**Final Report of Work on SBIR Contract HHSN271201700041C**

**NIDA Reference Number N43DA-17-1215**

**“Creation of an Accurate Model of the Topical Structure of PubMed and Associated Indicators”**

**SciTech Strategies, Inc., Contractor**

**Stanford University, Subcontractor**

**Executive Summary**

Research indicators, such as field-weighted citation impact and cites per paper, are currently based on high-priced citation databases such as the Web of Science and Scopus. Such indicators are out of reach financially for many institutions (e.g., agencies, universities, NPOs) and researchers. There is an established need for a low-cost source of scientific indicators that can be used by institutions and researchers.

In addition, there is a need for new indicators associated with the virtue of research that quantify things such as grant funding patterns, quality, reproducibility, and potential for translation. These new indicators can be associated not only with research evaluation, but with planning as well. Research planning also requires an accurate detailed model of scientific topics. Journal-based classification systems (e.g., journal subject categories) are not specific enough to enable decision making at the topic level.

Overall, there is a significant need for a low-cost web-based system of research topics and indicators aimed specifically at research planning.

This SBIR project is designed to meet these needs. The ultimate result of this project, if funded through Phase II, will be the creation of such a system. Phase I of this SBIR project has been focused on some technical challenges related to this overall goal, especially the use of open source data to avoid the costs associated with licensing of a name-brand citation database. The following challenges have been successfully addressed:

* An accurate, detailed and comprehensive model of the biomedical literature has been created from open source (PubMed) data. We have shown that the new model is as accurate as the current best-in-class model created using citation data ([Klavans & Boyack, 2017a](#_ENREF_21)). This model assigns over 23 million PubMed records to approximately 54,000 scientific topics.
* Using open source data from PubMed and PubMed Central (PMC), we have developed surrogate indicators that approximate two of the most popular citation-based indicators in use today (citations per paper and field weighted citation impact) for these 54,000 topics.
* Work to develop new indicators related to funding, quality, and reproducibility that reflect the virtue or responsibility of the science has also been done. For example, we have accurately assigned project-level funding data from the U.S. Star Metrics database to the topics in our model. Using PubMed and PMC data, we have developed a surrogate indicator that approximates funding levels by topic and can serve as a topic-level funding predictor. This new indicator is just as accurate as a similar indicator we recently developed based on citation and usage data from Scopus ([Klavans & Boyack, 2017a](#_ENREF_21)).

Results from the model, and impact and funding indicators above are based on comparison to values from Scopus and Scopus-related offerings. An accurate detailed model created from open source data that includes competitive impact and funding indicators will be competitive with the state-of-the-art and can serve as the backbone of our proposed system.

Results on the development of indicators related to the virtue of science are not yet definitive but are promising. We fully expect that these indicators will prove to be useful with further development. We also note that these new indicators represent capabilities that cannot currently be found anywhere, and we are thus at the forefront of development of these types of indicators.

Regarding the business opportunity that our proposed system will address, we note the following. Currently, there are three major providers of large-scale citation analytics: Clarivate (Web of Science and InCites), Elsevier (Scopus and SciVal), and Digital Sciences (Dimensions). These offerings represent two price/quality positions – Clarivate and Elsevier at the high end, and Digital Sciences at the low end). Our current business strategy, to be further developed in Phase II, is briefly reflected by the following points.

* Although our proposed system will be suitable for evaluation, we will market it with a focus on planning. We have independently been developing a proposal analytics capability that will be a new offering in the marketplace in a soon-to-be registered B-corporation. This capability can accurately assign proposals to topics in our model and thus enables planning at the proposal level using advanced indicators. There are currently no competitors in this market.
* We will price our system lower than Digital Sciences with comparable quality for a specialized application. Unlike Digital Science (which is broadly focused on all nations and all areas of research), we will initially focus on research funded in the United States that is of benefit to the health and well-being of all people.

We would appreciate the opportunity to provide more information to NIH/NIDA regarding our Phase 2 proposal to further describe our strategy and the social benefits we intend to provide.

**Final Report of Work on SBIR Contract HHSN271201700041C**

**NIDA Reference Number N43DA-17-1215**

**“Creation of an Accurate Model of the Topical Structure of PubMed and Associated Indicators”**

**SciTech Strategies, Inc., Contractor**

**Stanford University, Subcontractor**

**Background and objectives**

Research assessment and planning exercises at government agencies and research institutions are based predominantly on high-priced citation-based databases and associated indicators. The best indicators are based on normalization techniques that require an accurate delineation of the scientific literature into topics and disciplines ([Ioannidis, Boyack, & Wouters, 2016](#_ENREF_17); [Waltman, 2016](#_ENREF_29)). Citation-based databases (i.e., the Web of Science and Scopus) have a distinct and well-earned advantage in this market in that they have broad coverage of the literature, are of relatively high quality, and because they include cited references. Citation-based indicators, as a proxy for impact, dominate research assessment. Citation links have also been shown to produce very accurate delineations of the literature into topics ([Klavans & Boyack, 2017b](#_ENREF_22)) that can then be used as the basis for citation-based indicators and the identification of hot and/or emerging topics ([Small, Boyack, & Klavans, 2014](#_ENREF_27)). Useful and accurate alternatives to this citation-dominated situation could increase the availability of advanced analytics to those who cannot currently afford them.

At the same time, there is increasing recognition of the prevalence of a reproducibility crisis in biomedical research and beyond. Work at Stanford in the Meta-Research Innovation Center under the leadership of John Ioannidis has tried to promote the empirical study of research practices, promote best practices and enhance quality, reproducibility, and translational potential for research ([Goodman, Fanelli, & Ioannidis, 2016](#_ENREF_13); [Ioannidis, Fanelli, Dunne, & Goodman, 2015](#_ENREF_18)). Empirical data suggests that most biomedical fields currently score very low in these aspects ([Begley & Ioannidis, 2015](#_ENREF_2); [Ioannidis, 2014](#_ENREF_15), [2016](#_ENREF_16); [Iqbal, Wallach, Khoury, Schully, & Ioannidis, 2016](#_ENREF_19)), but there is a lot of interest to improve both the efficiency and transparency and quality of research investigations.

The ***opportunity*** exists to develop an alternative to citation-based modeling of science and indicators of science and technology that can be used for research assessment and funding decisions. We intend to create an accurate model of biomedical science at the topic level. We also intend to develop indicators of impactful science and of how virtuous the research is (e.g., transparency, reproducibility and potential for translation/application), and correlate them with existing citation-based indicators. This work will be based on PubMed data, an open literature database available to all, and the new metrics will be based on open data.

The main ***objectives*** of this project are:

1. To create a set of topics from the open biomedical literature that are as accurate as those than can be created from citation data,
2. To create and provide accurate impact indicators at the topic level using open data,
3. To create metrics addressing quality, transparency, reproducibility and the potential for translation, and to relate these to the topic structure.

Successful accomplishment of these objectives will result in important information that can be continuously updated and will provide an observatory of quality, transparency, reproducibility and translational potential for biomedical research, and for specific topics.

Creation of an accurate set of topics from the biomedical literature will also enable accurate identification of productive scientists, and potentially fruitful collaborations, and will help identify and monitor areas where transparent and reproducible research practices can thrive or attention is needed for their improvement.

**Contractor and subcontractor**

The contractor for this work is SciTech Strategies, Inc., a small business with extensive experience in bibliometrics research. The SciTech PI’s for this work are Dr. Kevin W. Boyack and Dr. Richard Klavans. A portion of the work is subcontracted to the Meta-Research Innovation Center at Stanford University. The Stanford PI for this work is Dr. John P. A. Ioannidis.

**Tasks and schedule**

The proposed tasks and schedule from the proposal are shown in Table 1. This task list will be used to document progress on the project, and the report will be structured by task.

**Table 1. Tasks and proposed schedule for this SBIR Phase 1 project.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Sept 17** | **Oct 17** | **Nov 17** | **Dec 17** | **Jan 18** | **Feb 18** |
| 1. Update databases (SciTech) |  |  |  |  |  |  |
| 2. Create models (SciTech) |  |  |  |  |  |  |
| 3. Impact measures (SciTech) |  |  |  |  |  |  |
| 4. Virtuous measures (Stanford) |  |  |  |  |  |  |
| Documentation (Both) |  |  |  |  |  |  |

All work will be documented monthly and at a level of detail that supports replication of work. This report will be of the running documentation style. In other words, this report will be updated and augmented each month, and will thus include all work to date on the project rather than accounting only for the work done in a particular month. This report will be available to all of those working on the project, and will be transmitted to NIDA each month as indicated in the contract.

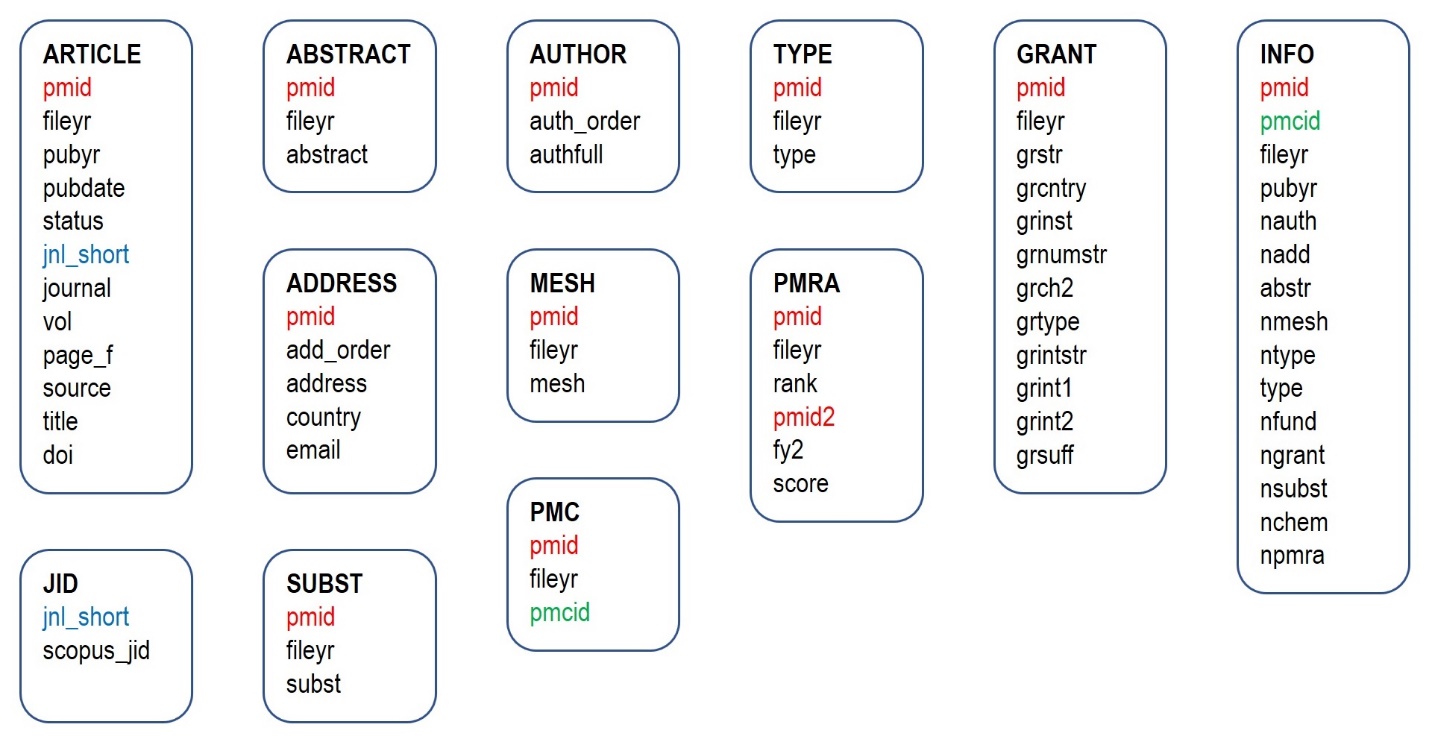
**Task 1: Update of the databases needed for Phase I work**

As of the proposal date, SciTech’s in-house versions of the PubMed and Scopus databases only contained records through September 2015 and December 2015, respectively. For the purposes of this project, both databases need to be current through roughly the start date of the project, and were updated accordingly as detailed below.

***1.a: Update SciTech’s PubMed database to be current as of the project start date***

SciTech’s in-house instance of the PubMed database was updated during the first two weeks of September 2017. The first step in this process was to obtain a complete set of PubMed records. This was done by querying PubMed through its online interface and downloading files containing annual sets of records. For example, a set of records for 2016 was retrieved using the query ("2016/1/1"[Date - Publication]: "2016/12/31"[Date - Publication]), and then downloaded in MEDLINE format using the “Send to: File” capability on the query webpage. This was done for each year back to 1960. Although XML is typically a preferred file format, we used the MEDLINE format because we have tools developed many years ago that are very good at extracting information from tagged record types of files.

Once the files were obtained, they were processed to extract data that could be fielded in a MySQL database. One of the key processing steps was to deduplicate the records. In some cases, the same record will be retrieved using queries for more than one publication year. For each record occurring more than once, the record in the most recent year was kept and loaded in the database. Deduplication was based on the unique PubMed identifier (PMID). The schema for the SciTech PubMed MySQL database is shown in Figure 1.



**Figure 1. Schema of the SciTech PubMed database. Fields used as keys to link tables are colored (red, green, blue).**

Although most fields were simply loaded in the database, some fields were calculated prior to loading. For instance, author order and address order are not fields in the raw data, but can be inferred from the order in which each occurs in the data records. The GRANT and INFO tables consist mostly of calculated fields. For the GRANT data, the raw grant strings are processed to separate country, institute (agency), and type, IC code and grant number for NIH grants. Some suffixes are also parsed. The INFO table contains counts of various features (e.g., authors, MeSH terms, grants, etc.) for each record.

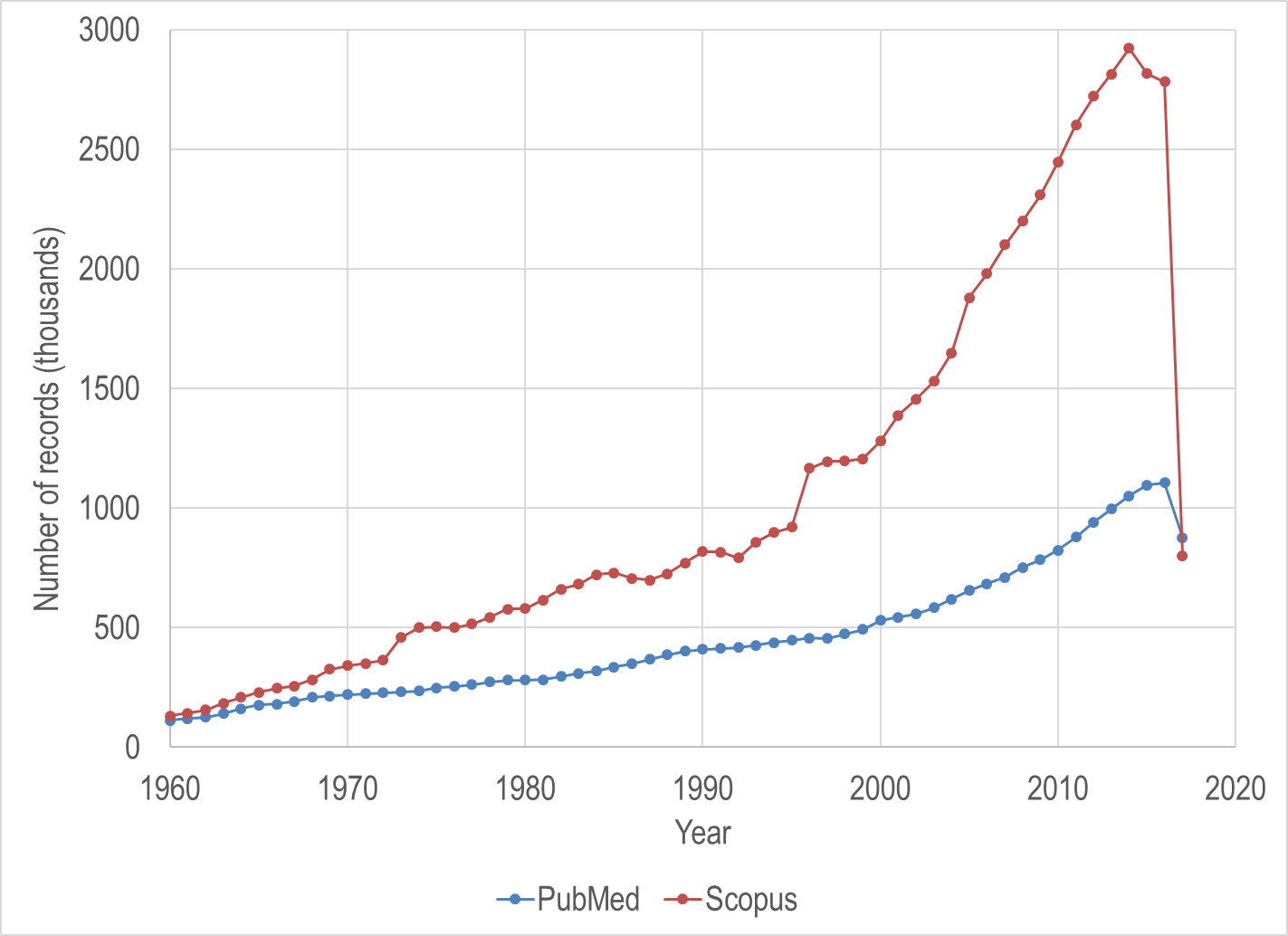
Two of the tables in Figure 1 need additional explanation. The JID table contains a list of 23,312 PubMed journals that we have specifically linked to Scopus journal IDs over many years of maintaining PubMed and Scopus databases. These matched pairs are used to help us accurately link individual PubMed and Scopus records, as will be described later.

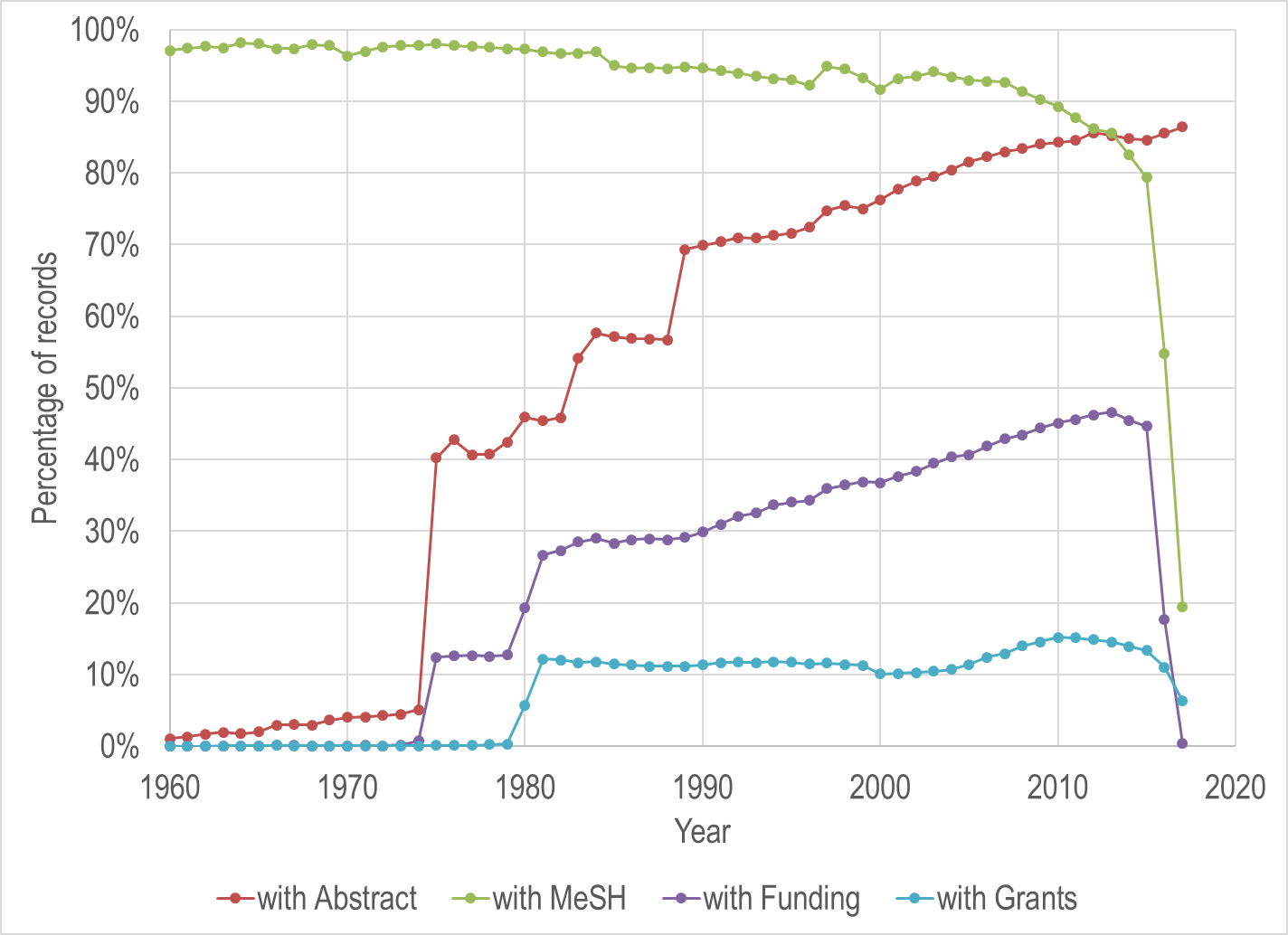
The PMRA table contains PubMed related article (or “similar article” or “neighbor”) scores. PubMed has calculated the top 100 scoring neighbors for each record from titles, abstracts and MeSH terms using the similarity method of Lin & Wilbur ([2007](#_ENREF_24)). Thus, for each record, the PMRA scores identify the top 100 most similar articles in PubMed based on textual similarity. Note, however, that while these scores are used by PubMed to populate the “similar articles” portion of their webpages for each article, the scores are not in the raw PubMed data. Fortunately, however, there is a way to retrieve these scores using an Entrez e-utility.[[1]](#footnote-1) For example, the query to retrieve the neighbor scores for PMID=20210808 is:

<https://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&db=pubmed&id=20210808&cmd=neighbor_score>

These scores cannot be retrieved all at once. Entrez has rules governing the size and frequency of queries that can be made. We retrieved the neighbor scores for all PMID in PubMed as of September 2015 over a three-month period from October - December 2015. In September 2017, we retrieved the neighbor scores for all PMID that have been added to PubMed since September 2015, thus bringing our PMRA tables current. Our PMRA tables currently contain 4.38 billion neighbor scores between pairs of PubMed records from 1960 through September 2017.

