A Multi-perspective Analysis of Retractions in Life Sciences

Bhumika Bhatt
contact@bhumikabhatt.org

Abstract

I explore trends in retracted publications in life sciences and biomedical sciences. Based on nearly seven thousand publications, which comprise the entirety of retractions visible through PubMed as of August 2019, I perform several analyses to understand trends over different axes, including time, countries, journals and impact factors, and topics. This work involved sophisticated data collection and analysis techniques to use data from PubMed, Wikipedia, and WikiData, and study the publications with respect to these axes. Importantly, I employ state-of-the-art analysis and visualization techniques from natural language processing (NLP) to understand the topics in retracted literature. To highlight a few results, the analyses demonstrate an increasing rate of retraction over time and noticeable differences in the publication quality (as measured by journal impact factors) among top publishing countries. Moreover, while molecular biology and cancer dominate retractions, we also see a number of retractions not related to biology. The methods and results of this study can be applied to continuously understand the nature and evolution of retractions in life sciences, thus contributing to the health of this research ecosystem.

1 Introduction

The cutthroat competition in academia, the rush to publish or just the greed of career advancement and scientific grants can lead scientists to publish flawed results and conclusions. While some of these errors can be unintentional flaws and honest mistakes, others are intentional scientific misconduct. According to a 2012 report, 67.4% of the then retractions in biomedical research literature were due to misconduct [1]. In this, 43.4% was due to fraud or suspected fraud, 14.3% due to duplicate publication, 9.8% due to plagiarism and rest due to unknown or other reasons. Another 2018 study found that the most common reasons for retraction in open access journals were errors, plagiarism, duplicate publication, fraud/suspected fraud and faked peer review process [2]. Flawed publications not only undermine the integrity of science but can also lead to the propagation of erroneous data to other genuine publications, and thus retraction of such work is a necessary step to preserve the integrity of the entire scientific research ecosystem.

There exist a few previous publications to understand the reasons, scope, and impact of retractions in life sciences [3, 4, 5, 6, 7, 8, 9, 10]. Researchers have analyzed retractions by intentions (i.e., mistakes, fraud, fabrication, etc.) as above as well as studied the relationships with the journals the retracted publications appeared in. Much of this work has has been manual and cannot benefit significantly from available computational techniques. My work in this manuscript differs from the previous work in both the end goals and the approach. I map the available retraction data to various dimensions such as countries and subject areas and discover previously unknown or unconfirmed trends in the data.

This work has addressed several challenges, most of which stem from the unstructured nature of the data. The affiliations and abstracts for publications are all unstructured. There is no easy way to obtain the authors' countries, for instance, which we need in our analysis. Consequently, I had to gather country data from other sources and join it with the PubMed data using non-trivial automatic and manual analysis. I also perform topic modeling on the abstracts to understand and characterize retracted literature by subject areas. This involved state-of-the-art natural language processing analyses and visualization. Finally, the entire work has involved significant data cleaning and processing to convert the data into a form amenable for analysis and drawing conclusions. In the next few sections, I discuss the methodology along with the findings in greater detail.

This work makes the following major contributions.

- *Comprehensiveness.* This manuscript reports on all retractions until August 2019, as reported on PubMed. To the best of the author's knowledge, this is the biggest dataset on retractions in biomedical literature to be studied.
- *Trends by time and journal impact factors.* The work shows that retraction rate is increasing over time and differs significantly by journals, with most retractions happening in lower impact factors.
- *Trends by countries.* I show that authors from India and China have more retractions in lower impact factor journals while authors from Germany, Japan, and the US have more retractions in higher impact factor journals when compared to India and China.

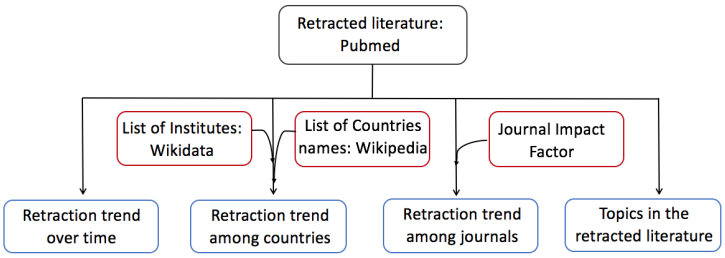


Figure 1: Data and analyses in this manuscript.

