models. We found that these predictions were much better calibrated than the original L-LDA scores. Furthermore, when we computed the weighted average of AUC and R-precision scores (as before for the L-LDA scores) we found significant improvements in performance: the average AUC was 0.89 and the average R-precision was 0.64.

In making predictions for the full NCI data set, we repeated the same procedure as above, consisting of 3 steps:

- L-LDA and logistic regression were fit using all of the 31,628 labeled documents (in the manner describe above for the training data);
- L-LDA was then used to infer $p(c|d)$ scores (via Gibbs sampling) for all 149,901 documents in the NCI corpus (across both labeled and unlabeled documents);
- Finally the trained logistic regression model was used to generate probabilities $p_l(c|d)$ for each of these documents and for each RCDC category $c = 1, \dots, 88$.

In this manner we were able to extrapolate estimates of RCDC categories from the labeled subset to the full data set.

4 Analyzing Funding Patterns over Time

4.1 Funding per RCDC Category per Year

We use the probabilities $p_l(c|d)$ produced by logistic regression, for each document d and each category c as described above, to infer how much funding to attribute to each category c . Let x_d be the amount of funding awarded to document d and let y_d be the year in which the funding was awarded. For a given document d we would like to be able

⁶The SciKit Learn Python library was used to obtain all logistic regression results. All settings were left at their defaults. Cross-validation runs showed that performance appeared to be insensitive to optimization and regularization parameter choices.

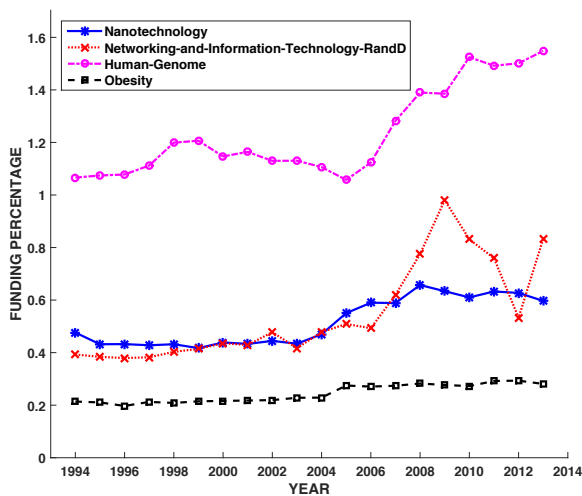


Figure 3: Estimated percentage of funding allocated to 4 general RCDC categories .

to distribute the funding amount x_d across the different categories. We adopt a simple approach and fractionally assign the funds in direct proportion to the logistic class probabilities $p_l(c|d)$. Specifically, we define a set of weights

$$w_{cd} = \frac{p_l(c|d)}{\sum_k p_k(c = k|d)}, \quad c = 1, \dots, 88$$

for each document d , where the weights sum to 1 by definition for each document across the categories. The amount of funding attributed in document d to category c is then defined as $w_{cd}x_d$.

In this manner we can compute the total estimated amount of funding for category c in a given year y by summing up all the attributed funding for category c across documents, i.e.,

$$F_c^y = \sum_{d:y_d=y} w_{cd}x_d,$$

with $c = 1, \dots, 88$, $y \in \{1994, \dots, 2013\}$. From these numbers we can for example compute the relative percentage of funding awarded in year y to the different RCDC categories (as we do in the next section).

In addition, if we wish to know what fraction of projects (independent of funding) is devoted to category c in year y , we can estimate this via $P(c|y) = \frac{1}{N_y} \sum_{d:y_d=y} w_{cd}$ where N_y is the number of projects that were funded in year y .

4.2 Funding Patterns over Time

Using the methodology described in the last section, we show in Figure 3 the estimated funding percentage as a function of time for 4 different RCDC categories, all of which are somewhat broader and more general than cancer research. All 4 categories exhibit systematic increases in their portion of funding over time, which is not surprising given that these are categories that have received increasing attention in recent decades for a variety of reasons. The