In total, our PubMed database is comprised of 26,034,934 records dating from 1960 with an annual distribution as shown in Figure 2. PubMed experienced roughly linear growth from 1960 to 1997, after which time the growth rate has increased substantially. Figure 2 (bottom) also shows that not all PubMed records contain the same types of information. For instance, prior to 1975, less than 5% of PubMed records contained an abstract. In 1975 that number jumped to 40%, and it has risen consistently since then to where it currently contains abstracts for over 85% of records in the most recent couple of years. Nevertheless, this suggests that some articles will be clustered less accurately than others because they do not contain as much information that could be used for clustering or classification. MeSH terms are available for most records, with over 90% coverage until 2010. The severe dropoff in the percentage of records with MeSH terms over the most recent couple of years suggests that indexing of records for the assignment of MeSH terms is not keeping pace with the addition of records to PubMed. General funding type acknowledgments and specific grant acknowledgments are available for an increasing fraction of records, but coverage of these features was not substantial until 1980.





**Figure 2. Numbers of PubMed and Scopus records per year (top) along with the percentage of PubMed records that contain abstracts, MeSH terms, general funding acknowledgments, and specific grant information (bottom).**

Although not shown here, we note that an average of nearly 12 MeSH terms have been assigned to each record that has MeSH terms for the past 15 years. Numbers were slightly lower before that time. The average number of specific grants acknowledged for papers with grant acknowledgments was constant at about 1.7 from 1975 through 1996, after which time it more than doubled to over 4.0 in 2010. This number has decreased again in recent years. We note that PubMed does not index all grants, but rather only those from a select set of agencies that are known to fund biomedical research.

***1.b: Update SciTech’s Scopus database***

SciTech’s in-house instance of the Scopus database was updated in June 2017. Although a September 2017 update would have been more optimal for this project, a June update was already scheduled with Elsevier, and shorter-term updates are not easy to obtain. The three-month difference between update should not negatively affect results of the project since we are likely to use citations counts through 2016 later in this project as a basis of comparison for the new impact metrics that will be developed. Numbers of Scopus records per year are shown in Figure 2 (top) along with the PubMed information. Scopus claims to incorporate all PubMed records, which suggests that PubMed has comprised around 35% of Scopus since the early 2000s.

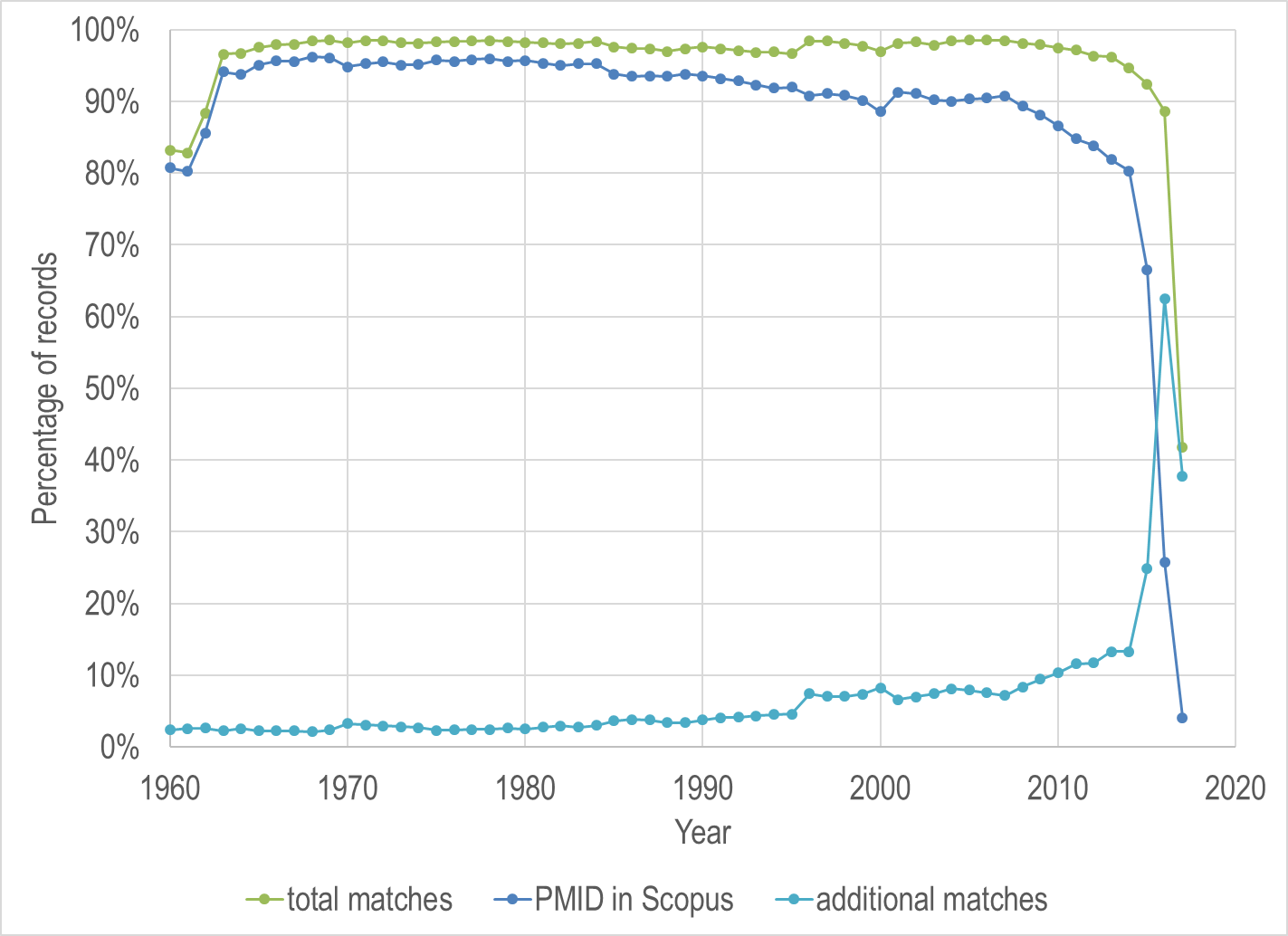
To verify this claim, and also to ensure that we have the best possible set of matches between PMID and Scopus article IDs, we have done our own matching of records from the two databases. Our matching process consists of a series of steps in which different matching criteria are used. If a match is found at one step, the matched Scopus record is not allowed to participate in subsequent matching steps. The most specific criteria is used first, and the subsequent criteria become looser. The criteria used, in order, are:

1. Journal (using the JID lookup table) AND (Volume OR Year) AND Page AND Title
2. Journal AND (Volume OR Year) and Title
3. Journal AND Title
4. Journal AND (Volume OR Year) and Author

Although most matches using these criteria produce one-to-one matches between PubMed and Scopus records, there some cases of one-to-many matches. For these, a final deduplication step is applied that scores each instance using all of the criteria listed above, and selects the match with the highest score. In addition, Scopus does contain PMID for a large majority of PubMed records, and these are compared to and merged with our match results.

Note also that we do not use the raw title in the matching steps listed above, but use the “soundex” function on the titles. This strips all non-alphanumeric characters from text, and converts the remaining text to a string based on phonetics. Two strings that sound the same, but that have different spellings, can thus have the same soundex value. Use of the soundex function allows us to match some records where there are simple misspellings or punctuation differences in the article titles between the two databases.

Figure 3 shows the overall match rate by year, which averages over 98% from 1965 to 2013. The match rate has decreased slightly in recent years. We suspect that this is primarily due to differences in the timing of when records are added to each database. In addition, the percentage of Scopus records containing PMID has decreased over time, which has necessitated our additional matching. We do not know why Scopus does not have complete PMID information for its records, but suspect that it derives from difficulties in merging records from PubMed and from publishers. Despite the lack of complete agreement, the overall matching rate is sufficiently high (97% from 1975 through 2016) to enable accurate comparisons based on Scopus data.



**Figure 3. Percentage of PubMed records that are matched to a Scopus record.**

**Task 2: Create and analyze multiple models of the PubMed database**

The purpose of this task is to create a model of biomedical science using open data from PubMed records that is just as accurate as the large-scale models of science that we have created in the past using citation data from Scopus ([Klavans & Boyack, 2017b](#_ENREF_22)). We expect to be able to create a highly accurate model of PubMed because our previous experiments have shown that the best citation-based and best text-based methods gave results of roughly similar accuracy using a dataset of 2.15 million PubMed/Scopus documents ([Boyack & Klavans, 2010](#_ENREF_5); [Boyack et al., 2011](#_ENREF_7)). Our target is to create a PubMed model that is within 5% of the accuracy that we have attained using citation-based models.

***2.a: Calculate similarity scores between pairs of documents to use in clustering***

Three separate similarity files have been created. The first two use PubMed related article scores exclusively, while the third uses the related article scores as a baseline that is then augmented in cases where the two related articles have one or more authors in common. PubMed related articles scores are based on the algorithm of Lin & Wilbur ([2007](#_ENREF_24)), which has been used at PubMed for nearly a decade, and is known as a highly accurate text measure.

We have also made the decision to cluster only those records from 1975 to the present – to exclude publications prior to 1974. The reason for this is that PubMed contains abstracts for less than 5% of the articles published prior to 1975, and we trust the related article scores less for articles published prior to 1975 than for those published beginning in 1975. Our calculations are thus limited to the 23,234,923 PubMed records from 1975 to present for which we have related article scores. The three similarity files that we have created can be characterized as follows.

1. The first similarity file contains the top 12 related article scores for each PubMed record, referred to hereafter as **12PM**. We are not aware of any clustering method that will handle the entire set of 4.38 billion textual similarities. Thus, some filtering of similarities is needed. In previous work in clustering journals, we found that not all similarities needed to be included while clustering, but rather that the accuracy of journal clusters was maximized when the top 10-15 similarities were included (unpublished work). Thus, we chose to use the top 12 related article scores for the first model. This similarity file was comprised of a total of 241,643,546 related article scores.
2. The second similarity file contains the top 40 related article scores for each PubMed record, referred to hereafter as **40PM**. In recent (unpublished) work with citation data, we found that models created using a full citation graph of 830 million similarities – the full set of references from 35 million articles – gave much more accurate results than clustering with a filtered set of references. This result is different from that of our unpublished journal studies, and suggests that paper-level clustering may give very different results than journal-level clustering. Given this result, we feel it wise to use a much larger set of related article edges to see if it will similarly increase the accuracy of the PubMed model. Given that we know that the clustering algorithm will handle around 800 million similarities, we chose to use the top 40 scores for this second model, resulting in a similarity file comprised of 790,047,772 related article scores.
3. The third similarity file is very similar to the first. It contains the top 12 related article scores (*Rij*) for each PubMed record, but augments that score by up to 1/3 for related articles that have authors in common (referred to hereafter as **12AU**). Author-based similarities have not been widely used in creating models of science for several reasons. Among these reasons are the author disambiguation problem, and the fact that prolific authors are known to have research portfolios that can span many topics. Thus, clustering based on authors can easily link papers that should not be linked. To reduce the effects of these two issues, we identified the set of 6.95 million authors (based on full names) with a minimum of two papers and a maximum of 50 papers in PubMed. A similarity between papers *i* and *j* based on their authorship was calculated as *Aij = Nij / √Ni Nj* , where *Nij* is the number of authors (from our filtered set) in common between the two papers; *Aij* varies between 0 and 1. The final similarity for each pair of papers was calculated as *scoreij = Rij \* (1 + Aij /3)*. Thus, scores remain unchanged for pairs of papers without a common author, and can be augmented by a factor of up to 1/3 for pairs of papers with common authors. Within this file, 12,364,441 pairs of paper had authors in common. Thus, only 5.12% of the papers with high textual similarity have authors in common. Given this low fraction, we do not expect the clustering result to be very different from the first similarity file.

One might ask the question as to why we did not use the author-based similarity on its own. The answer is that even though there was a large number of author-based similarity pairs (401.5 million), these authors are only associated with 19,196,910 of the documents. We do not consider 82.5% coverage to be high enough for the model. We could increase the upper threshold to include authors with more than 50 papers. This would increase coverage, but it would also significantly increase the number of similarity pairs in the calculation, and would also increase the likelihood of improperly linking papers from very different topics that are authored by researchers with broad research interests and portfolios.

***2.b: Create models (sets of clusters) of the PubMed database***

The three similarity files mentioned above were each clustered using the modularity-based smart local moving (SLM) clustering methodology and algorithm from CWTS at Leiden University ([Waltman & van Eck, 2013](#_ENREF_31)) that has recently been shown to be among the most accurate clustering algorithms available ([Emmons, Kobourov, Gallant, & Börner, 2016](#_ENREF_11)). The SLM algorithm is written in Java and can be freely downloaded (<http://www.ludowaltman.nl/slm/>) for use by anyone. It allows a resolution value to be specified – these can be tuned to produce a desired number of clusters. We desire solutions with approximately 50,000 clusters given that these PubMed models will contain roughly half as many records as our most recent citation-based models with 100,000 clusters. This should result in cluster size distributions that are similar to the cluster size distribution of our Scopus model. This is important since we have already shown that models with this type of distribution can be used to successfully identify emerging topics ([Small et al., 2014](#_ENREF_27)) and can be successfully linked to funding data ([Klavans & Boyack, 2017a](#_ENREF_21)). These are both important features associated with topics because they allow topic-level information to be associated with policy and decision making.

We also clustered each similarity file hierarchically to create more highly aggregated models with roughly 5,000 and 500 clusters. This was not done because we envision using these models with fewer clusters, but rather because they provide a point of comparison with other approaches where we have done similar hierarchical clustering. The resolutions, desired numbers of clusters, and minimum cluster sizes used for the clustering runs are shown in Table 2.

**Table 2. Input characteristics for the SLM clustering runs. Values in italics are the input resolution values.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Set** | **# Similarities** | **~105 clusters** | **~104 clusters** | **~103 clusters** |
| 12PM | 241,643,546 | *12787.5* | *4.25 x 108* | *2.0 x 109* |
| 40PM | 790,047,772 | *23250* | *2.5625 x 109* | *2.5625 x 1010* |
| 12AU | 241,643,546 | *12785.5* | *4.25 x 108* | *4.25 x 109* |
| Desired # clusters |  | ~50,000 | ~5,000 | ~500 |
| Min cluster size |  | 30 | 300 | 3000 |

For each case, sample calculations of one iteration were run with different resolution values to determine which resolution would give a solution with the desired number of clusters. Once this value was determined, then final runs were done using 10 iterations to generate the cluster solutions. Calculations were performed on an Amazon r3.4xlarge EC2 instance with 16 cores and 122 GB memory. The largest of the calculations took roughly one day of server time to complete.

Table 3 gives the number of papers and clusters for each of the three models, along with numbers of clusters above several size thresholds. The 12PM5 and 40PM5 models are quite different. Although they have a similar number of clusters overall, the 40PM5 model has larger clusters and significantly fewer clusters with at least 100 and 500 papers, respectively. This is clarified by Figure 4, which shows the cluster size distributions. Although both models have around the same number of clusters with at least 1000 papers, the 12PM5 model has far more clusters with between 100 and 1000 papers. Flatter cluster size distributions have always been attractive to us. Our experience when examining cluster contents is that mid-sized clusters are very coherent in terms of their topical focus. Thus, we favor models with larger numbers of such clusters in that they provide large numbers of well-differentiated topics that correlate well with the perceptions of researchers regarding their research topics ([Boyack, Klavans, Small, & Ungar, 2014](#_ENREF_6)).

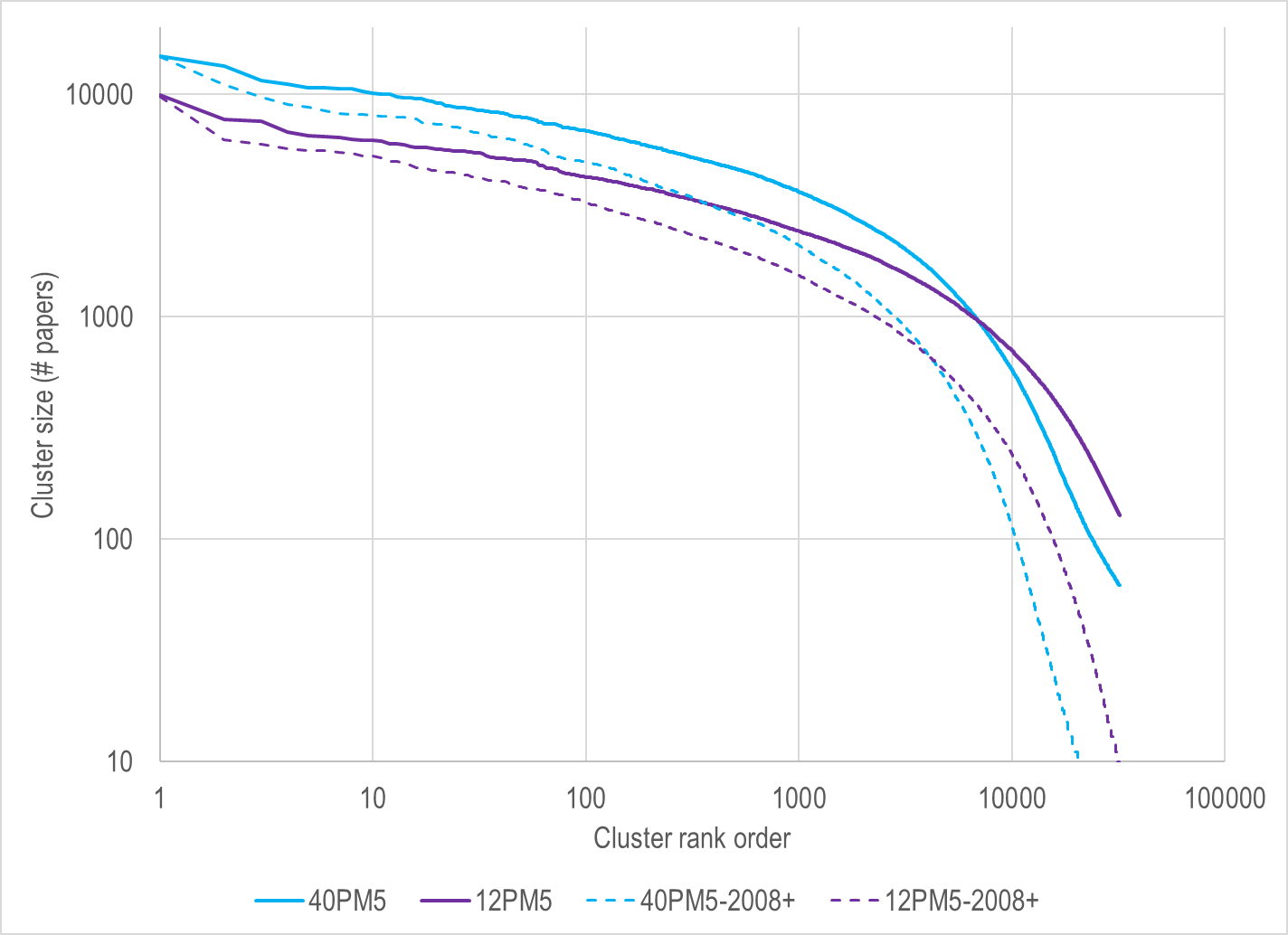
**Table 3. Characteristics of the three models with ~50,000 clusters. The superscript “5” in the model names indicates that the model has roughly 105 clusters.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | **# Clusters** | | | | |
| **Model** | **# Papers** | **All** | **>30 (% coverage)** | **>100** | **>500** | **Largest** |
| 12PM5 | 23,234,841 | 55,202 | 54,018 (99.89%) | 36,466 | 13,829 | 9,892 |
| 40PM5 | 23,234,923 | 53,819 | 46,067 (99.32%) | 23,855 | 10,897 | 14,871 |
| 12AU5 | 23,234,841 | 54,374 | 53,185 (99.90%) | 36,064 | 13,793 | 9,808 |

Table 3 also shows that 12PM5 and 12AU5 models give very similar results – the addition of co-authorship information to the textual similarity has a very small effect on the cluster size distribution in that the numbers of clusters with at least 100 and 500 papers are very similar for both models.