- *Trends by subject areas.* This work implements a topic model based on state-of-the-art statistical techniques to identify the subject areas of retracted publications. This technique overcomes the limitations of MeSH (Medical Subject Headings)-based analysis such as the absence of MeSH headings for some articles. While molecular biology, containing genetics and biochemistry, and cancer dominate the topics among retracted papers, we also find a significant number of papers that are not related to life sciences.

2 Methods

I collected the data from PubMed (the PubMed/MEDLINE data is freely available courtesy of the U.S. National Library of Medicine) in August 2019 using Eutilities, an API (application programming interface) for accessing National Centre for Biotechnology Information (NCBI) databases (this includes PubMed) in XML (eXtensible Markup Language) format. From this, I extracted articles' titles, journals' names, publication years, months, dates, authors' affiliations, abstracts and dates of retraction. The entire extraction and processing was done in Python. The obtained data contained 6,940 retracted articles. There were four duplicate entries which had duplication for article title and journal and these were dropped leaving 6,936 unique entries. Figure 1 depicts the data used and the analyses conducted here.

2.1 Processing Dates

My time-based analysis for retractions used publication and retraction dates. The publication dates can be found most often under the `PubDate` node (a node is an element format). However, in 216 cases they were available under the `Medlinedate`. Additionally, there is a `PubmedPubDate`. I chose the minimum of these dates. After this operation, 832 entries had their day missing and so were assigned with day 15 (mean number of days in a month). Another 58 entries, which were missing the month, were dropped for the time-based analysis (I still used these for other analyses).

Retraction dates are available under the `CommentsCorrections` elements as unstructured text. So, I used regular expressions to parse the year, month, and day of the dates. Most (4,211) entries were missing the day, which was again assigned with 15. 44 entries were missing retraction year and 709 entries were missing month or had month entry reported as quarter of the year (summer, fall, winter, and spring). These were dropped for time-based analysis. In 23 cases the retraction date was earlier than the publication date. Further investigation revealed that the correct publication date was not available for them. All such entries were also dropped. Additionally, in seven other cases, the retraction date appeared to be before the publication date but this was due to assigning the day artificially to 15. In total, 765 entries were dropped to calculate time taken for retraction.

2.2 Processing Country Names

Author affiliations can be used to identify which countries the authors belong to. However, this poses a few challenges.

- Affiliations are unstructured and do not follow a common format, it is inherently challenging to derive structured information such as countries from them.
- A country may be known by multiple names, for example, Germany and Deutschland; and Brazil and Brasil. These examples appear in our dataset.
- A country name may not be present in an affiliation but instead must be inferred. For example, if an affiliation mentions "Stanford University" only, it must be inferred that the country is United States.

I addressed the first challenge through an algorithm that scans an affiliation in the reverse order. It uses heuristics to tokenize it into group of words, which it then uses to identify or infer a country. To solve the second challenge, I

obtained a list of alternative country names from Wikipedia. When looking for countries, I searched for all country names, including the alternative ones, and then map the alternative ones to a canonical name.

The third challenge is more complicated to solve. It would be desirable to have a mapping of all the universities and organizations in the world to their home countries. Fortunately, a close approximation of this mapping is possible through ontological databases such as WikiData, which contain crowd-sourced information relating various entities and can be queried to obtain desired mappings. Using the SPARQL (a query language) interface of WikiData, I obtained a list of 84,476 organizations including universities, research institutes, engineering colleges, university systems, international organization, hospitals, businesses, research centers, and academies of sciences and their corresponding countries. While certainly not exhaustive, it would likely cover most of the affiliations we are likely to see in our dataset. Moreover, while neither the list of alternative country names nor the list of organizations may be exhaustive, together they would intuitively provide better resolution of countries than those provided when using one of these methods alone. After processing affiliations in such an automatic fashion, I manually checked them (about 100 were updated).

My methodology involved using all countries appearing in author affiliations, counting each country at most once for a given paper. There were 25 cases where no individual authors but only collective names were mentioned. I manually searched for the countries these groups belong to. The data also lacked affiliations for 439 retracted papers.