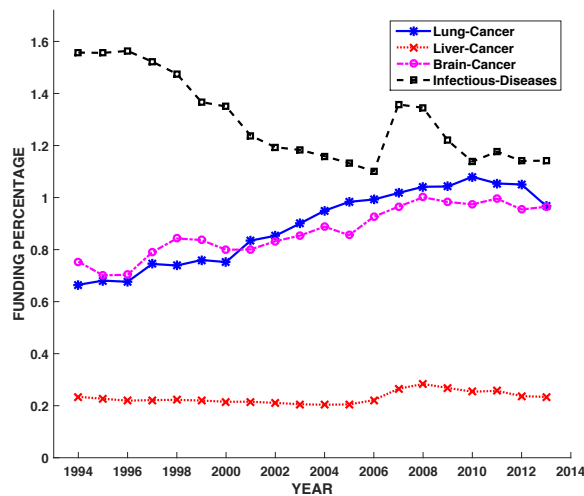


Figure 4: Estimated percentage of funding allocated to 4 specific RCDC disease categories.

Human-Genome and *Nanotechnology* exhibit systematic increases from the mid 2000s onwards. The *Networking-and-Information-Technology* category also begins to increase its funding share from 2007 onwards although it decreases after 2009—the two peak years of 2009 and 2010 correspond to the years of ARRA funding in NIH. The *Obesity* category accounts for a smaller percentage of NCI funding than the other categories, but exhibits a visible systematic shift in level of funding from 2005 onwards.

Figure 4 focuses on RCDC categories associated with specific diseases. *Brain Cancer* and *Lung Cancer* show consistent and sustained increases in their share of funding over most of the 20-year period. The *Infectious Disease* category shows a steady decline until 2006, with a subsequent increase in funding in 2007, but declining again from 2008 onwards. The *Liver Cancer* category has a much smaller share of funding than the others, and appears to have undergone a systematic increase in funding between 2006 and 2008.

Figure 5 shows plots for three of the RCDC categories that each had a relatively stable share of funding but then experienced significant shifts around 2006-2008. *Breast Cancer* starts a steady decline in 2007 that continues through 2013. *Translational Research* and *Epidemiology-and-Longitudinal-Studies* both appear to have seen rapid increases in their share of funding around 2009 and 2010, again coinciding with the two years of NIH ARRA funding.

The changes that we see in these plots may be due to a number of different factors, for example, known interventions such as the ARRA funding program, or other less well-known policy changes and shifts within NIH. The time-series plots in these figures cannot on their own provide precise understanding of the factors that are governing NIH funding over time. Nonetheless, the patterns are suggestive of systematic changes in NIH funding and can provide decision-makers and policy-makers (both inside and outside NIH) with a broad quantitative look at how funding allocations have been changing over time.

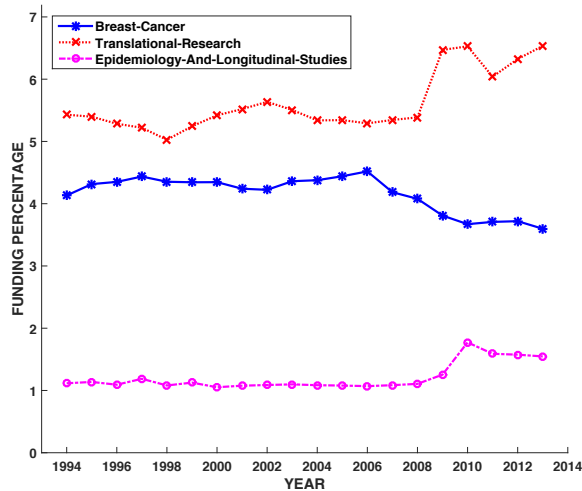


Figure 5: Estimated percentage of funding allocated to 3 RCDC categories with changes in percentages during the late 2000's.

4.3 Funding Ratios over Time

We can also use this data to analyze how the share of funding that is being assigned to a category each year *relative to the fraction of documents assigned to the category in the same year*, by computing the ration of (a) the normalized F_c^y score (normalized across categories) and (b) $P(c|y)$. For example, if a particular category has a 3% share of funding, and also a 3% share of grants assigned to this category, then the ratio is 1. If a category is getting a greater share of funding than grants the ratio will be above 1, and if it is getting less of a share of funding than grants, then the ratio will be less than 1. We explore some of these ratios in Figure 6 for 5 RCDC categories. The dotted black line at $y = 1$ is what we would expect to see if all categories had the same funding and document shares over time. What we see instead is that there is significant and systematic variation in these ratios. The two cancer categories have ratios lower than 1 indicating a lower ration of funding share to document (grant) share, and for the category *Breast Cancer* there is evidence of a systematic decline over time. Conversely, for the 3 other categories plotted in Figure 6 the ratio is consistently above 1 indicating a greater share of funding for these categories. There is also a quite striking increase in this ratio over time for all 3 of these categories, with the ratio for the *Networking* category effectively doubling over time. These types of patterns could (for example) reflect systematic differences and shifts in project budgets across different categories, with some categories tending to have more expensive budgets than others.