***2.c: Measure the accuracy of the different models and compare with the established accuracies of the best existing citation-based models***

In accordance with best practices, we seek to establish the relative accuracy of the models created in this project by comparing them with existing models of known accuracy. Best practices for comparing accuracies are still being established, and such comparisons should be done using a principled approach ([Waltman, Boyack, Colavizza, & Van Eck, 2017](#_ENREF_30)). For example, solutions should be compared using a basis that is independent of all approaches if possible. In practice, this condition is difficult to achieve. Furthermore, solutions with different granularities (i.e., different numbers and sizes of clusters) should not be directly compared because most metrics will naturally favor solutions with few large clusters (low granularity) over those with many small clusters (high granularity). Granularity should thus be accounted for in any comparative analysis.



**Figure 4. Cluster size distributions for the 12PM5 and 40PM5 models (solid lines). Cluster size distributions when limiting clusters to papers published from 2008 to present (dashed lines) are also shown for comparison.**

In practice, it is nearly impossible to generate multiple solutions with the same granularity. We overcome this by using a graphical approach that presents results as granularity-accuracy (GA) plots. Using a GA plot, the accuracies of different solutions can be compared despite differences in granularity. We define the granularity of a clustering solution obtained using relatedness measure X as

. (1)

where N is the number of papers in the solution and is the number of publications belonging to cluster for that solution. Thus, granularity is the ratio of the number of papers to the sum of the square of the cluster sizes.

In this project, we compare the accuracy of our three PubMed-based models (12PM, 40PM, 12AU) to several citation-based models from our earlier research ([Klavans & Boyack, 2017b](#_ENREF_22)), including direct citation, bibliographic coupling, and co-citation models created using Scopus data. All comparisons reported here are not based on Scopus content, but are rather based only on those documents that are available in both Scopus and PubMed. We restrict the comparison in this way to avoid biasing the results by including content not available in PubMed.

A selection of the models compared here – those with around 50,000-100,000 clusters – are listed in Table 4, along with their organizing principles or bases. Three different citation-based models are included, using direct citation, co-citation, and bibliographic coupling as their bases for document similarity. While the direct citation model covered a relatively long time period, the co-citation and bibliographic coupling models used source items from shorter time periods to keep the calculations to a manageable size. Note that these models include fewer papers than the direct citation model because not all papers are cited in any given year or time window. Two different bibliographic coupling models were created – one using a single year of source items, and one using three years of source items – to give us an idea of the effect of the time window length on accuracy. The final three items in the table are the three text-based models described above. Models with lower granularity based on these same organizing bases are also included in our graphical comparisons, but since they are more aggregated, we do not describe them here, nor do we focus much attention on them in the analysis. In addition, our graphical comparisons also include three journal-based models. The UCSD ([Börner et al., 2012](#_ENREF_3)) and ScienceMetrix ([Archambault, Caruso, & Beauchesne, 2011](#_ENREF_1)) journal classification systems and a model where each journal is considered its own category, are shown in the figures below. In nearly all cases, the journal-based models are far less accurate than any of the article-based models that we are comparing.

**Table 4. Characteristics of highly granular models created using different organizing bases.**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Basis** | **Source Years** | **Cited Years** | **# Clusters** | **#Doc < 2009** |
| DC5 | Direct citation | 1996-2012 | 1960-2012 | 91,726 | 25,723,095 |
| CC5 | Co-citation | 2011-2013 | 1960-2013 | 92,571 | 14,904,560 |
| BC5 | Bib coupling | 2010 | 1960-2010 | 101,337 | 5,599,379 |
| 3BC5 | Bib coupling | 2011-2013 | 1960-2013 | 98,313 | 11,310,070 |
| 12PM5 | Text | 1975-2017 |  | 55,202 | 15,487,903 |
| 40PM5 | Text | 1975-2017 |  | 53,819 | 15,487,937 |
| 12PM5 | Text + author | 1975-2017 |  | 54,374 | 15,487,903 |

In this work, we report three separate measures of relative accuracy. As previously mentioned, in an ideal world the comparative measures would be independent of the solutions being compared. In practice, however, this can rarely be achieved. Thus, to provide balance to the results, our comparisons are based on 1) a citation-based measure, which will likely bias toward citation-based models, 2) a text-based measure, which will likely bias toward text-based models, and 3) a grant linkage-based measure, which is relatively independent of both citation-based and text-based models.

***2.c.1: Citation-based measure***

In our previous work, we compared the accuracy of different models of science using a concentration index (i.e., Herfindahl index) based on the references from review articles ([Klavans & Boyack, 2017b](#_ENREF_22)). The premise behind this measure of accuracy is that those writing review articles are proficient in their topics, and that each review thus serves as an expert opinion of the contents of a topic in science. Assuming this premise to be true, the model that most closely duplicates the topic structure suggested by thousands of reviews is the most accurate. We note that this measure is citation-based, and thus not independent of our models.

For this measure, we identified 17,873 papers from PubMed published in 2010 that had at least 100 references each, and for which at least 80% of the references were available in PubMed and could be assigned PubMed article IDs. The reference information was obtained by matching PubMed and Scopus records, both for the review papers and for the references. The Herfindahl index, which has a natural range from 0 to 1, was then calculated for each paper and model combination. For the P review papers (*p* = *1 … P*), the Herfindahl value for each paper *p* in each taxonomy *i* is calculated as:

. (2)

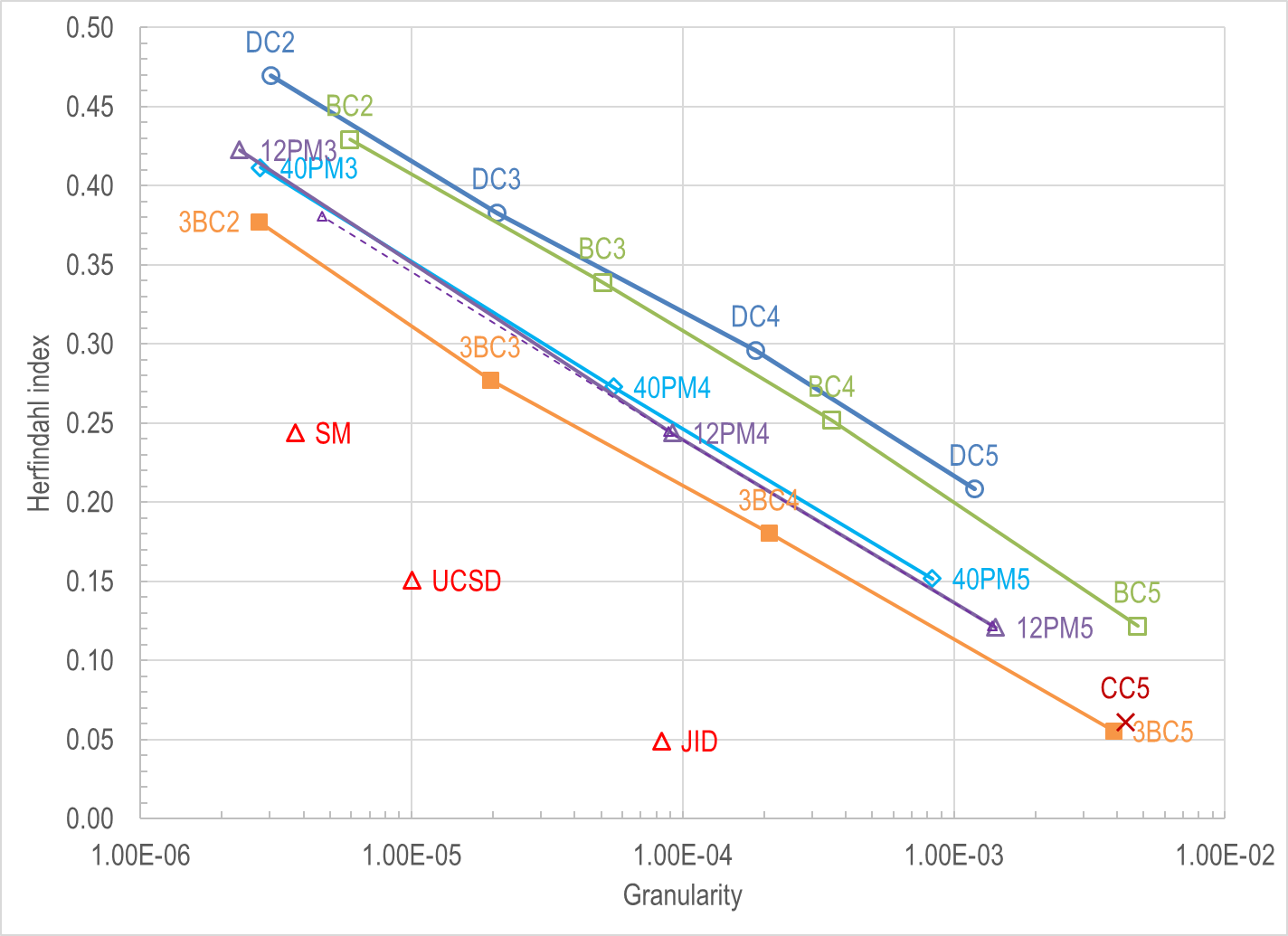
where the standard share value for paper *p* in cluster *j* of taxonomy *i* is calculated as , is the number of references from paper *p* in cluster *j* of model *i*, and  is the total number of references in paper *p* that are available in the DC5 and PubMed models. The overall Herfindahl index value for a model is the average of the index values over the 17,873 individual papers.

Figure 5 shows these values as a function of granularity for each model considered in this analysis. Although the figure shows results for models with a wide range of granularities, we are most interested in the results with a granularity of around 10-3, because these are the models where the average cluster contains roughly 400 papers, with each cluster representing a topic in science. Note that for those models where hierarchical clustering was done to obtain aggregated models with lower levels of granularity (e.g., DC, BC, 12PM, 40PM), lines are drawn connecting the models created using the same organizing principle. For instance, the 12PM5, 12PM4, and 12PM3 models are connected by a line. This gives us an estimate of accuracy for a model with a different granularity value. The slope of these lines is significant. It is roughly the same for all model types, and thus shows the natural relationship between granularity and reference concentration as calculated using the Herfindahl index. When taking the slope into consideration, this also suggests that the 12PM5 and 40PM5 models have very similar accuracy because the 12PM and 40PM lines are nearly on top of each other, even though the absolute value for 40PM5 is somewhat higher than that for 12PM5.

Figure 5 also shows that the direct citation and single-year bibliographic coupling models are the most accurate when using the Herfindahl index based on references in review articles. However, the PubMed-based models (12PM and 40PM) also perform relatively well. The 12AU models are represented by the dashed purple line and small purple triangles, and are virtually indistinguishable from the 12PM results. The difference between DC5 and 12PM5 models is perhaps not as large as the figure might suggest. The fact that the Herfindahl index is based on squared values accentuates difference between curves. Thus, the differences in accuracy between models based on Figure 5 should not be judged in a linear sense.

***2.c.2: Text-based measure***

For our second measure to compare the accuracies of different models, we use a text-based measure. The PubMed related article scores are a convenient measure to use because they are available through an Entrez e-utility and thus do not need to be calculated again. However, this measure is not independent of the models in our comparison. In fact, this is the exact same measure used to cluster the 12PM, 40PM and 12AU models. In that sense, this measure does not provide a fair comparison between text-based and citation-based models. Nevertheless, we include it for two reasons. First, the Herfindahl index (our first measure) is inherently biased to citation-based similarities, and so we feel we can include a measure that is biased toward text-based similarities as well for balance. Second, since this measure was used for clustering as well as for analysis, it provides an upper bound to accuracy based on textual characteristics ([Waltman et al., 2017](#_ENREF_30)).



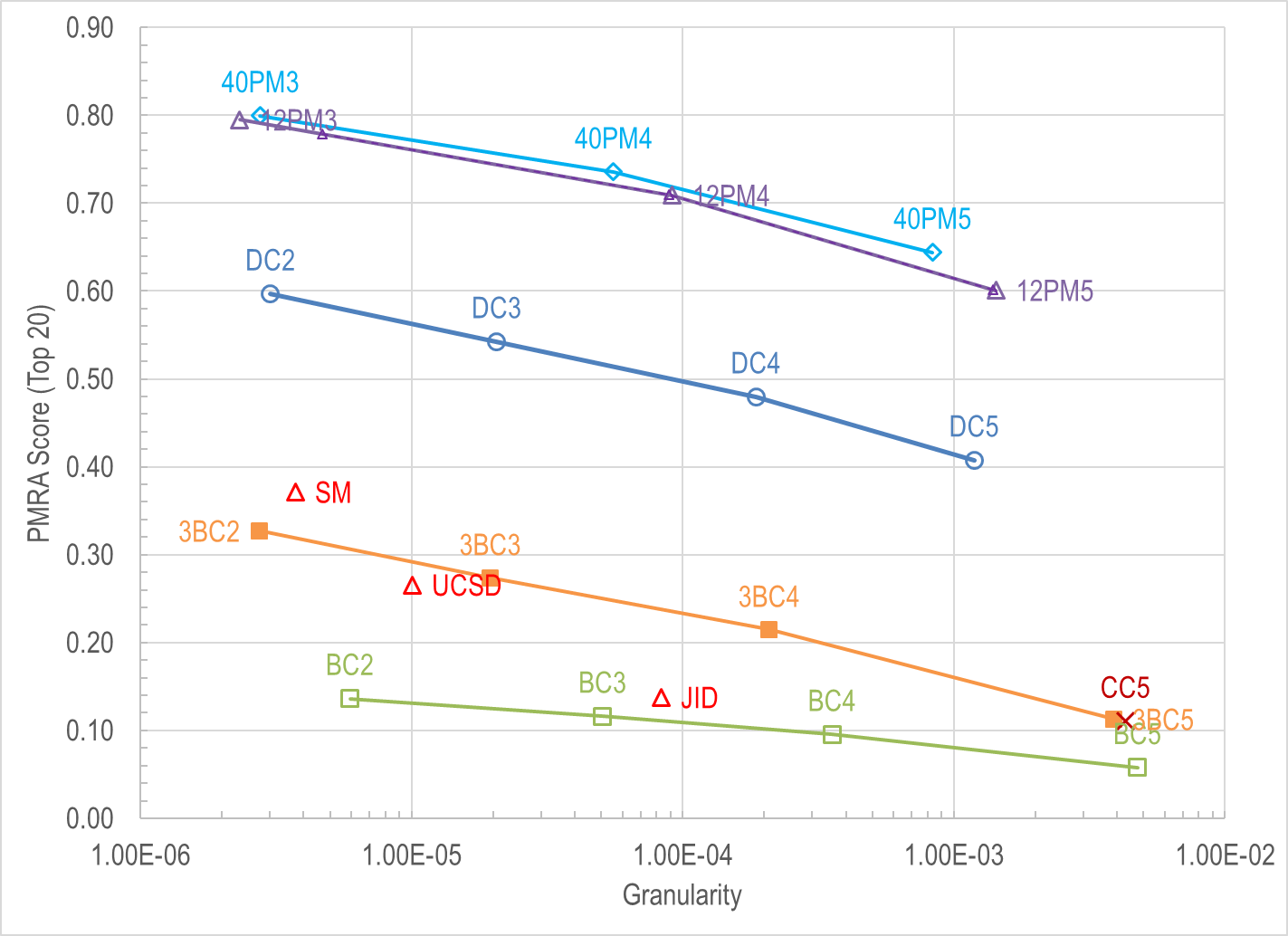
**Figure 5. Herfindahl index as a function of granularity for models of science created using different organizing principles. DC = direct citation, BC = bibliographic coupling, CC = co-citation, 12PM = PubMed related articles top 12, 40PM = PubMed related articles top 40. SM, UCSD and JID are the journal-based models.**

To calculate this metric, we take the top 20 scoring similarity pairs for each article published from 1996 to 2012. Pairs where the linked document was outside this time window were later excluded. Thus, for many papers, fewer than 20 similarity pairs were included in the basis set, which ultimately consisted of a list of 203,537,938 pairs of documents with PMRA scores. For each model, the metric value was then calculated as the sum of the scores for paper pairs that were in the same cluster divided by the total sum of the scores. This metric thus reflects the fraction of the overall textual signal that ends up within clusters of a model of science. Table 5 shows example scores and whether the pairs end up in the same cluster for three different models for a single PubMed article.

The overall results for this metric for all models are shown in Figure 6. As expected, the two sets of models based on PMRA scores do the best job of creating clusters where most of the text similarity is within clusters. The 40PM5 and 12PM5 models both preserve over 60% of the overall textual similarity within clusters, while the DC5 model only preserves 40% of the overall textual similarity within clusters. We note that, while the 40PM5 has the best overall score among models with granularity of around 10-3, it is not substantially higher than the score for the 12PM5 model. Thus, including the additional similarities (going from top 12 to top 40) does not appear to substantially increase the accuracy of the model.