2.3 Journals

While journal names are easy to obtain from the PubMed XML, they need augmentation with external data to perform further meaningful analysis. One of the analyses in this manuscript used impact factors, which were obtained from 2019 edition of Journal Citation Reports (JCR) [11]. Before matching the journal names to those in the impact factor list, the names were normalized by making them lowercase, removing spaces and punctuations, and so on. The matching is still not trivial: for instance, an original journal name in the XML is 'JAMA', which is available in the impact factor list as 'JAMA-Journal of the American Medical Association'. Such mismatches were identified and fixed manually.

Furthermore, a few journals, e.g., Biochimica et Biophysica Acta, have several sections, each of which has its own impact factor in JCR. However, the PubMed entries list the journal only without mentioning the section in which the article was published. In these cases, I manually resolved the sections to assign the right impact factor for the journal for the given article. The two manual steps allowed assigning impact factor to 153 additional articles.

2.4 Topic Modeling

Understanding the subject areas and themes of the retracted articles is a major effort in this work. An obvious approach is to use MeSH (Medical Subject Headings) [12] identifiers assigned to each article. MeSH consists of qualifiers, which denote broad subject areas and descriptors, which denote detailed topics appearing in the articles, and are manually assigned to each article. Their primary purpose is to enable efficient search for articles [13, 14]. The data provided by the PubMed APIs contains MeSH, where available. Such a curated list of subject areas and topics would undoubtedly make our work trivial: we could just use the qualifiers to identify the themes in articles. Unfortunately, this does not work for us for at least two reasons: (a) there are 1,713 articles (i.e., about a quarter of the dataset), which do not have any qualifiers assigned; and (b) MeSH provides multiple qualifiers for a given article but does not indicate the relative importance of the different qualifiers, thus making it difficult to identify a single, most-important subject area for the article. An alternative to curated labeling as afforded by MeSH is performing topic modeling, which uses statistical techniques to discover abstract "topics" in given documents (here, retracted literature).

Topic models learn topics from an unlabeled corpus in an unsupervised way. Latent Dirichlet Allocation (LDA) [15] is a state-of-the-art method used for topic modeling. It is a generative probabilistic topic model, based on the idea that each document is a probability distribution over topics and each topic is a distribution over words. Thus, hidden topics in documents can be found out from the collection of words that co-occur frequently. LDA has been used in a number of natural language processing and information retrieval tasks and has even been explored in life sciences literature. For instance, it has been an appropriate tool to discover relationships among drugs, genes, proteins, and pathways in several articles [16, 17, 18, 19]. This work, on the other hand, uses LDA to understand the broad topics appearing in the literature under consideration. Next, I discuss *data preprocessing* to obtain the "words" for the model, and *building the LDA model* to understand the topics.

2.4.1 Data Preprocessing

This step was used to clean and augment the available data to allow building an effective model. Articles that either lacked an abstract or had an abstract describing only their retraction were excluded, resulting in 6,417 articles. Wherever possible, abbreviations were replaced with their respective full forms. I used a simple heuristic for this – upon

encountering a possible abbreviation within parentheses, its letters are matched with the first letters of the immediately preceding words. If a match is found, this abbreviation is expanded all over the abstract. This heuristic only helps with the common case and may miss cases like 'hookworm' abbreviated as 'HW' and 'tuberculosis' abbreviated as TB. The resulting abstracts of these articles were processed through the following steps.

Tokenization and lemmatization. This step broke text into words, removed some words based on their part-of-speech usage, and then normalized. *Part of speech (PoS) tagging* is used to enable *lemmatization* or normalization of words. For this work, part of speech tagging was additionally used to extract only nouns, proper nouns, and adjectives (e.g, 'cell', 'Parkinson's disease', and 'pulmonary'), while ignoring other parts of speech, such as verbs (e.g., 'examined' and 'discovered'), which do not appear useful to our current task. The obtained words were then lemmatized to extract roots of the words. This normalizes the words according to their use as different parts of speech. For instance, the plural 'symptoms' would be transformed to the singular 'symptom'. As life science literature frequently contains phrases, such as 'cell line', 'reactive oxygen species', which have domain-specific connotations not conveyed by the comprising words such as 'cell' and 'line' (when considering 'cell line'), I augmented the derived terms with such phrases. To identify these phrases, I used Named Entity Recognition. The techniques discussed here were implemented through Scispacy [20], a python library extending spaCy [21] for processing scientific and biomedical texts.