5 Conclusions

In this paper we have shown how labeled topic-modeling and logistic classifiers can be combined to analyze NIH grant funding data and to extract potentially useful information. Directions for further research include both methodological issues such as investigating how best to calibrate the L-LDA model, as well as broader analysis of the economic

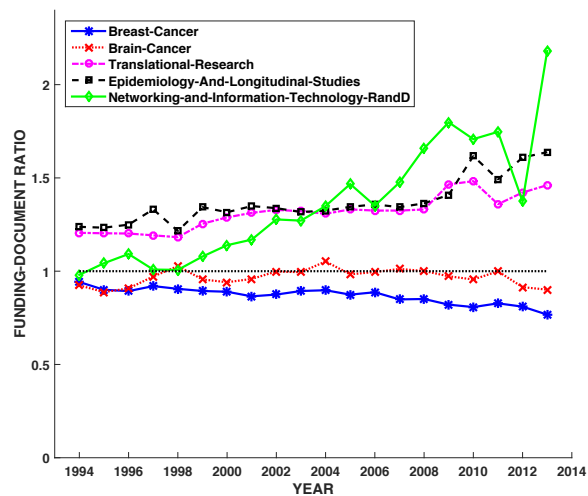


Figure 6: Ratio of funding percentage to document percentage over time for 5 RCDC categories.

and policy implications of the extracted information. Finally, it is also of significant interest to policy analysts to not only look at grant abstracts but to other relevant data, such as scientific articles that resulted from the funded grants, citation data, and so on.

Acknowledgements

This work was supported by the US National Science Foundation under award number NSF 1158699. The authors also gratefully acknowledge the suggestions of the reviewers.

References

- Baker, K.; Bhandari, A.; and Thotakura, R. 2009. An interactive automatic document classification prototype. In *Proc. of the Third Workshop on Human-Computer Interaction and Information Retrieval*.
- Gaut, G.; Steyvers, M.; Imel, Z. E.; Atkins, D. C.; and Smyth, P. 2016. Content coding of psychotherapy transcripts using labeled topic models. *IEEE Journal of Biomedical and Health Informatics* in press.
- Hanna, M. 2015. Matching taxpayer funding to population health needs. *Circulation Research* 116(8):1296–1300.
- Hughes, V. 2013. The disease olympics. *Nature Medicine* 19(3):257–260.
- Lane, J., and Bertuzzi, S. 2011. Measuring the results of science investments. *Science* 331(6018):678–680.
- Mimno, D.; Wallach, H. M.; Talley, E.; Leenders, M.; and McCallum, A. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 262–272. Association for Computational Linguistics.
- Punyakanok, V., and Roth, D. 2001. The use of classifiers in sequential inference. In *Advances in Neural Information Processing Systems*, 995–1001.

- Ramage, D.; Hall, D.; Nallapati, R.; and Manning, C. D. 2009. 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*, 248–256. Association for Computational Linguistics.
- Rubin, T. N.; Chambers, A.; Smyth, P.; and Steyvers, M. 2012. Statistical topic models for multi-label document classification. *Machine Learning* 88(1-2):157–208.
- Sampat, B. N.; Buterbaugh, K.; and Perl, M. 2013. New evidence on the allocation of NIH funds across diseases. *Milbank Quarterly* 91(1):163–185.
- Talley, E. M.; Newman, D.; Mimno, D.; Herr II, B. W.; Wallach, H. M.; Burns, G. A.; Leenders, A. M.; and McCallum, A. 2011. Database of NIH grants using machine-learned categories and graphical clustering. *Nature Methods* 8(6):443–444.
- Talley, D.; Hortin, J.; and Bottomly, J. 2011. Information needs of public policy lobbyists. In *Proceedings of the iConference*, 781–782.
- Wallach, H. M.; Mimno, D. M.; and McCallum, A. 2009. Rethinking LDA: Why priors matter. In *Advances in Neural Information Processing Systems*, 1973–1981.