**Table 5. PMRA scores and clustering results for pairs associated with PMID=25168145.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **PMID-1** | **PMID-2** | **Year-2** | **PMRA Score** | **DC5** | **40PM5** | **12PM5** |
| 25168145 | 22461425 | 2012 | 22464054 |  | YES | YES |
| 25168145 | 20075867 | 2010 | 22117189 |  | YES | YES |
| 25168145 | 17565368 | 2007 | 22112204 | YES |  | YES |
| 25168145 | 23146937 | 2012 | 21724206 |  | YES | YES |
| 25168145 | 21304548 | 2011 | 19668356 | YES | YES | YES |
| 25168145 | 21277195 | 2011 | 19572506 |  | YES | YES |
| 25168145 | 20808441 | 2010 | 19530867 |  |  |  |
| 25168145 | 11793144 | 2001 | 19519581 |  |  |  |
| 25168145 | 15843607 | 2005 | 19022356 |  |  |  |
| 25168145 | 18419827 | 2008 | 19020558 |  |  |  |
| 25168145 | 11850451 | 2002 | 18953939 | YES |  | YES |
| 25168145 | 16817857 | 2006 | 18531906 | YES |  | YES |
| 25168145 | 19683913 | 2009 | 18427798 | YES | YES | YES |
| 25168145 | 21720312 | 2011 | 18414653 | YES | YES | YES |
| 25168145 | 22072513 | 2012 | 18322214 |  |  |  |
| 25168145 | 15909178 | 2005 | 18312712 |  |  |  |



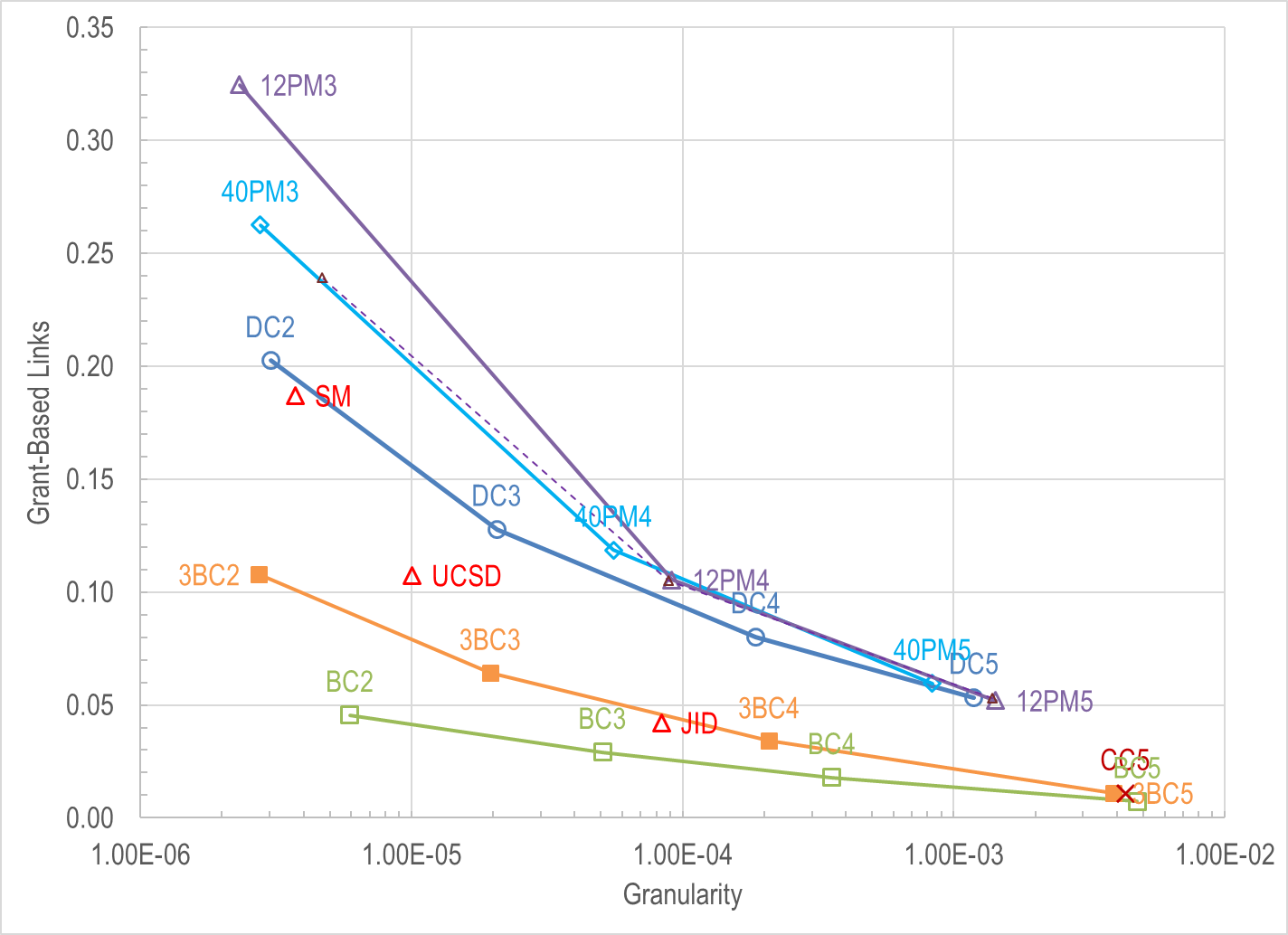
**Figure 6. Fraction of textual similarity preserved within clusters as a function of granularity for models of science created using different organizing principles.**

***2.c.3: Grant linkage-based measure***

Our third accuracy metric is based on grant-article linkages mined from the acknowledgments of papers. From these we create a list of pairs of papers that acknowledge the same grant. The fraction of those pairs of papers that end up in the same cluster for each model is then computed. This is somewhat different from, but related to a measure we used in previous work ([Boyack & Klavans, 2010](#_ENREF_5); [Boyack et al., 2011](#_ENREF_7)) where grant-article linkages were used to calculate a Herfindahl index based on cluster assignments. This grant-based measure is the most objective of our three measures in that these data were not used to create any of our models.

The grant-article linkage data were obtained from the NIH RePORTER website, and comprise the 3,686,992 links between PubMed article IDs and NIH project numbers for papers published from 1980 to 2012. From these links, we found that there are 263,334,240 pairs of papers that reference the same NIH project number. The fraction of these pairs of papers that appear in the same cluster has been calculated for each model.

The results for this metric for all models are shown in Figure 7. Among models with a granularity of around 10-3, the relative accuracies of the DC5, 12PM5 and 40PM5 models are virtually indistinguishable in that they all fall roughly along the same sloped line. For models with lower granularity (12PM3 and 40PM3), the text-based models clearly reproduce the patterns associated with NIH grant data more accurately than do citation-based models. However, this could also reflect the fact that the DC2,3 models contain all science rather than just biomedical science, and the biomedical signal may be diluted by related papers in other sciences that are not covered by PubMed at the more aggregated levels.



**Figure 7. Fraction of pairs of papers referencing the same NIH project number appearing within clusters as a function of granularity for models of science created using different organizing principles.**

***2.d: Select the best model for further refinement and to use as the basis for indictors***

With three different accuracy metrics, one balanced, one biased toward citation-based similarities, and one biased toward text-based similarities, we now have enough information to draw observations about the relative accuracies of different models. Table 6 contains results comparing the DC5, 12PM5 and 40PM5 models, and shows that overall, one can consider all three models to have comparable accuracy.

**Table 6. Comparison of accuracy metrics for models near granularity = 10-3.**

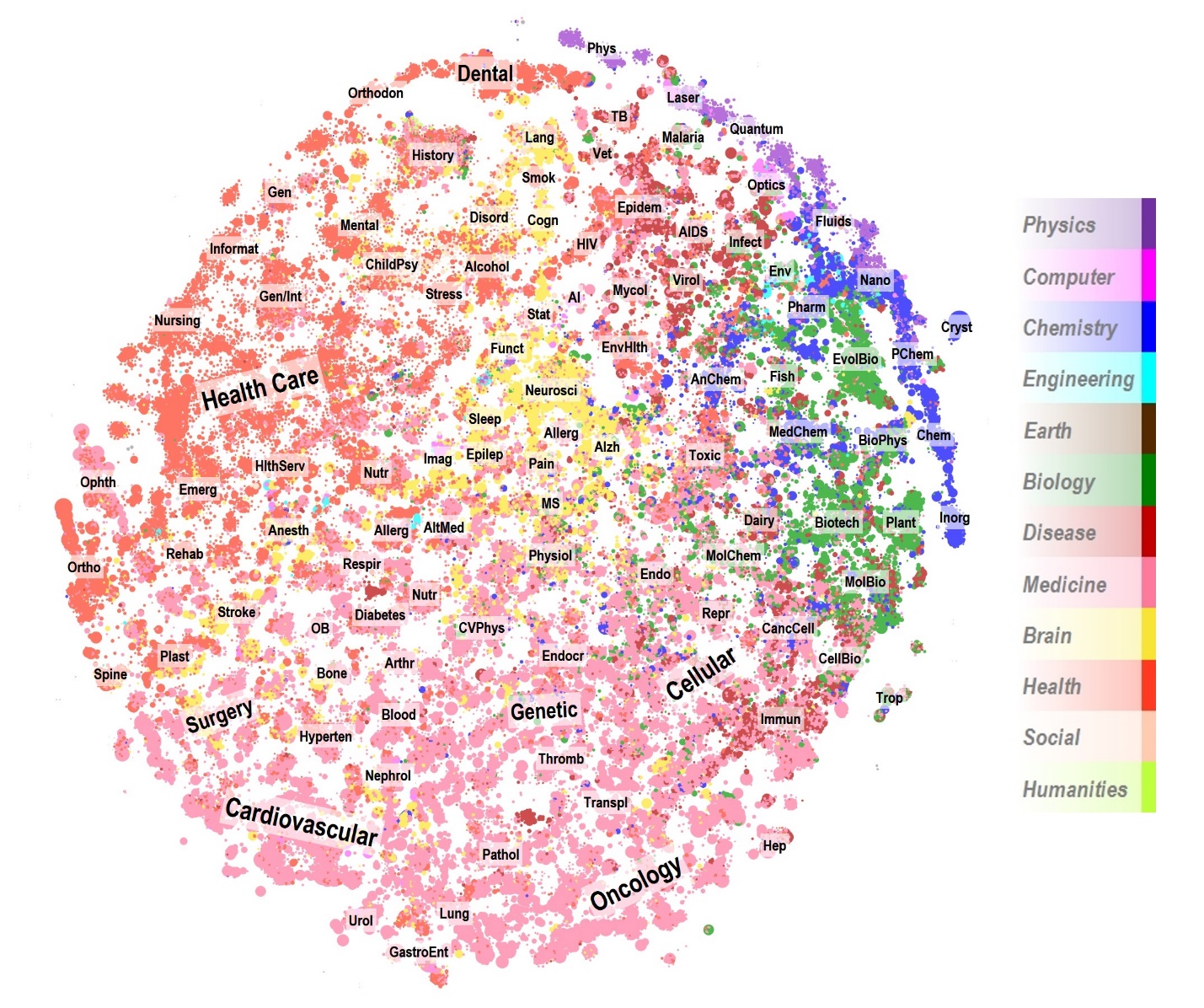
|  |  |
| --- | --- |
| **Measure** | **Result** |
| Citation-based measure | DC5 > (12PM5, 40PM5) |
| Text-based measure | (12PM5, 40PM5) > DC5 |
| Grant-based measure | (12PM5, 40PM5) = DC5 |
| Overall | (12PM5, 40PM5) ≈ DC5 |

We now need to choose one model from among the PMRA-based models to use for future development. Figures 5-7 show that the metric values for the three models (12PM5, 40PM5 and 12AU5) are very similar. In terms of absolute accuracy, the 40PM5 model scores highest. However, when slopes of the curves are considered, this absolute difference is much reduced. When granularity is considered, we feel that the 12PM5 model is the best one to move forward with because it is 1) sufficiently accurate, and 2) has higher granularity than the other models, which reflects a higher level of differentiation between clusters.

Now that we have chosen a model upon which the rest of the analysis will be based, we will no longer talk about “clusters”, but will refer to the clusters as “topics”.

To facilitate future visualizations of the 12PM5 topic space, we have created a visual map of the model. This was done by taking the top 15 topic-topic similarity values for each of the 55,202 topics, and using them as input to the DrL graph layout algorithm ([Martin, Brown, Klavans, & Boyack, 2011](#_ENREF_25)), which assigns each node an x,y position in 2D space. The resulting visualization is shown in Figure 8. Each topic is colored using journal-to-field assignments from the UCSD model of science ([Börner et al., 2012](#_ENREF_3)) based on topic contents.

As can be seen from Figure 8, although biomedical fields (Biology, Disease, Medicine, Brain, Health) dominate the map, there are also topics of documents that are based in Chemistry, Physics, Computer Science, Engineering and the Social Sciences. This map will be used as a template for overlays of metric values in later stages of the project. The map also shows the interdisciplinary nature of much of science. Even at the very low resolution of the figure, one can see local areas of the map where there are four or more colors, indicating that these are areas of science populated by researchers from multiple disciplines.



**Figure 8. Visual map of the topics in the 12PM5 model. Each dot represents a single topic, and dot sizes reflect the number of papers per topic. Labels based on manual inspection.**

**Task 3: Create and test indicators of scientific and economic impact at the topic level**

Now that the PubMed model has been created, and articles have been assigned to topics, we turn to the task of developing measures of scientific, clinical and economic impact at the topic level that can be used for decision making.

The indicators that are currently used for research assessment and funding decisions are typically of two main types – productivity indicators (such as paper counts) and indicators of scientific impact (such as citation counts or journal impact factors). The now ubiquitous h-index ([Hirsch, 2005](#_ENREF_14)) combines productivity and impact into a single indicator. Most indicators of scientific impact in current use are ultimately based on citation counts to authors, articles, institutions, etc. These cannot be calculated directly from PubMed data since PubMed does not contain reference lists, which are traditionally used to determine citation counts.

One goal of this research is to develop indicators based on data other than citations that will serve as a proxy for traditional citation-based indicators. In this section we describe and calculate many different non-citation indicators, some of which are original. We also calculate several more traditional indicators to will serve as a baseline for impact, and that will serve as the dependent variables in correlations to which the non-citation-based indicators will be compared.

The full list of indicators calculated is listed in Table 7. Each indicator, along with distributional characteristics and overlay maps, will be described in the sections below.

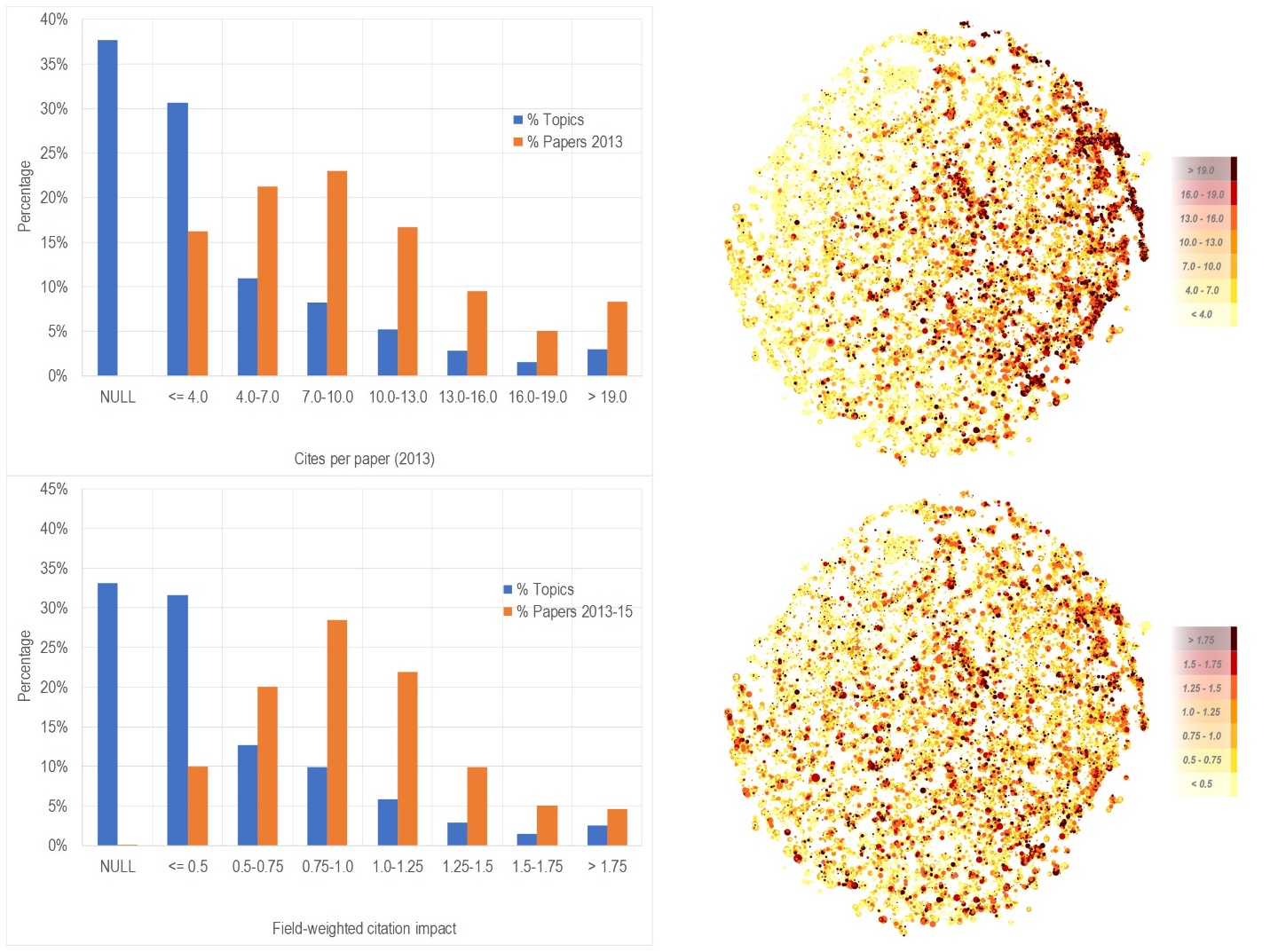
**Table 7. Indicators calculated by topic, including the years for which each indicator was calculated. The years used for the correlations and analysis in subsequent sections are shown in bold green font.**

|  |  |  |  |
| --- | --- | --- | --- |
| **INDICATOR TYPE** | **NAME** | **DESCRIPTION** | **YEARS** |
| **Dependent** |  |  |  |
| Impact | CPP\_SC | Cites per paper | 08,09,10,11,12,**13**,14 |
|  | FWCI | Field-weighted citation impact | **13-15** |
| Funding | STAR | Funding dollars using linked Star Metrics data | 08,09,10,11,12,**13**,14  **08-10,11-13,14-16** |
| **Independent** |  |  |  |
| Open citation | CPP\_PM | Cites per paper from PubMed Central Open Access subset | 08,09,10,11,12,**13**,14 |
| Size | NP | Number of records | 10,11,12,**13**,14,15,16 |
| Industrial | INDFRAC | Fraction industry participation (address) | **14-17** |
|  | PATFRAC | Fraction papers cited by a patent | **14-16** |
| Clinical | CLINFRAC | Fraction clinical participation (address) | **14-17** |
|  | RLEV | Research level | **14-17** |
| Age | VIT | Vitality – average of 1/(age+1) | 10,11,12,**13**,14,15,16 |
|  | TXTPOT | Prospective (forward) text linkage | 07,09,11,**13**,15 |
| Author | NAUTH | Authors per paper | 10,11,12,**13**,14,15,16 |
| Funding type | NFTYP | Funding types per paper (max 6) | 10,11,12,**13**,14,15,16 |
| Grant | NGR | Grant numbers per paper (limited to agencies indexed by PubMed) | 10,11,12,**13**,14,15,16 |
| Substance | NSUB | Substances indexed per paper | 10,11,12,**13**,14,15,16 |
| Atypical links | NATYP | First-year co-occurrences between indexed substances | 10,11,12,**13**,14,15,16 |
| Similarity | MAXPM | Average maximum relatedness value | 10,11,12,**13**,14,15,16 |

Although this task was originally structured in the order of creating indicators from open source data, and then correlating those indicators with citation-based and funding-based indicators, we will discuss indicators roughly in the order given in Table 7. Thus, we discuss and profile the citation and funding indicators (dependent variables) first, and will then describe indicators calculated from PubMed and related open source data (independent variables). The structure below for task 3 thus does not match that listed in the proposal.

***3.a: Description of traditional citation and funding indicators for topics***

We have calculated two traditional citation impact indicators to use as a basis of comparison. First, we calculated the average citations per paper (CPP) for each topic and for a number of years as shown in Table 7. To obtain these values, we linked Scopus citation counts for each paper as of the end of 2016 to their PubMed counterparts by matching PMIDs. This was facilitated by a table that we maintain with PubMed ID to Scopus ID matches that cover roughly 98% of all PMIDs, the bulk of which are obtained from Scopus records, and the balance of which were obtained through detailed matching of article metadata (e.g., title, author, journal, volume, year, page, etc.) The distribution of average CPP published in 2013 by topic is shown in Figure 9 along with an overlay map of the distribution. 2013 was chosen as the year for analysis since citations to most papers (especially in biomedical fields) peak within three years. Nearly 38% of the topics had no papers in 2013, and thus no cites per paper for that year. For those topics with papers, the average cites per paper was 6.35. As shown by Figure 9, the majority of papers are in topics where the average CPP are between 4 and 10. The tail on the high end is notable – nearly 5% of topics have at least 16 cites per paper. Topics with the most highly cited papers tend to be on the right half of the map.



**Figure 9. Distribution of topics and papers as a function of average cites per paper in 2013 (top) and field-weighted citation impact (bottom). Overlay maps are also shown.**

Field-weighted citation impact (FWCI) was also calculated for each topic. However, this was done not just for 2013, but for all papers from 2013 to 2015. To enable this, the expected number of citations for each paper was calculated based upon its field assignment and publication year. For field assignments, we used the 849 fields in our DC3 model created from over 40 million Scopus documents ([Klavans & Boyack, 2017b](#_ENREF_22)). Once expected counts were available, FWCI was calculated as the sum of actual counts divided by the sum of expected counts for each topic and year. These distributions are also shown in Figure 9. For topic with papers in the 2013-2015 time window, the average FWCI value is 0.647. However, when topics are weighted by the number of papers, the average value is 0.980, which is very close to the expected value of 1.0. The FWCI overlay map in Figure 9 shows that relatively high values of FWCI are reasonably well distributed throughout the map, and do not correlate visually with the CPP distribution.