Stopword elimination. Stopwords are words that do not have enough meaning to differentiate between two texts. The NLTK [22] library contains 179 stopwords. In addition to these, I added more words such as 'proof', 'researcher', 'record' that have no differentiating effect in life sciences literature, making the stopword list 544 words long. Moreover, I made all words in lowercase to make the subsequent steps case-insensitive and removed all words consisting of digits only or one letter only. In addition, symbols that are common in scientific literature such as $=,<,>,-$ were removed.

2.4.2 Building an LDA model

I now discuss the implementation of the LDA model. The first step is to construct a vocabulary from the terms derived from data processing. Only those terms that appeared in at least twenty articles and in not more than 15% of the abstracts were used. The rationale behind this is that terms appearing in too few documents would not convey any meaningful pattern while those appearing in too many documents would not be restricted to a few topics. This vocabulary is used to construct a *document-term matrix*, where each row corresponds to an article abstract (our documents) and each column corresponds to a word or term. Each cell in the matrix represents the number of occurrences of the term in the document. This is also known as a *bag-of-words (BoW)* model and is an input to the LDA algorithm.

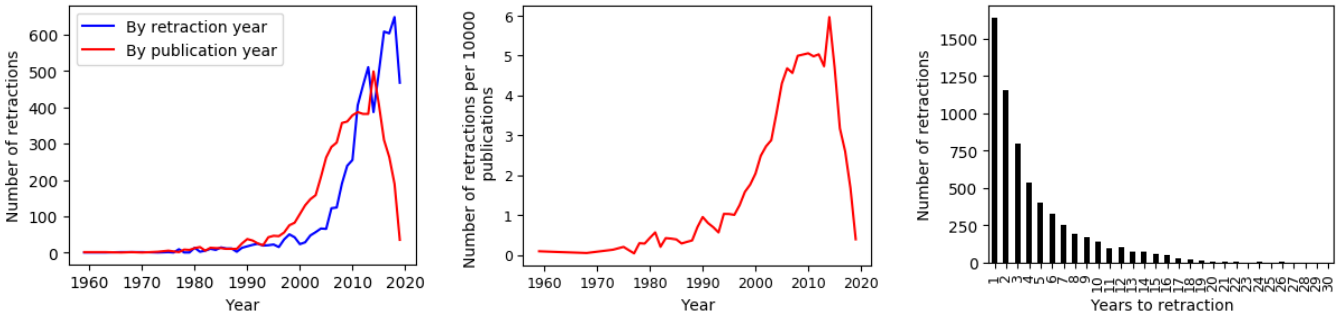
LDA provides as many topics K as defined by the user. A low K can provide broad topics while a high K can give topics with words repeated in multiple topics, thus making them difficult to interpret. To arrive at an appropriate K, I started with a target $K = 15$ and built models for values of K around it to identify the model with the most interpretable topics. I finally selected $K = 16$. The entire implementation was done using the Gensim library [23].

3 Results And Discussion

3.1 Trend over time

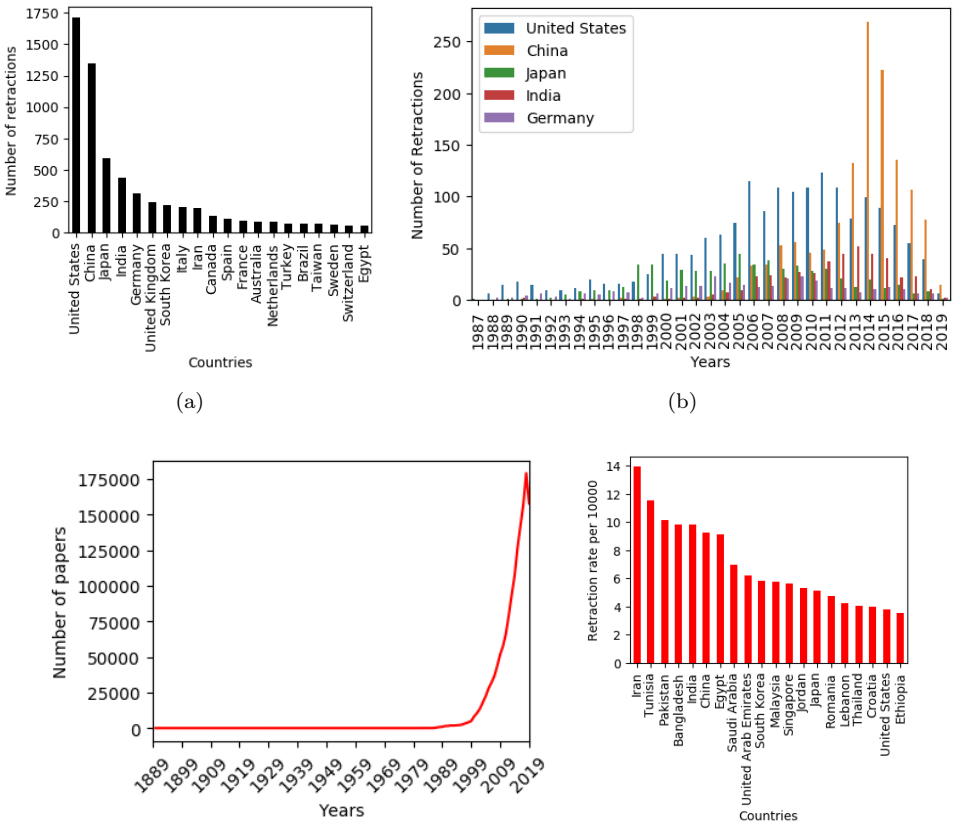
This section derives results from 6,936 unique retracted papers. The earliest publication to be retracted dates back to 1959. As seen in Figure 2a, the number of retractions for the newly published papers has been increasing each year. One may legitimately ask if this rise can be attributed to the increasing number of publications per year. To answer this question, I collected per-year publication statistics from PubMed and plot the retraction rate in Figure 2b. The retraction rate per 10,000 publications was 0.38 in 1985 (here, a retraction is attributed to the year of publication) and it rose to 2.03 in 2000 and 5.95 in 2014. This indicates that the number of retractions has been increasing even after accounting for the increase in the number of publications. Although about six retractions out of 10,000 publications might appear as a small number, the disruptive potential of papers claiming to be breakthrough in their respective fields, including in clinical trials and medical treatments, cannot be underestimated. This trend may possibly be due to a lowering of research paper quality, attributable to fraud, honest mistake, and so on but may also be due to increase in vigilance.

Next, I checked how many publications are retracted in a given year irrespective of their publication year. From Figure 2a, blue curve, it can be confirmed that starting mid to late 2000s there is a steady increase in number of retractions. Additionally, the difference between retraction time and publication date provides the time taken for each retraction. As observed in Figure 2c, maximum retractions happen within 1 year of publication and the number



(a) (b) (c)

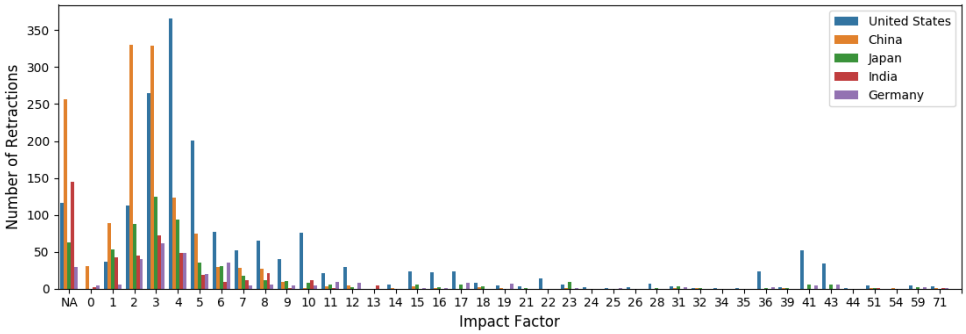
Figure 2: Retraction trend over time. (a) The number of retractions show a generally increasing trend both by the year of publication (red curve) and by the year of retraction (blue curve). Note that the red curve falls off in recent years because there will likely be more retractions in future among publications published in recent years. The blue curve falls off in 2019 because we have only partial data for 2019. (b) Similar analysis as (a), except the number retractions is per 10,000 publications. This shows that even after accounting for the increasing number of publications, the rate of retraction is increasing. (c) Most retractions happen soon after publication.