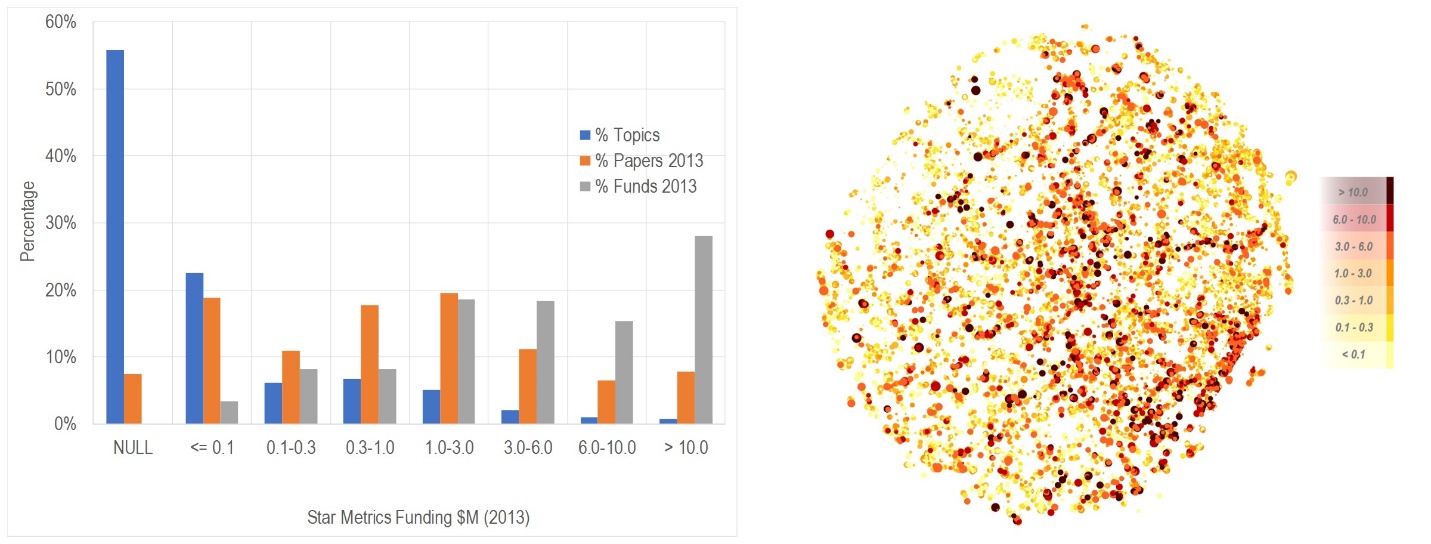
We have also calculated amounts of funded grants per topic and year using project level funding data. Star Metrics is a federal data repository that contains project level funding data by year from NSF, NIH, CDC, NASA, and several other agencies of which we used funding data for 2008-2014. Star Metrics also hosts linkage data, specifically lists of PMIDs and the grant numbers that they reference which have been extracted from PubMed records. Using links from years 2008-2016, we found linkages between 2,503,252 PMID and 110,975 NIH grant numbers. Since each PMID is found in a single topic, we assigned dollars fractionally to topics and years by grant and year using these grant-to-article linkages.

NSF funding data were also assigned to topics using the same process. However, the NSF grant-to-article linkages were not available through Star Metrics. Rather, using outputs associated with NSF annual reports that were downloaded through an NSF API, we matched those outputs (text strings, most of which were bibliographic references to scientific papers) to a listing of Scopus papers using an ElasticSearch index and search routine. Scopus IDs were then converted to PMID using our lookup tables, resulting in a set of 105,540 PMID linked to 23,044 NSF grant numbers. In total, we assigned $186.565 billion dollars (NIH: $176.450 B, NSF: $10.115 B) in funding to topics, including $23.52 billion in 2013.

Figure 10 shows the distribution of funding by topic in 2013 along with an overlay map of the distribution. 55% of the topics had no funding assigned from NIH and NSF sources. It is highly likely that many of these topics (at least those with papers in 2013) had funding from other sources such as the CDC, industry, or foreign health-related agencies. A small number of large topics are associated with a large fraction of NIH funding. Figure 10 shows that these large well-funded topics are not evenly distributed across the map.

***3.b: Description of indicators for topics calculated from open source data***

In this section we consider indicators that have the potential to be proxy for citation counts. Some of these indicators are ones that have been considered by others when predicting citation counts, while others seem to have not yet been proposed or tested. Note that here we are interested in indicators that will correlate with citation-based indicators or funding amounts at the topic level, and not at the individual paper level. The indicators described here are listed in Table 7 in the “independent variable” section.

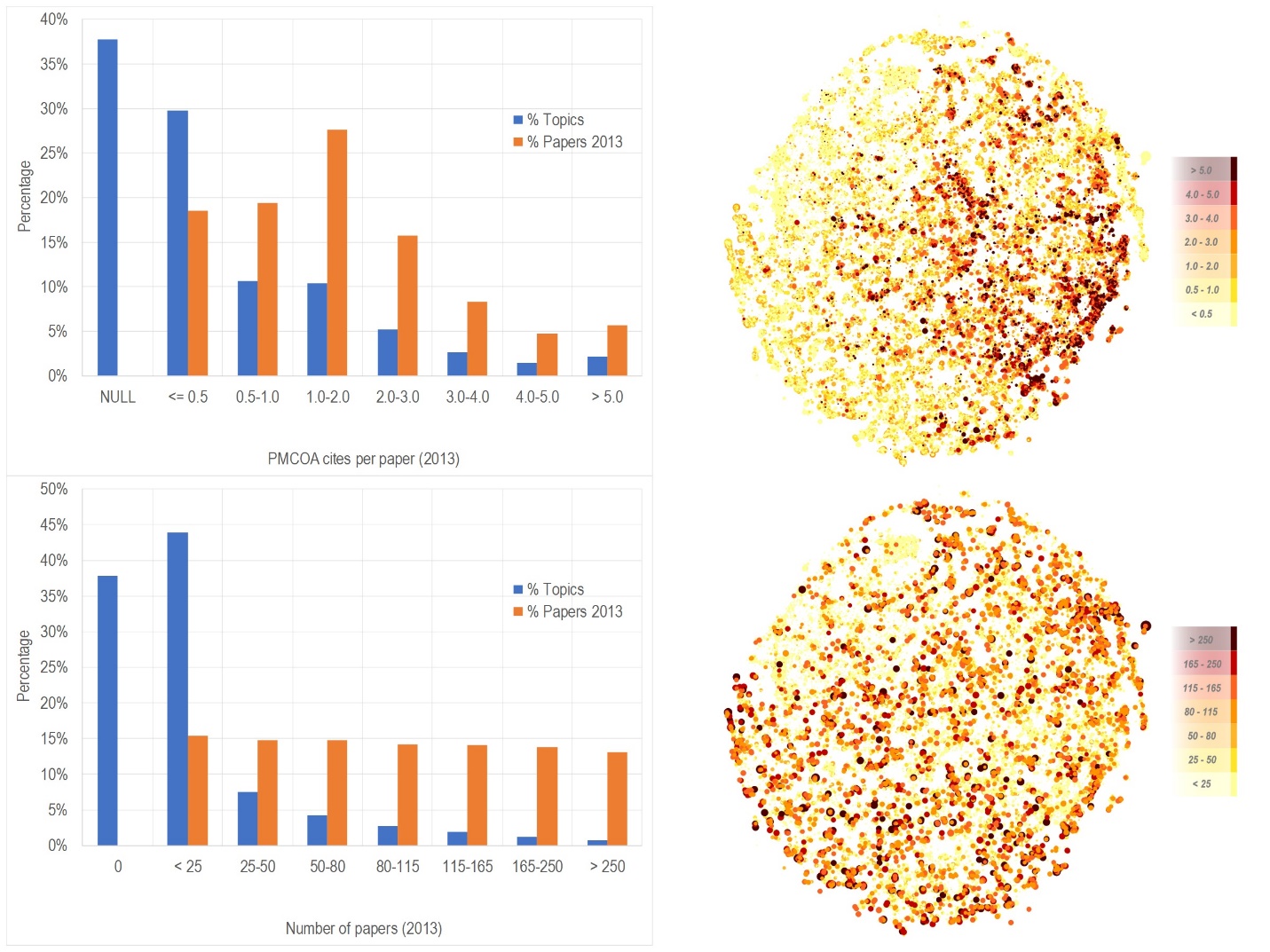


**Figure 10. Distribution of topics and papers as a function of Star Metrics funding per topic in 2013. An overlay map is also shown.**

The first indicator we consider here, perhaps unexpectedly, is based on partial citation counts. It is generally thought that citation counts are only available from the two comprehensive citation databases – Web of Science or Scopus. If one needs the most complete set of citation counts possible for paper-level metrics, it is best to use one of these sources. However, there may be some use cases where partial citation counts may be acceptable. In our case, we are interested in relative citation impact by topic. Thus, it is possible that the citation counts that can be calculated from the PubMed Central Open Access (PMCOA) subset, which comprises roughly 20% of PubMed articles over the past several years ([Boyack, Van Eck, Colavizza, & Waltman, 2018](#_ENREF_9)), may be acceptable for this purpose. Thus, we have calculated average PMCOA cites per paper (pmCPP) using the extracted references from the 1.54 million full text documents available from PMCOA through 2016.

The distribution of average pmCPP published in 2013 by topic is shown in Figure 11 along with an overlay map of the distribution. For those topics with papers, the average cites per paper was 1.22, which is roughly five times less than the full CPP from Scopus. The visual distribution of topics with relatively high pmCPP in Figure 11 is quite similar to that shown for CPP in Figure 9. However, there is a band of topics at the outer edge of the map between 1:30 and 3:00 (using a clock metaphor) with very high CPP in Figure 9 that has relatively low pmCPP in Figure 11. Comparison with the map of Figure 8 suggest that these are topics in Chemistry (blue dots in Figure 8), and thus suffer from the fact that they do not receive citations from a very distinct subset of chemistry papers rather than from a representative sample.

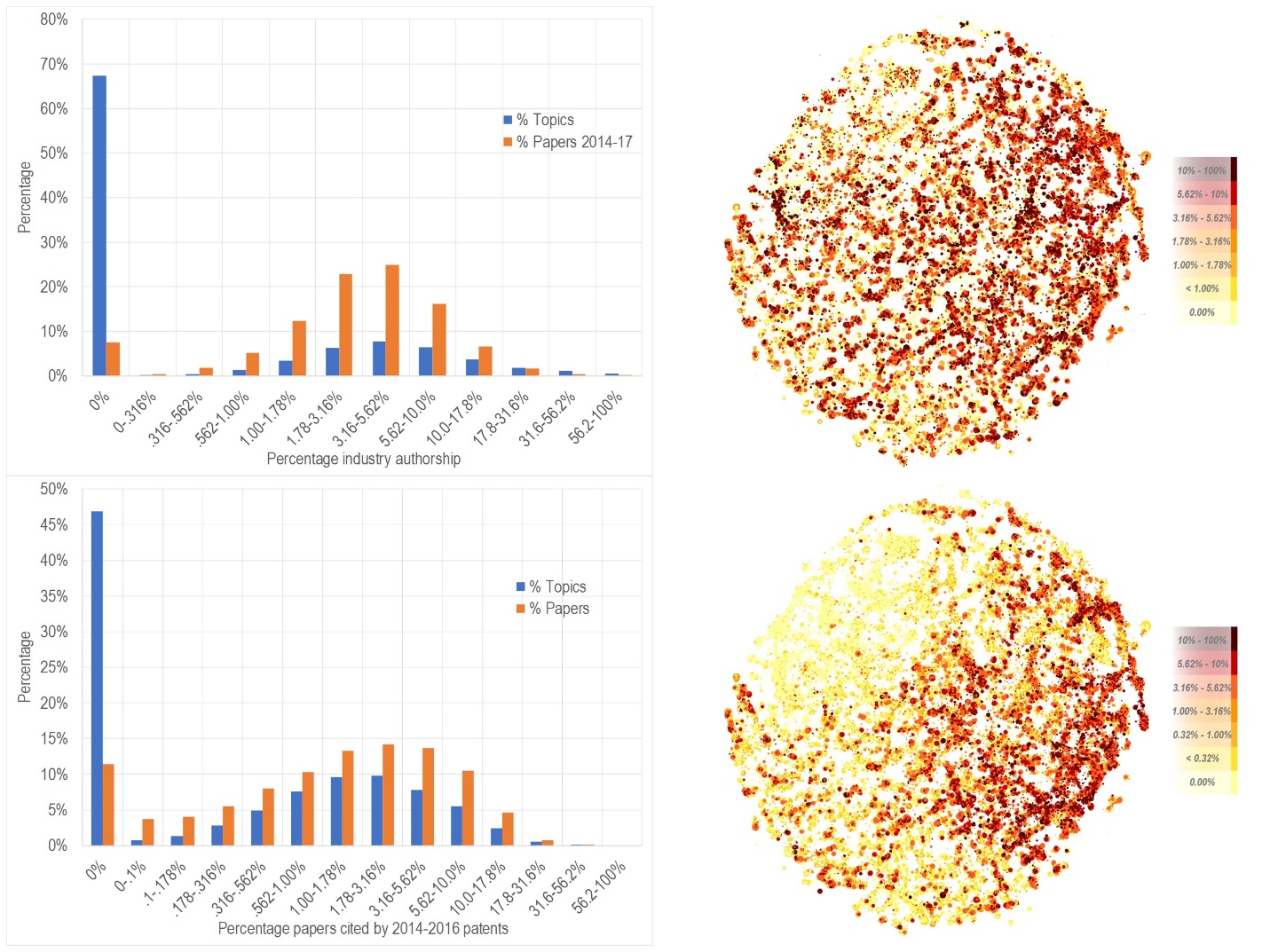
Topic size is the next indicator that we consider here. In previous research we have noted that larger topics tend to attract more citations than smaller topics. This seems to be a natural outcome of clustering papers into topics using citation links in the sense that a dense citation network tends to result in a larger topic. We suspect that this is also true for topics created using text-based linkages. Topic sizes have been calculated for each year from 2010-2016, and the distributions and overlay map for 2013 topic sizes are shown in Figure 11. In 2013, nearly 38% of topics contain no papers. Roughly 20% of the topics, those with more than 25 papers in 2013, contain 85% of the papers. Large topics occur in most areas of the map. However, there are some small areas of the map (e.g., near the upper left) that are comprised mostly of smaller topics.



**Figure 11. Distribution of topics and papers as a function of average cites per paper in 2013 from PMCOA (top) and topic size in 2013 (bottom). Overlay maps are also shown.**

The next several indicators are thought to relate to the commercialization or economic potential of a topic. Economic potential is likely related to industry involvement in a topic ([Tijssen, 2010](#_ENREF_28)). High industry authorship in a topic reflects high attractiveness of that topic to industry. Topics with high industry participation should thus have more impact economically than topics with low industry participation. We have calculated the fraction of papers per topic with an industry author.

Papers with industry authors were identified in two different ways, using common industrial abbreviations (e.g., Co., Ltd.), and using names of companies that are known to publish large numbers of papers (e.g., Merck, Pfizer). The list of industry-related abbreviations and companies used is given in Appendix A. Address strings from PubMed papers within the year range of 2014-2017 were queried to identify papers containing any of these strings. Only recent years were used for two reasons: 1) metrics need to be relatively current to be actionable, and 2) PubMed typically only indexed the first address for most articles through 2013. The common industrial abbreviations were much more effective than company names at identifying papers authored by industry. Of the 3,732,749 papers in the model from this year range, 165,769 (4.44%) contained an industry address. Only 1/3 of the topics contained any industry-authored papers; however, these accounted for only 8% of the papers. The distribution of topics and papers by percentage of industry authorship is shown in Figure 12, along with a map overlay showing where the locations of topics with high industry authorship percentages. Topics with relatively high industry authorship are found in many locations within the map, although there are some areas of the map (e.g., upper left, lower middle) with predominantly low industry authorship.



**Figure 12. Distribution of topics and papers as a function of industry authorship (top) and papers cited by patents (bottom). Overlay maps are also shown.**

Second, economic potential is likely related to the number of patents that cite papers in a given topic. We had previously linked non-patent reference strings in U.S. patents to scientific papers using an ElasticSearch index and search routine, and maintain a list of patent-to-paper references for patents issued from 1976 to 2016. From this list we identified all PubMed articles published between 2000 and 2017 that were cited by patents, and from that have calculated the percentage of papers by topic cited by patents issued from 2014 to 2016. We focus on recent patents to maintain consistency with the metric for industry authorship. However, we use a much wider time window for papers because patent lag times (from application to grant) are typically long, and very few recent papers are cited in patents. Thus, long publication windows are needed to accrue citations from patents to papers.

There are 14,160,965 papers in the model from 2000-2017. Of these, 380,401 (2.69%) are cited by patents that were granted between 2014 and 2016. As with industry authorship, most topics (60%) are not cited by any patents. However, these are typically small topics. Among those topics that are cited by patents, the largest numbers of topics have between 1.78 and 10% of their papers cited by patents, as shown in Figure 12. The figure also shows a map overlay based on patent citation percentages. Here we see that there are only select portions of the map with relatively high citation rates, and other areas of the map with low citation rates. Comparison of the industry authorship and patent citation maps shows that while there is some visual correlation, that correlation is not strong.

In addition to indicators related to economic impact, we have calculated two indicators that are related to application or clinical potential.

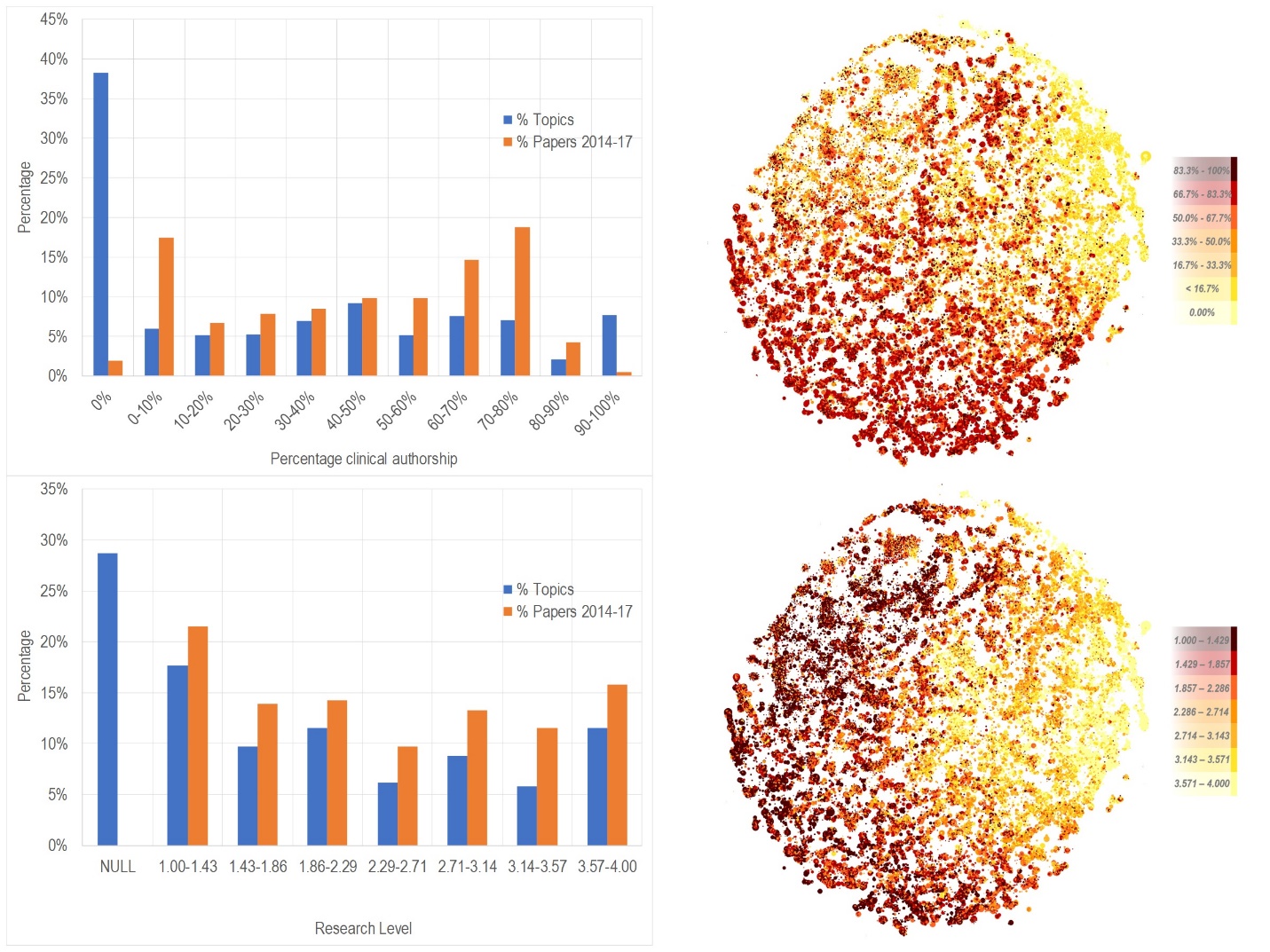
Clinical potential is likely related to hospital involvement in a topic ([Tijssen, 2010](#_ENREF_28)). A topic where a high fraction of the papers is authored by hospitals is likely to be more clinically relevant than a topic with low hospital involvement. We have calculated the fraction of papers per topic with an author from a clinical address (e.g., hospital, medical center, etc.) using a list of strings associated with clinical institutions. This list is provided in Appendix A. Address strings from PubMed papers within the year range of 2014-2017 were queried to identify papers containing any of these strings. Of the 3,732,749 papers in the model from this year range, 1,640,469 (43.95%) contained a clinical address. 38% of the topics contained papers with clinical addresses; however, these accounted for only 2% of the papers. The distribution of topics and papers by percentage of clinical authorship is shown in Figure 13, along with a map overlay showing the locations of topics with high clinical authorship percentages. Topics with relatively high clinical authorship are found in the lower half of the map, and in the top center portion. The right-hand side of the map has relatively low clinical participation, which is the same portion of the map that has the higher industry participation (compare Figure 12). Thus, it appears that there may be a negative correlation between industry and clinical participation in topics.