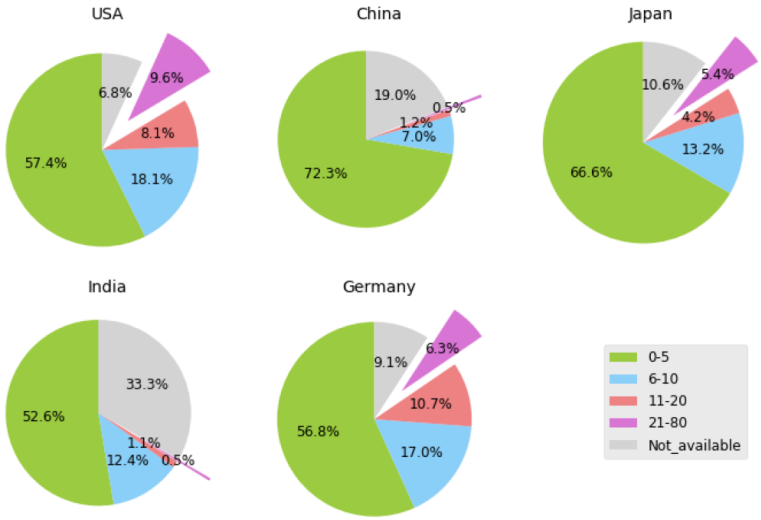
(a) (b)

(c) (d)

Figure 3: Retraction trend among countries. (a) Top 20 countries with highest number of retractions. (b) Retraction trend over time for the five countries with highest retractions. (c) Publication over years for China showing very high increases in recent decades. This is a factor for the high number of retractions for China in recent years. (d) Top 20 countries with highest retraction rate per 10,000 publications.



(a)



(b)

Figure 4: Distribution of retracted publications over their journal's impact factor for the top five countries with highest retractions. (a) Number of retracted publications by their corresponding journal's impact factor. Note: NA impact factor indicates journals without impact factor. (b) Pie charts indicating individual country's distribution of retracted publications according to impact factor. Impact factors have been clustered into five groups: Not available, 0-5, 6-10, 11-20, 21-80.

decreases as years pass by. It takes, on an average, 3.8 years for a publication to be retracted (with standard deviation of 4.01 years) and this explains why we see less number of retractions for the papers published in year 2015 onwards compared to the year 2014 (Figure 2). Additionally, the 25th percentile for years to retraction is about 1 year and the 75th percentile is 5.3 years. The median is 2.3 years. It is worth mentioning that it has in some cases taken over 25 years for retraction.

3.2 Trend among countries

In the available data, 98 countries were found that contributed to the retracted publications, United States, China, Japan, India, and Germany occupy the first five ranks (Figure 3a). Focusing more on these 5 countries revealed that retracted publications from Chinese authors (i.e., authors from China affiliations) soared in mid 2010's especially for the years 2014-2015 (Figure 3b). This increase in retractions correlates well (pearson correlation coefficient = 0.9, for the time span of 2000-2015) with the increase in publications coming from China (Figure 3c) which had a nearly 16 fold increase in 2015 when compared to the turn of this millennium.

While the above record has provided country-wise retractions in absolute terms, it is important to check which countries have high retraction rate (number of retractions per total number of papers published). In order to avoid getting countries that are not publishing actively, I set a threshold that a country should have at least 10,000 publications. This provided the countries that have highest retraction rates – Iran, Tunisia, Pakistan, Bangladesh and India (Figure 3d). According to a 2018 report [24], 80% of the retracted papers from Iran were retracted due to scientific misconduct, ringing an alarm to control scientific fraud.