The second metric in this category is the average research level of the papers in each topic. Narin et al. ([1976](#_ENREF_26)) and CHI Research introduced a classification scheme representing the basic-to-applied spectrum when they classified 900 biomedical journals into four research levels (RL). Additional journals were added to the research level system at various intervals ([Carpenter et al., 1988](#_ENREF_10)). In biomedicine, research levels are defined as follows:

1. Clinical observation
2. Clinical mix
3. Clinical investigation
4. Basic research

Journals were assigned to a RL based on a combination of expert knowledge and citation patterns, using the assumption that clinical research would cite basic research but that the reverse would not be true. For example, given the research levels listed above, journals in RL1 would cite journals in RL2, RL3, and RL4, but journals in RL4 would only cite other RL4 journals.

Several years ago, we created an algorithm to calculate RL for individual papers based on words in titles and abstracts ([Boyack, Patek, Ungar, Yoon, & Klavans, 2014](#_ENREF_8)). Using a machine learning approach, words were associated with the different research levels using the original journal-to-RL assignments. The resulting lists of word-to-topic weightings have been used to estimate the RL of each PubMed article from 2014-2017. Using average research levels for each topic, the distribution of topics and papers by research level is shown in Figure 13, along with a map overlay showing areas dominated by clinical research (dark) and basic research (light).



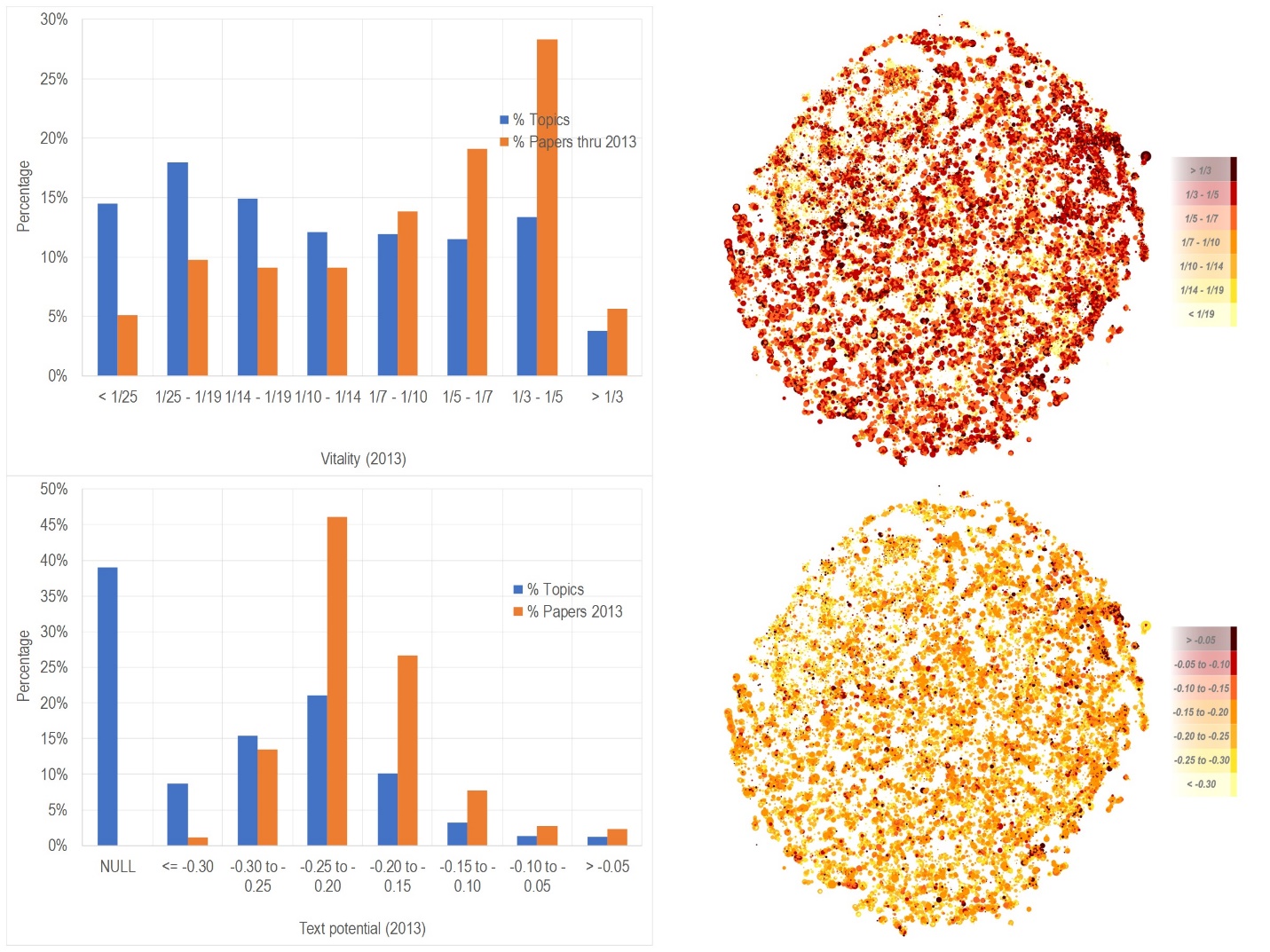
**Figure 13. Distribution of topics and papers as a function of percentage of papers with clinical addresses (top) and research level (bottom). Overlay maps are also shown.**

The two maps in Figure 13 correspond in some ways but not in others. In both maps, the region from roughly 2:00 to 4:00 (using a clock metaphor) associated with basic research (low percentage of clinical addresses, and high research levels). In both maps, the regions of the map from roughly 4:00 to 9:00 are associated with clinical research (high percentage of clinical addresses, and low research levels). However, the two maps have lower correspondence from 9:00 to 12:00 – addresses indicate a mix of basic and clinical research, while the research levels suggest that the research is nearly entirely clinical in nature.

It is also interesting to compare the research level map of Figure 13 with the patent citation map of Figure 12. The two maps are nearly opposites of each other in terms of their color distributions, suggesting that patents primarily cite research in basic research topics.

The next two indicators are associated with topic age. These inherently assume that topics follow S-curve-like trajectories over time – young, rapidly growing topics are likely to have higher short-term citation counts than older mature (and perhaps declining) topics. We calculate topic age in two different ways.

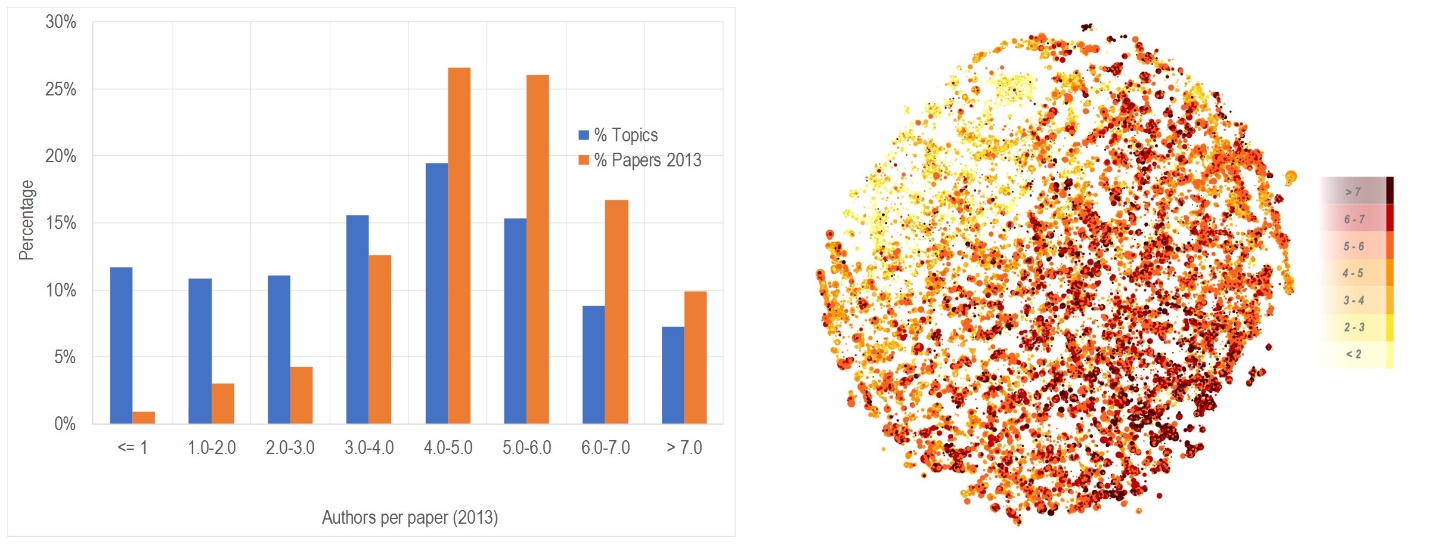
The first age-related indicator is called *vitality*, which is the average inverse age of papers in a topic at a particular point in time. This indicator is related to the technology cycle time indicator used with patents ([Kayal & Waters, 1999](#_ENREF_20)), which is the median age of cited patents. We use 1/(age+1) for each paper rather than age to reduce the effect of very old papers on the indicator value. Figure 14 shows the distribution of topic vitality as of 2013, along with an overlay map of topic vitalities. One third of the topics are quite old with vitalities of less than 1/19 (which corresponds to median ages of over 20 years old). However, these topics only account for 15% of the papers. In contrast, only 16% of topics have a vitality of greater than 1/5 (median age of less than four years), but these topics account for one third of the papers. The overlay map shows that the distribution of topic vitalities is roughly constant over the map space.



**Figure 14. Distribution of topics and papers as a function of vitality in 2013 (top) and text potential in 2013 (bottom). Overlay maps are also shown.**

The second age-related indicator is a new indicator of our design, is called *text potential* and is specific to PubMed data. This indicator is based on the hypothesis that papers whose strongest textual links are to subsequent papers will be more highly cited than those whose strongest links are to older papers. In other words, does the text of the paper project forward in time more than expected by chance or not. *Text potential* is calculated as the fraction of forward links (among the top 100 PMRA relatedness links) divided by the expected value of the fraction of forward links using a time window of 10 years forward and 10 years backward. Expected values vary by year. For instance, for 2013, there are 7,422,490 papers in the prior (backward) 10 years, and 4,114,185 papers in the forward 10 years. The distribution of average *text potential* values is shown in Table 14 along with an overlay map of the data. Interestingly, very few topics have a value greater than zero, and the average text potential value for all papers in 2013 is -0.201. Given that expected values are accounted for, we expected the average to be near zero, and have no explanation for the negative average value. Visually, the text potential map looks similar to the vitality map in that the highest values are dispersed throughout the map (although with different color scales).

The next few indicators are all simple counts of average numbers of features per paper for each topic. Figure 15 shows a distribution and overlay map of the number of authors per paper (for 2013). Number of authors has been shown to correlate positively with citation counts in a study using the entire Web of Science database ([Larivière, Gingras, Sugimoto, & Tsou, 2015](#_ENREF_23)). The average number of authors per paper in 2013 for PubMed was 5.24. Figure 16 shows that topics averaging between four and six authors per paper account for the over half of the papers in 2013. The visual overlay shows that the distribution of authors per paper is relatively constant over most of the map, but that most of the topics in the region from 9:00 to 11:00 are smaller than average.



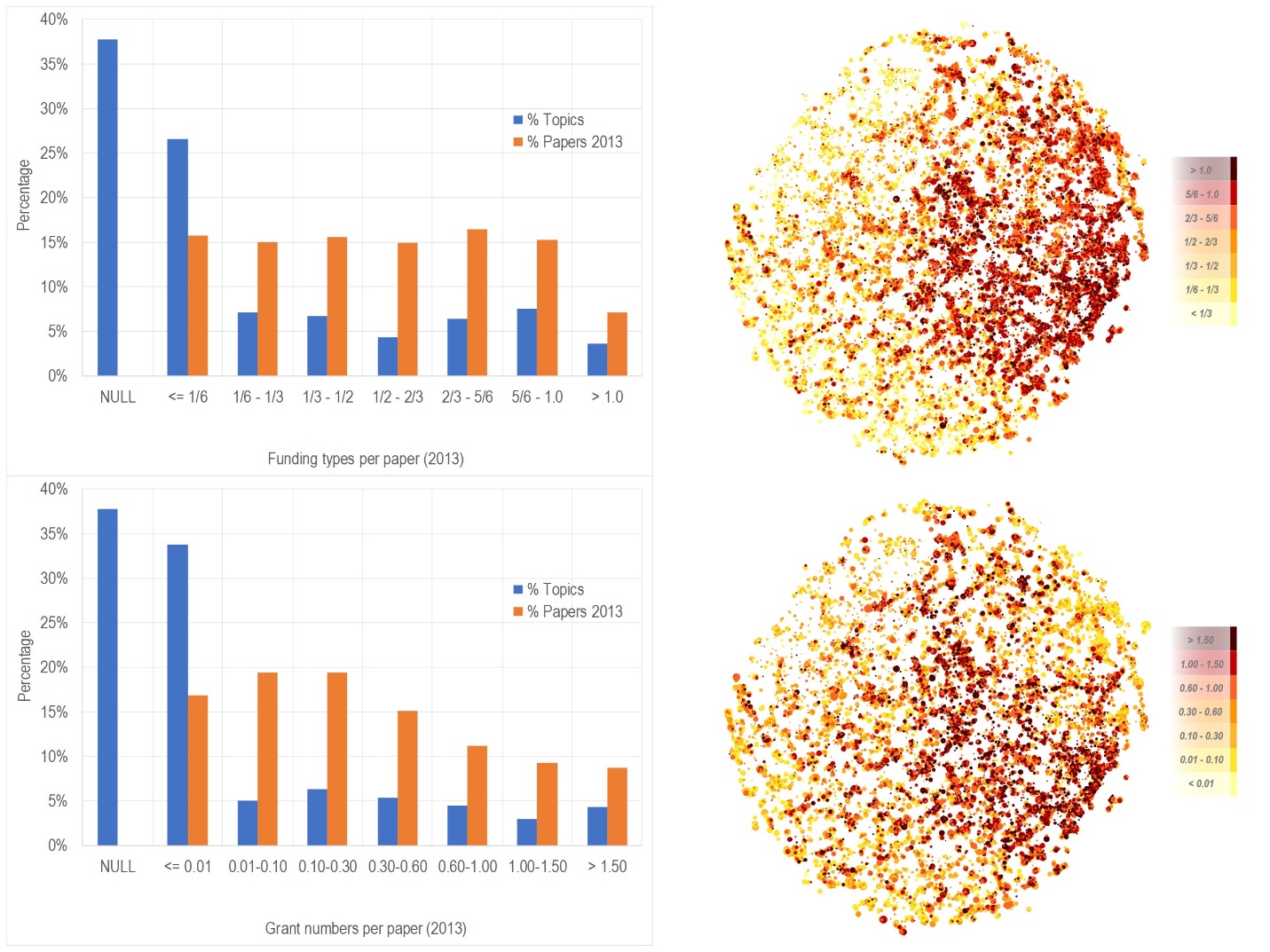
**Figure 15. Distribution of topics and papers as a function of average number of authors per paper in 2013. An overlay map is also shown.**

The next two indicators are associated with funding acknowledgments. In previous research we have shown that citation counts increase with the number of funding sources ([Boyack & Jordan, 2011](#_ENREF_4)). There are two features in PubMed data that are related to the number of funding sources – funding types and grant numbers.

Six different funding types are listed in the article type field:

* Research Support, American Recovery and Reinvestment Act
* Research Support, N.I.H., Extramural
* Research Support, N.I.H., Intramural
* Research Support, Non-U.S. Gov't
* Research Support, U.S. Gov't, Non-P.H.S.
* Research Support, U.S. Gov't, P.H.S.

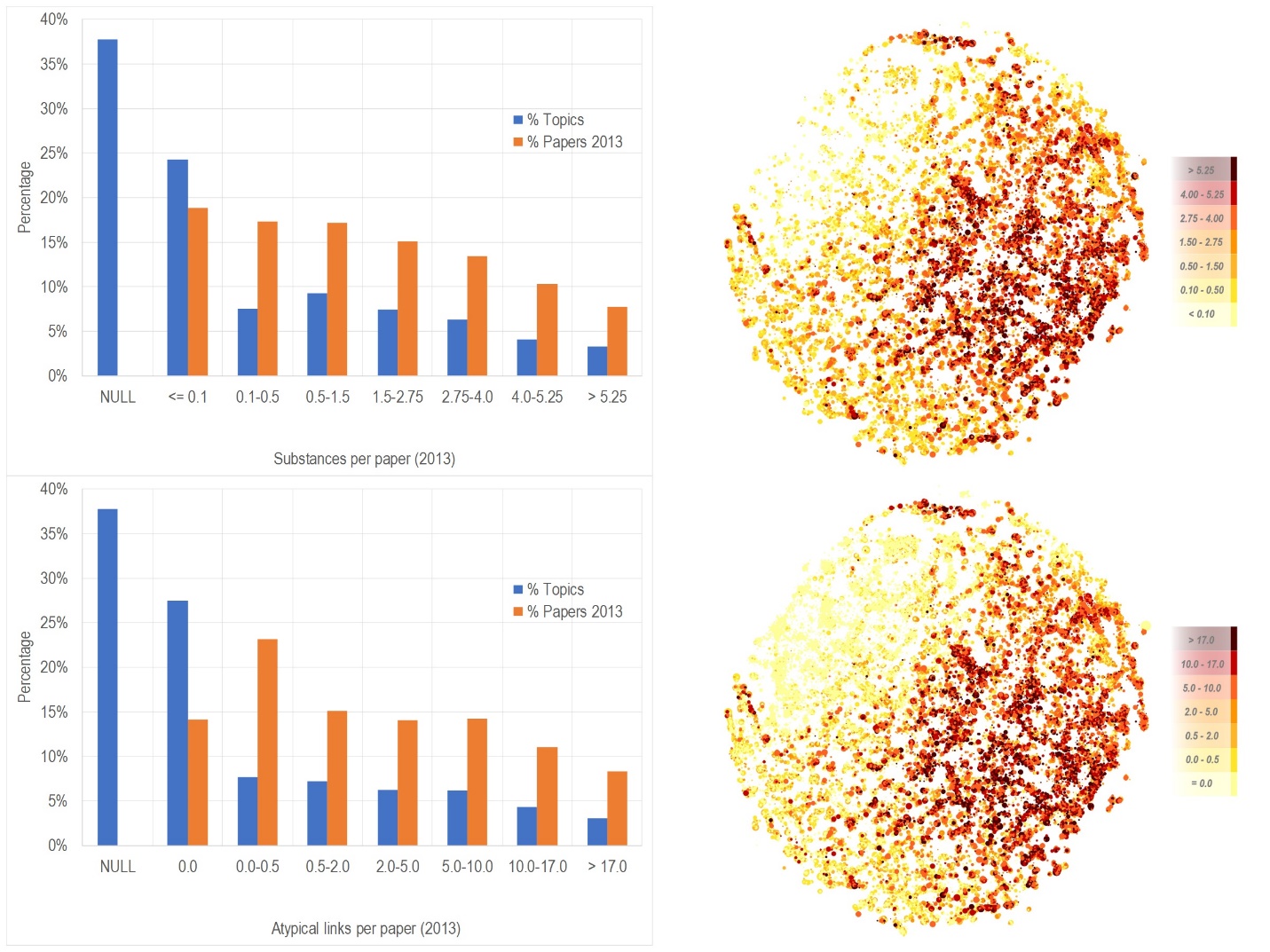
Each article in PubMed can be tagged with anywhere from none to all of these funding types. We count the average number of funding types per article for each topic and year. Figure 16 shows the distribution and overlay map of the number of funding types per paper (NFTYP) for 2013. Roughly 38% of the topics have no articles in 2013, and thus no funding types. However, for the balance of the topics, NFTYP is relatively evenly distributed over the range from 0 to 1. The distribution of funding type counts in Figure 16 is visually similar to the distributions of CPP (Figure 9), PATFRAC (Figure 12) and the high values (light colors) of RLEV (Figure 13).



**Figure 16. Distribution of topics and papers as a function of average number of funding types per paper in 2013 (top) and average number of grant numbers per paper in 2013 (bottom). Overlay maps are also shown.**

We also count the average number of unique grant numbers per paper for each topic and year. PubMed only indexes grant number for a small set of funding agencies. Nearly 90% of the grant numbers listed in PubMed are associated with NIH. Other agencies indexed include the UK Medical Research Council (2.1%), Wellcome Trust (1.7%), US PHS (1.6%), Canadian Institutes of Health Research (0.8%), and various foundations and non-US research councils at much lower frequencies. Thus, the grants numbers are very NIH centric, and not comprehensive. Nevertheless, grant number counts may distribute in a representative fashion, and if so, could be very useful for the purposes of approximating impact. The distribution of grant numbers per paper (NGRANT) for 2013 are shown in Figure 16 along with an overlay map. For most topics, NGRANT is a small number. The distribution of NGRANT is visually similar to that of pmCPP (Figure 11) in that the highest values are on the right side, but both are highly underrepresented in the Chemistry areas.

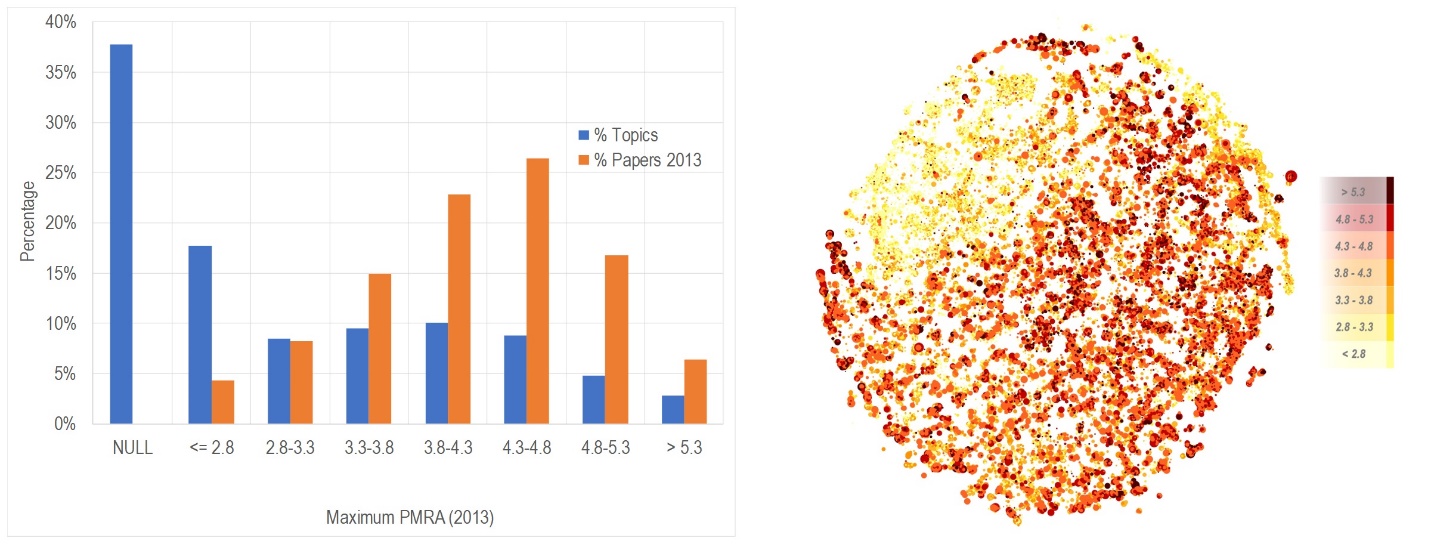
The next two indicators are associated with indexed substances. First, we count the average number of substances per paper by topic and year. We hypothesize that broader use of substances (e.g., chemicals) makes a paper more referenceable, and thus may correlate with citation counts. Figure 17 shows the distribution of the average number of substances per paper (NSUBST) for topic in 2013. Once above a threshold of 0.1, NSUBST is relatively evenly distributed over the range from 0.1 to 6.0.



**Figure 17. Distribution of topics and papers as a function of average number of substances per paper in 2013 (top) and average number of atypical substance-based links per paper in 2013 (bottom). Overlay maps are also shown.**

The second substance related indicator that we calculate is based on pairs of substances. It was recently shown that papers with unique pairs of indexed substances were more innovative than those without unique pairs ([Foster, Rzhetsky, & Evans, 2015](#_ENREF_12)). We define a pair of substances as being unique (or atypical) only in the first year that those two substances co-occur across the entire corpus, and count the average number of atypical substance pairs per paper (NATYP) by topic and year. Figure 17 shows the distribution of this indicator along with its overlay map. We note that the visual overlays of both maps in Figure 17 are very similar to that of NFTYP in Figure 16 (and thus also to CPP, PATFRAC and RLEV). Thus, these indicators may be self-reinforcing and indicative of the same phenomenon.

The final indicator that we present and calculate is the based on the PMRA relatedness values of the papers in a topic. We hypothesize that papers with high relatedness to other papers are likely to be more referenceable than papers with low relatedness values. For this indicator, we consider only the maximum PMRA relatedness for each paper (MAXPMRA), and average that by topic and year. Figure 18 shows the distribution of the indicator values. The accompanying overlay map is quite similar to the map for NAUTH (Figure 15), and also has some similarities with other maps.



**Figure 18. Distribution of topics and papers as a function of average textual relatedness in 2013. An overlay map is also shown.**

***3.c: Correlate independent indicators with dependent citation-based indicators and funding amounts by topic***

The independent variables detailed above have all been correlated with our two dependent indicators – cites per paper in 2013 and field-weighted citation impact – to identify which are the best candidates for composite surrogate indicators of impact that accurately mimic citation indicators based on full citation data. The 33,618 topics with papers in 2013 were used for the correlations. Given that the distributions for most of the variables are not normally distributed (see Figures 9-18), logarithmic and square root transforms were applied to reduce their skewness, and correlations were done with various combinations of raw and transformed variables to determine the best correlations. Log transformed variables are done using ln(value+1). Table 8 shows the subset of the full correlations with the highest values. Note that the open source citations (CPP\_PM) and number of funding types (NFTYP) have the highest correlations with the log-transformed values of both citation-based indicators, and that these correlations are quite high. We also note that the correlations between many pairs of the independent variables are also quite high, and thus are not independent of each other.

**Table 8. Correlations between open-source based indicators and citation-based indicators. Transformed variables are indicated by log (L:) and square root (S:).**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | L:cpp13\_sc | L:fwci | L:cpp13\_pm | L:np13 | rlev13 | vit13 | S:nftyp13 | S:ngr13 | S:nsub13 | S:maxpm13 |
| L:cpp13\_sc | 1 |  |  |  |  |  |  |  |  |  |
| L:fwci | 0.825 | 1 |  |  |  |  |  |  |  |  |
| L:cpp13\_pm | 0.778 | 0.703 | 1 |  |  |  |  |  |  |  |
| L:np13 | 0.555 | 0.499 | 0.451 | 1 |  |  |  |  |  |  |
| rlev13 | 0.403 | 0.243 | 0.349 | 0.068 | 1 |  |  |  |  |  |
| vit13 | 0.503 | 0.485 | 0.422 | 0.675 | 0.088 | 1 |  |  |  |  |
| S:nftyp13 | 0.685 | 0.528 | 0.595 | 0.463 | 0.509 | 0.361 | 1 |  |  |  |
| S:ngr13 | 0.496 | 0.429 | 0.558 | 0.360 | 0.272 | 0.261 | 0.602 | 1 |  |  |
| S:nsub13 | 0.488 | 0.333 | 0.441 | 0.307 | 0.535 | 0.188 | 0.588 | 0.370 | 1 |  |
| S:maxpm13 | 0.524 | 0.412 | 0.428 | 0.530 | 0.229 | 0.371 | 0.558 | 0.322 | 0.572 | 1 |

It is interesting to examine the correlations in Table 8 in the context of what we know about inter-rater reliability (IRR). IRR is the degree of agreement between raters and thus indicates the level of consensus among ratings given by judges. Although benchmark scales differ, in general an IRR value of greater than 0.7 or 0.75 is considered good or substantial, while values above 0.8 are considered very good or excellent. Thus, variables that have correlations in this range (or higher) can be considered as proxies for one another – they qualify as expert-level indicators with respect to each other. Thus, the open source version of citation counts (L:CPP\_PM) extracted from the full text of PubMed Central Open Access subset articles, which comprise perhaps 15-20% of the citations to PubMed articles, can on its own be considered a proxy variable (R=0.778) to full citation counts from Scopus (L:CPP\_SC) at the topic level. We note also that CPP and FWCI are highly related (R=0.825). We note, however, that these two variables, despite their high correlation, do measure different things. FWCI accounts for significantly different citation norms across fields, while CPP does not.

Correlations between the same independent variables and the amount of funding per topic for two different time periods were also calculated and are shown in Table 9. A different mix of variables have the strongest correlations with funding – topic size (NP) and the numbers of funding types (NFTYP) and grants (NGR).

**Table 9. Correlations between open-source based indicators and funding amounts. Transformed variables are indicated by log (L:) and square root (S:).**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | L:star1113 | L:star1416 | L:cpp13\_pm | L:np13 | rlev13 | vit13 | S:nftyp13 | S:ngr13 | S:nsub13 | S:maxpm13 |
| L:star1113 | 1 |  |  |  |  |  |  |  |  |  |
| L:star1416 | 0.920 | 1 |  |  |  |  |  |  |  |  |
| L:cpp13\_pm | 0.587 | 0.587 | 1 |  |  |  |  |  |  |  |
| L:np13 | 0.678 | 0.661 | 0.451 | 1 |  |  |  |  |  |  |
| rlev13 | 0.357 | 0.311 | 0.349 | 0.068 | 1 |  |  |  |  |  |
| vit13 | 0.473 | 0.462 | 0.422 | 0.675 | 0.088 | 1 |  |  |  |  |
| S:nftyp13 | 0.641 | 0.599 | 0.595 | 0.463 | 0.509 | 0.361 | 1 |  |  |  |
| S:ngr13 | 0.644 | 0.646 | 0.558 | 0.360 | 0.272 | 0.261 | 0.602 | 1 |  |  |
| S:nsub13 | 0.440 | 0.425 | 0.441 | 0.307 | 0.535 | 0.188 | 0.588 | 0.370 | 1 |  |
| S:maxpm13 | 0.463 | 0.441 | 0.428 | 0.530 | 0.229 | 0.371 | 0.558 | 0.322 | 0.572 | 1 |

***3.d: Calculate regression equations based on open source indicators and determine if they can be used as proxy for funding and citation-based indicators***

Given the high correlations shown in Tables 8 and 9, we then performed a series of regressions to design surrogate indicators based on open source metrics that can mimic indicators based on full citation data. Regressions were performed in a bottom-up manner. In other words, a surrogate indicator was constructed using the following steps:

1. A regression was done using the single variable with the highest correlation.
2. Correlations were then done between the remaining variables and the residual from the correlation to find the variable that had the highest correlation with the residual. This variable may or may not be the variable with the next highest correlation overall because of intercorrelation between independent variables.
3. A regression was then done using two variables – the one from the first regression, and the one resulting from (2).
4. This pattern was repeated a third or fourth time until it was found that the R2 value of the regression was not increasing significantly with the addition of another variable.

Regression models for L:CPP\_SM and L:FWCI are shown below in Figure 19. The four variables that form the models for these two indicators are the same. However, their orders are slightly different. L:CPP\_PM is the most important variable in each model. NFTYP is the second most important variable to predicting L:CPP\_SC while VIT13 (i.e., vitality, age) is the second most important variable to prediction L:FWCI. We note that in both cases, all variables are strongly inter-correlated and thus not truly independent. Also, in both cases, four variables were sufficient to maximize the R2 value (i.e., the fraction of variance accounted for by the variables).

**L:CPP13\_SC**

Source | SS df MS Number of obs = 33618

-------------+------------------------------ F( 4, 33613) =20664.82

Model | 23290.5284 4 5822.63211 Prob > F = 0.0000

Residual | 9470.9828 33613 .281765472 R-squared = 0.7109

-------------+------------------------------ **Adj R-squared = 0.7109**

Total | 32761.5112 33617 .9745519 Root MSE = .53082

------------------------------------------------------------------------------

lcpp13\_sc | Coef. Std. Err. t P>|t| [95% Conf. Interval]

-------------+----------------------------------------------------------------

lcpp13\_pm | .8622663 .0064553 133.57 0.000 .8496137 .874919

nftyp13 | .6655107 .0098351 67.67 0.000 .6462335 .6847879

vit13 | 1.427734 .0309185 46.18 0.000 1.367132 1.488335

maxpm13 | .2740474 .0102863 26.64 0.000 .2538859 .2942089

\_cons | -.0242791 .0166865 -1.46 0.146 -.0569851 .008427

------------------------------------------------------------------------------

**L:FWCI**

Source | SS df MS Number of obs = 33618

-------------+------------------------------ F( 4, 33613) =10157.88

Model | 1715.07727 4 428.769317 Prob > F = 0.0000

Residual | 1418.82191 33613 .042210511 R-squared = 0.5473

-------------+------------------------------ **Adj R-squared = 0.5472**

Total | 3133.89918 33617 .093223642 Root MSE = .20545

------------------------------------------------------------------------------

lfwci | Coef. Std. Err. t P>|t| [95% Conf. Interval]

-------------+----------------------------------------------------------------

lcpp13\_pm | .2832084 .0024985 113.35 0.000 .2783112 .2881056

vit13 | .5834309 .011967 48.75 0.000 .5599751 .6068866

nftyp13 | .0812723 .0038067 21.35 0.000 .073811 .0887335

maxpm13 | .0275558 .0039813 6.92 0.000 .0197523 .0353593

\_cons | .1000266 .0064585 15.49 0.000 .0873677 .1126855

------------------------------------------------------------------------------

**L:STAR1416**

Source | SS df MS Number of obs = 33618

-------------+------------------------------ F( 3, 33614) = 9117.10

Model | 35784.0911 3 11928.0304 Prob > F = 0.0000

Residual | 43977.6702 33614 1.3083141 R-squared = 0.4486

-------------+------------------------------ **Adj R-squared = 0.4486**

Total | 79761.7614 33617 2.37266149 Root MSE = 1.1438

------------------------------------------------------------------------------

lstar1416 | Coef. Std. Err. t P>|t| [95% Conf. Interval]

-------------+----------------------------------------------------------------

lcpp13\_pm | .9014714 .0134573 66.99 0.000 .8750946 .9278483

nftyp13 | 1.277913 .0211536 60.41 0.000 1.236451 1.319375

maxpm13 | .4767359 .0217657 21.90 0.000 .4340744 .5193974

\_cons | -3.537209 .0359558 -98.38 0.000 -3.607683 -3.466734

------------------------------------------------------------------------------

**Figure 19. Regression models for surrogate indicators to mimic traditional impact indicators L:CPP13\_SC and L:FWCI, and funding L:STAR1416.**

We note that the surrogate (regression equation) for L:CPP13\_SC does an extremely good job of representing actual cites per paper at the topic level, with an adjusted R2 of 0.711, which is roughly equivalent to a correlation of 0.843. The surrogate for L:FWCI not as good at R2=0.547, but given that this is equivalent to a correlation of 0.740, this is still quite good. Both qualify as expert-level indicators in the context of our discussion of inter-rater reliability above. These data show that we have accomplished our technical goal of creating acceptable surrogate indicators for traditional impact indicators.

The surrogate indicators for impact can be approximated by the following equations, which are based on the coefficients in Figure 19. For all topics where np13>0:

L:cpp13\_sm = 0.8622663\*lcpp13\_pm + 0.6655107\*nftyp13 + 1.427734\*vit13 + 0.2740474\*maxpm13 - 0.0242791

L:fwci = 0.2832084\*lcpp13\_pm + 0.5834309\*vit13 + 0.0812723\*nftyp13 + 0.0275558\*maxpm13 + 0.1000266

We note that these equations are specific to 2013, but given the stability in year-to-year correlations we have done using similar types of calculations ([Klavans & Boyack, 2017a](#_ENREF_21)), we suspect that the coefficients above will be representative for other years as well.

A regression models for L:STAR1416 is also shown in Figure 19. Only three variables are needed here to maximize the variance accounted for in the approximating funding amounts per topic. These are three of the same variables that are used in the impact surrogates, and occur in the same order as for L:CPP13\_SM. The surrogate for L:STAR1416 does a reasonable job of predicting funding by topic in subsequent years with an adjusted R2 of 0.449, which is equivalent to a correlation of 0.670. While not as good as our impact indicators, it is still a very reasonable and useful indicator to use to predict funding by topic. The equation for this indicator is:

L:star1416 = 0.9014714\*lcpp13\_pm + 1.277913\*nftyp13 + 0.4767359\*maxpm13 - 3.537209

Note that these equations all generate values for the logarithms of the dependent variables. To convert to actual values, an exponential transform is required, for instance:

Star1416 = exp(L:star1416)

To summarize this section, using open source data from PubMed and PubMed Central (PMC), we have developed surrogate indicators that approximate two of the most popular citation-based indicators in use today (citations per paper and field weighted citation impact) at the topic level. In addition, we have accurately assigned project-level funding data from the U.S. Star Metrics database to the topics in our model, and have developed a surrogate indicator that can serve as a topic-level funding predictor. This new indicator is just as accurate as a similar indicator we recently developed based on citation and usage data from Scopus ([Klavans & Boyack, 2017a](#_ENREF_21)).

**Task 4: Create and enhance indicators of the virtue of scientific research, including indicators of quality, transparency, reproducibility, and potential for translation**

We used a sampling process to generate a new random sample of 100 papers published in 2015 and 2016 and indexed in PubMed. The sample was stratified according to whether there is publicly available data for the full-text of the article in PMC or not.

A set of 500 biomedical papers published in 2000-2014 was previously analyzed ([Iqbal et al., 2016](#_ENREF_19)). The full text of these papers was manually analyzed, and a number of features related to quality, transparency, reproducibility and potential for translation were identified and recorded. However, the same analysis was not done using the titles and abstracts of those papers, thus we examined this aspect as well so as to do a similar analysis on the original set of 500 papers using titles and abstracts only. This enabled comparison of the level of signal that can be obtained using full text with that from metadata.

We followed an approach similar to the one that we used recently to evaluate randomly a sample of 500 papers from PubMed published in 2000-2014 ([Iqbal et al., 2016](#_ENREF_19)). The additional 100 papers were identified for 2015-2016 and verified manually for eligibility and then the relevant data were extracted.

The sample was divided into seven study categories, as in our previous evaluation: (1) no research (items with no data such as editorials, commentaries, news, comments and non-systematic expert reviews), (2) models/modeling or software or script or methods without empirical data (other than simulations), (3) case report or series (humans only, with or without review of the literature), (4) randomized clinical trials (humans only), (5) systematic reviews and/or meta-analyses (humans only), (6) cost effectiveness or decision analysis (humans only), and (7) other (empirical data that includes uncontrolled study [human], controlled non-randomized study [human], or basic science studies).

We used the enhanced field-classification process that allows a better categorization and exploration of both the 500 previous articles and the 100 new ones.

Publications with data and analyses were assessed for publicly available full protocols and datasets, patterns of reproducibility (whether the study claimed to be a replication effort, whether subsequent citing papers had tried to replicate the analyses, and whether data were included in systematic reviews and/or meta-analyses), conflicts of interest, and funding. For published items without data and analyses, only statements of conflict and funding were investigated, since protocols, datasets, and reproducibility are not relevant.

In more detail, we captured the following:

* Protocol availability: We reviewed the eligible papers for any mention of the protocol, possible hyperlink, or reference to the source for available protocol. For the studies that have publicly available protocols, we captured whether or not the available protocols cover all or part of the presented analyses.
* Dataset availability: Articles were scrutinized for any mention of access to the datasets that stand behind the analyses presented in the paper. If studies had datasets, we recorded whether the available datasets cover all or part of the presented analyses.
* Funding: For each eligible paper, we assessed whether any mention of funding was made, and if so, whether funding was from public and/or private sources.
* Replication: Abstracts from papers that include data and analyses were examined for statements regarding study novelty or replication. We used the following categories, all based on using text from the abstract and/or introduction of a paper:
  + The paper claims that it presents some novel findings,
  + The paper clearly claims that it is a replication effort trying to validate previous knowledge or it is inferred that the index paper is a replication trying to validate previous knowledge,
  + The paper claims to be both novel and replicate previous findings,
  + The paper makes no statement or unclear statement about whether it presents a novel finding or replication (or no distinct abstract and introduction exists).

In addition, the number of citations to each of the index papers with data and analyses as of end-2016 was identified. The citing papers of each index paper were examined to identify whether any of them are systematic reviews and/or meta-analyses and/or studies that claim to try to replicate findings from the index paper. Citing papers were screened at the title level, and those that were potentially relevant were also screened at the abstract, introduction, and full-text level.

**Description of Assessed Sample of Articles, 2015-2016**

Among the 100 randomly selected articles pushed in 2015 and 2016, nearly half (46 [46.0%]) were publications in the research field of “Medicine” (**Table 10**). The vast majority had some form of empirical data (n=80 [80.0%] – n = 71 excluding case studies and case series, in which protocol and raw data sharing may not be pertinent, and n = 67 excluding also systematic reviews, meta-analyses and cost-effectiveness analyses where replication in studies with different data would not be pertinent). The sample included one (1.0%) cost-effectiveness or decision analysis, 9 (9.0%) case studies or case series, 3 (3.0%) randomized clinical trials, 3 (3.1%) systematic reviews or meta-analyses, and 63 (63.0%) ‘other’ articles with empirical data (including cross-sectional, case-control, cohort, and various other uncontrolled human or animal studies). Nearly one-fifth (18 or 100 [18.0%]) of the sample was classified as research without empirical data or models/modeling studies. There were 43 (43.9%) with a PMCID, of which 21 were also PMCOA.

|  |  |
| --- | --- |
| **Table 10. Characteristics of 100 New Assessed Articles** | |
| **Characteristic** | **Median (IQR)** |
|  | **N (%)** |
| **Impact Factor (2013)** |  |
| *Impact Factor – Median (Interquartile Range)* | 3.19 (1.88-5.10) |
| *No 2013 JCT Impact Factor* | 40 |
| **Article Study Field** |  |
| Medicine | 46 (46.0) |
| Health Sciences | 18 (18.0) |
| Biology | 7 (7.0) |
| Infectious Disease | 13 (13.0) |
| Brian Sciences | 18 (18.0) |

**Protocol and Raw Data Availability**

Excluding case studies or case series and models/modeling studies (in which a protocol would not be relevant), none of the 80 articles with empirical data included a link to a full study protocol. However, two articles referenced their clinical trial identifiers (PMIDs: 27391533 and 25682436). There were three articles that provided additional methods sections, figures, or supplementary materials either as a detailed appendix at the end of the paper or online (27791002, 26572063, and 27348411).

There were 12 (12 of 80, 15.0%) articles that discussed publicly available data. Among the 20 eligible PMCOA articles, one stated that data had been deposited on the Gene Expression Omnibus database repository and include a dataset identifier (26484203), one included an OpenfMRI number (27096608), one stated that TCR sequence data had been deposited in the NCBI Sequence Read Archive [a valid reference number was missing from the article (27348411)], and one was actually a data brief outlining the raw data for an existing study (27617276). Although another article mentioned that all relevant data were within the supporting information files, the supplementary files did not contain any raw data (26413900). Among the 60 non-PMCOA articles, one article included supplementary figures and primers used for qPCR analyses and excel documents with RNA-sequence transcriptome analyses (27214551), one disclosed that data was reported in the Gene Expression Omnibus (27791002), one reported data on ProteomeXchange (26238763), one stated that genome sequence data were submitted to Genbank (27108998), one stated that coordinates and structure factors were deposited in the Protein Data bank (27768894), one said that fMRI data from the Human Connectome Project was utilized (25252277), and another said that ITGB2 sequence variants were submitted to GenBank (26639818).

**Funding**

Nearly one-third (30 of 100 [30.0%]) of the 100 biomedical articles did not include any information on funding. Half (50 of 100 [50.0%]) were publically funded, either alone or in combination with other funding sources. Of these, 21 had National Institutes of Health (NIH) funding and two received National Science Foundation (NSF) support, either alone or in combination with other sources of funding. There were no funding differences between PMCOA and non-PMCOA articles (**Table 11**).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Table 11. Articles in the PMCOA Category Versus Non-PMCOA** | | | | | |
| **Variable** | **PMCOA** | | **Non-PMCOA** | | ***P* value** |
|  | N | % | N | % |  |
| **Funding** | **N = 21** | | **N = 79** | |  |
| No mention | 6 | 28.6 | 24 | 30.4 | 0.14 |
| No Funding | 3 | 14.3 | 7 | 8.9 |
| Public | 3 | 14.3 | 30 | 38.0 |
| Private | 0 | 0.0 | 2 | 2.5 |
| Other | 2 | 9.5 | 5 | 6.3 |
| Some combination of Public, Private, or Other | 7 | 33.3 | 11 | 13.9 |
| **Replication** | **N = 15** | | **N = 52** | |  |
| Novel Findings | 7 | 46.7 | 24 | 46.2 | 0.80 |
| Replication | 0 | 0.0 | 4 | 7.7 |
| Novel Findings and Replication | 2 | 13.3 | 9 | 17.3 |
| No State on Novelty or Replication | 6 | 40.0 | 15 | 28.9 |
| **Article Citation** | **N = 15** | | **N = 52** | |  |
| ***Replication of Index Study*** |  | |  | |  |
| No Citing Article | 15 | 100.0 | 50 | 96.2 | 1.0 |
| At Least One Citing Article | 0 | 0.0 | 2 | 3.8 |
| *Systematic Review/Meta-Analysis* | | | | |  |
| No Citing Article | 15 | 100.0 | 51 | 98.1 | 1.0 |
| At Least One Citing Article, No Data Included | 0 | 0.0 | 1 | 1.9 |
| At Least One Citing Article, Data Excluded | 0 | 0.0 | 0 | 0.0 |
| At Least One Citing Article, Data Included | 0 | 0.0 | 0 | 0.0 |
| **Statement of Conflict** | **N = 21** | | **N = 79** | |  |
| No Statement | 4 | 19.0 | 34 | 43.0 | 0.10 |
| Statement, No Conflict Exists | 15 | 61.9 | 38 | 48.1 |
| Statement, Conflict Exists | 2 | 9.5 | 7 | 8.9 |
| Based on Fisher’s Exact Test | | | | | |

**Articles Claiming to Contain Novel Findings Versus Replication Efforts**

Among the 67 biomedical articles with empirical data, excluding case studies and case series, systematic reviews/meta-analyses, and cost effectiveness/decision analyses studies, only four clearly claimed or were inferred to be replication efforts trying to validate previous knowledge (**Table 11)**. Nearly half (31 of 67, 46.3%) claimed to present some novel findings and 11(16.4%) had statements of both study novelty and some form of replication. Approximately one-third (21 of 67, 31.3%) either had no statement or an unclear statement in the abstract and/or introduction about whether the article presented novel findings or replication efforts. There were no observable differences between PMCOA and non-PMCOA articles (**Table 11**).

**Subsequent Citing by Replication Studies**

Of the 67 biomedical articles with empirical data, there were two non-PMCOA articles that had at least some portion of their findings replicated (**Table 11**). One of the replicating articles used an “almost comparable study design but over a long period” and included some patients with different characteristics. The second was a partial replication effort with a longer follow-up. Only one article was included in a subsequent systematic review.

**Reporting of Conflicts of Interest**

Among the 100 articles, there were 38 (38.0%) that had no conflict of interest statement. There were 52 (52.0%) that reported no conflicts of interests and 9 (9.0%) that had a clear statement of conflict. There were no observable differences between PMCOA and non-PMCOA articles (**Table 11**).

**Open Source Data (PubMed Level)**

**Protocol and Raw Data Availability**

Among the 541 articles published between 2000-2016 in eligible research fields directly related to biomedicine, 488 were non-PMCOA articles. Of these, there were 297 (60.9%) where protocol or data sharing would be relevant (excluding model/modeling studies and case studies or case series). Although none of the articles included a link to a full study protocol, there were seven articles that either referenced their clinical trials identifier, included a link to ClinicalTrials.gov, or stated that a Clinical Trials repository link was available on the journal website. Nearly all (6 of 7, 85.7%) also had clinical trial identifiers at the PubMed level. Among the 17 articles that discussed supplementary data and/or figures, database identifiers, or claimed that data was available upon request, only two included this information at the at the PubMed level, under the “Publication type, MeSH terms, Secondary source ID” tab.

**Funding and Reporting of Conflicts of Interest**

Among the 488 non-PMCOA articles published between 2000-2016, 245 listed funding sources at the full text level. Nearly half (122 of 245, 49.8%) of these articles included some funding information under “Publication type, MeSH terms, Secondary source ID” tab on PubMed (e.g., Research Support, Non-US Govt). There were 88 additional articles where PubMed provided at least one specific funding source (i.e. a specific grant number). No information regarding conflicts of interest were provided at the PubMed level.

**Articles Claiming to Contain Novel Findings Versus Replication Efforts**

There were 287 biomedical articles with empirical data, excluding case studies and case series, systematic reviews/meta-analyses, and cost effectiveness/decision analysis studies published between 2000-2016. Of these, five articles did not have an abstract on PubMed. Among the nine articles classified as replication studies based on information provided in the abstract and/or introduction, five had enough information in the abstract alone to establish whether they were replication studies. Less than half (68 of 143; 47.6%) of the articles claiming to present some novel findings based on the abstract and/or introduction could be classified as novel accord to the abstract only. Among the 14 articles that had statements of both study novelty and some form of replication, only six could be classified based on the abstract only.

**Comparison of Results from 2000-2014 and 2015-2016**

A comparison of articles published from 2000-2014 versus 2015-2016 revealed some distinctive patterns (**Table 12**). Between 2000-2014, articles were more likely to include any information related to funding. Furthermore, there were more replication attempts published in recent years (either alone or combined with addition novel analyses) and articles were more likely to contain statements of conflict.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Table 3. Comparison of 2000-2014 and 2015-2016 Samples** | | | | | |
| **Variable** | **2000-2014 Sample** | | **2015-2016 Sample** | | ***P* value** |
|  | N | % | N | % |  |
| **Funding** | **N = 441** | | **N = 100** | |  |
| No mention | 226 | 51.3 | 30 | 30.0 | <0.001 |
| No Funding | 12 | 2.7 | 10 | 10.0 |
| Public | 87 | 19.7 | 33 | 33.0 |
| Private | 19 | 4.3 | 2 | 2.0 |
| Other | 29 | 6.6 | 7 | 7.0 |
| Some combination of Public, Private, or Other | 68 | 15.4 | 18 | 18.0 |
| **Replication** | **N = 259** | | **N = 67** | |  |
| Novel Findings | 133 | 51.4 | 31 | 42.3 | <0.001 |
| Replication | 5 | 1.9 | 4 | 6.0 |
| Novel Findings and Replication | 5 | 1.9 | 11 | 16.4 |
| No Statement on Novelty or Replication | 111 | 42.9 | 21 | 31.3 |
| No abstract | 5 | 1.9 | 0 | 0.0 |
| **Article Citation** |  | |  | |  |
| ***Replication of Index Study*** |  | |  | |  |
| No Citing Article | 251 | 96.9 | 65 | 97.0 | 1.0 |
| At Least One Citing Article | 8 | 3.1 | 2 | 3.0 |
| *Systematic Review/Meta-Analysis* | | | | |  |
| No Citing Article | 221 | 85.3 | 66 | 98.5 | 0.023 |
| At Least One Citing Article, No Data Included | 19 | 7.3 | 1 | 1.5 |
| At Least One Citing Article, Data Excluded | 3 | 1.2 | 0 | 0.0 |
| At Least One Citing Article, Data Included | 16 | 6.2 | 0 | 0.0 |
| **Statement of Conflict** | **N = 441** | | **N = 100** | |  |
| No Statement | 305 | 69.2 | 38 | 38.0 | <0.001 |
| Statement, No Conflict Exists | 110 | 24.9 | 53 | 53.0 |
| Statement, Conflict Exists | 26 | 5.9 | 9 | 9.0 |
| **PMCID** |  |  |  |  |  |
| Yes | 85 | 19.3 | 57 | 57.0 | <0.001 |
| No | 356 | 80.7 | 43 | 43.0 |
| Based on Fisher’s Exact Test | | | | | |

***4.d: Determine correlation between value indicators and impact measures***

The number of Scopus citations of the 541 analyzed papers was correlated with the number of citations obtained in the PMCOA set, as shown in Figure 20 below.

**Figure 20. Correlation between number of citations derived from the PMCOA set until the end of 2016 and the total number of citations in Scopus**

The median number of Scopus citations was 11 when there was no statement of conflict of interest, 9 when there was conflict(s) of interest were disclosed, and 19 when there was a statement that there are no conflicts of interest. Differences need to be tempered by the fact that conflict of interest reporting has improved over time, and recent papers have less time to accrue citations.

The median number of Scopus citations was 6 when there was no mention of funding, 1 when it was explicitly stated that there was no funding, and 27 when there was statement of some funding. Thus explicitly funded work attracted more citations.

The median number of Scopus citations was 13 when the paper claimed to have novel findings, 2 when it claimed to be a replication, 1 when it claimed to present both novel findings and replications, 11 when it was unclear, and 0 when there was no abstract/introduction. However, the appearance of fewer citations for replication papers is spurious because replication papers are almost all very recent and did not have time to accrue citations.

**References**

Archambault, E., Caruso, J., & Beauchesne, O. (2011). Towards a multilingual, comprehensive and open scientific journal ontology. *Proceedings of the 13th International Conference of the International Society for Scientometrics and Informetrics*, 66-77.

Begley, C. G., & Ioannidis, J. P. A. (2015). Reproducibility in science: improving the standard for basic and preclinical research. *Circulation Research, 116*(1), 116-126. doi:10.1161/CIRCRESAHA.114.303819

Börner, K., Klavans, R., Patek, M., Zoss, A. M., Biberstine, J. R., Light, R. P., . . . Boyack, K. W. (2012). Design and update of a classification system: The UCSD map of science. *PLoS ONE, 7*(7), e39464.

Boyack, K. W., & Jordan, P. (2011). Metrics associated with NIH funding: A high-level view. *Journal of the American Medical Informatics Association, 18*, 423-431.

Boyack, K. W., & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology, 61*(12), 2389-2404. doi:10.1002/asi.21419

Boyack, K. W., Klavans, R., Small, H., & Ungar, L. (2014). Characterizing the emergence of two nanotechnology topics using a contemporaneous global micro-model of science. *Journal of Engineering and Technology Management, 32*, 147-159. doi:DOI: 10.1016/j.jengtecman.2013.07.001

Boyack, K. W., Newman, D., Duhon, R. J., Klavans, R., Patek, M., Biberstine, J. R., . . . Börner, K. (2011). Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches. *PLoS ONE, 6*(3), e18029.

Boyack, K. W., Patek, M., Ungar, L. H., Yoon, P., & Klavans, R. (2014). Classification of individual articles from all of science by research level. *Journal of Informetrics, 18*(1), 1-12.

Boyack, K. W., Van Eck, N. J., Colavizza, G., & Waltman, L. (2018). Characterizing in-text citations in scientific articles: A large-scale analysis. *Journal of Informetrics, 12*(1), 59-73. doi:10.1016/j.joi.2017.11.005

Carpenter, M. P., Gibb, F., Harris, M., Irvine, J., Martin, B. R., & Narin, F. (1988). Bibliometric profiles for British academic institutions: An experiment to develop research output indicators. *Scientometrics, 14*(3-4), 213-233.

Emmons, S., Kobourov, S., Gallant, M., & Börner, K. (2016). Analysis of network clustering algorithms and cluster quality metrics at scale. *PLoS ONE, 11*(7), e0159161. doi:10.1371/journal.pone.0159161

Foster, J. G., Rzhetsky, A., & Evans, J. A. (2015). Tradition and innovation in scientists' research strategies. *American Sociological Review, 80*(5), 875-908.

Goodman, S. N., Fanelli, D., & Ioannidis, J. P. A. (2016). What does research reproducability mean? *Science Translational Medicine, 8*(341), 341ps312. doi:10.1126/scitranslmed.aaf5027

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the USA, 102*(46), 16569-16572.

Ioannidis, J. P. A. (2014). How to make more published research true. *PLoS Medicine, 11*(10), e1001747. doi:10.1371/journal.pmed.1001747

Ioannidis, J. P. A. (2016). Why most clinical research is not useful. *PLoS Medicine, 13*(6), e1002049. doi:10.1371/journal.pmed.1002049

Ioannidis, J. P. A., Boyack, K. W., & Wouters, P. (2016). Citation metrics: A primer on hot (not) to normalize. *PLoS Biology, 14*(9), e1002542. doi:10.1371/ journal.pbio.1002542

Ioannidis, J. P. A., Fanelli, D., Dunne, D. D., & Goodman, S. N. (2015). Meta-research: Evaluation and improvement of research methods and practices. *PLoS Biology, 13*(10), e1002264. doi:10.1371/journal.pbio.1002264

Iqbal, S. A., Wallach, J. D., Khoury, M. J., Schully, S. D., & Ioannidis, J. P. A. (2016). Reproducible research practices and transparency across the biomedical literature. *PLoS Biology, 14*(1), e1002333. doi:10.1371/journal.pbio.1002333

Kayal, A. A., & Waters, R. C. (1999). An empirical evaluation of the technology cycle time indicator as a measure of the pace of technological progress in superconductor technology. *IEEE Transactions on Engineering Management, 46*(2), 127-131.

Klavans, R., & Boyack, K. W. (2017a). Research portfolio analysis and topic prominence. *Journal of Informetrics, 11*(4), 1158-1174.

Klavans, R., & Boyack, K. W. (2017b). Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge? *Journal of the Association for Information Science and Technology, 68*(4), 984-998. doi:10.1002/asi.23734

Larivière, V., Gingras, Y., Sugimoto, C. R., & Tsou, A. (2015). Team size matters: Collaboration and scientific impact. *Journal of the Association for Information Science and Technology, 66*(7), 1323-1332.

Lin, J., & Wilbur, W. J. (2007). PubMed related articles: A probabilistic topic-based model for content similarity. *BMC Bioinformatics, 8*, 423.

Martin, S., Brown, W. M., Klavans, R., & Boyack, K. W. (2011). OpenOrd: An open-source toolbox for large graph layout. *Proceedings of SPIE - The International Society for Optical Engineering, 7868*, 786806.

Narin, F., Pinski, G., & Gee, H. H. (1976). Structure of the biomedical literature. *Journal of the American Society for Information Science, 27*(1), 25-45.

Small, H., Boyack, K. W., & Klavans, R. (2014). Identifying emerging topics in science and technology. *Research Policy, 43*, 1450-1467. doi:10.1016/j.respol.2014.02.005

Tijssen, R. J. W. (2010). Discarding the 'basic science/applied science' dichotomy: A knowledge utilization triangle classification system of research journals. *Journal of the American Society for Information Science and Technology, 61*(9), 1842-1852.

Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics, 10*, 365-391.

Waltman, L., Boyack, K. W., Colavizza, G., & Van Eck, N. J. (2017). *A principled approach for comparing relatedness measures for clustering publications*. Paper presented at the 16th International Conference of the International Society on Scientometrics and Informetrics, Wuhan, China.

Waltman, L., & van Eck, N. J. (2013). A smart local moving algorithm for large-scale modularity-based community detection. *European Physical Journal B, 86*, 471.

1. https://www.ncbi.nlm.nih.gov/books/NBK25499/ [↑](#footnote-ref-1)