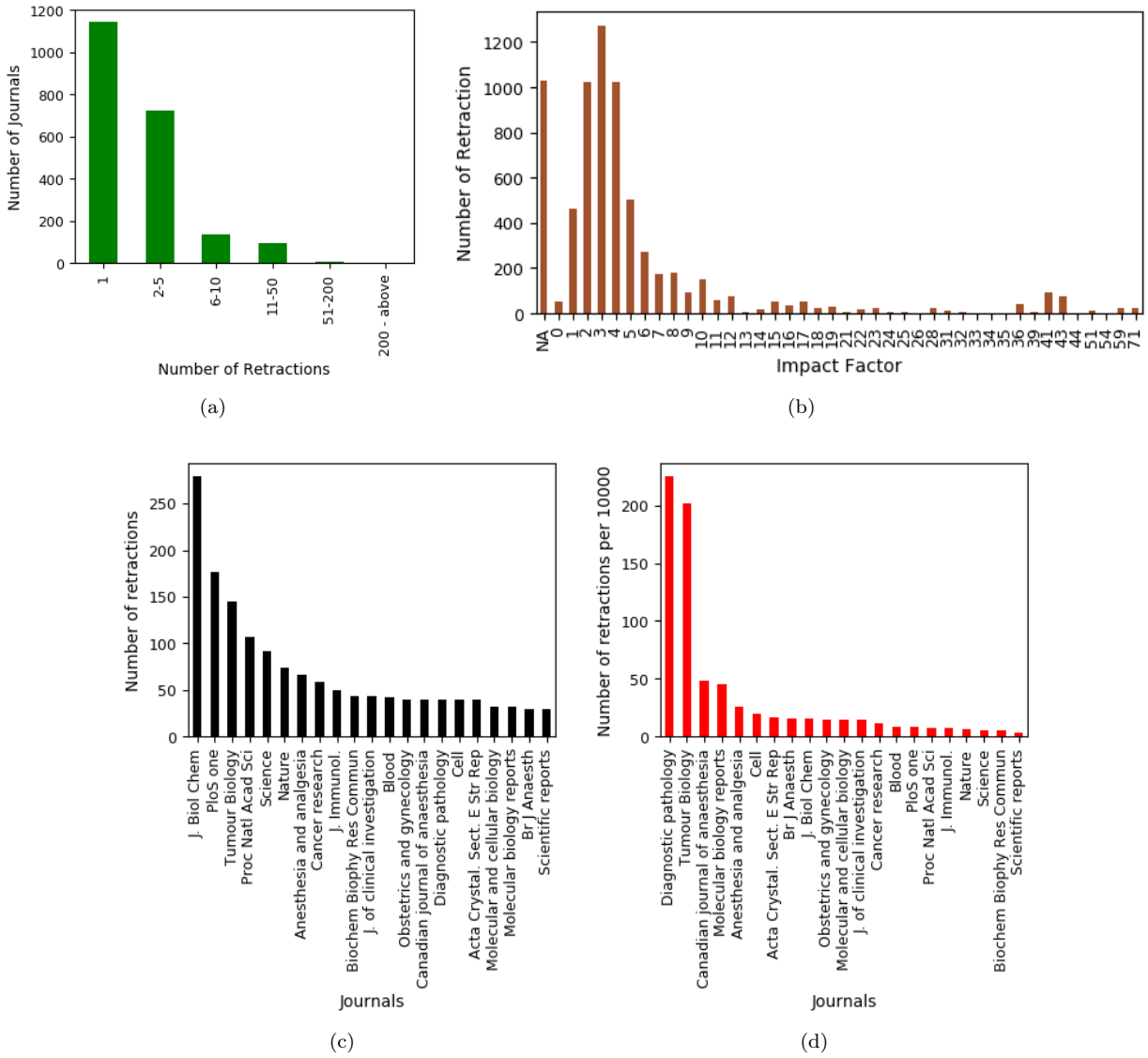


Figure 5: Retraction trend among journals. (a) Number of journals against the number of papers retracted from them. (b) Number of retractions coming from journals with different impact factors (NA - Not available impact factor). (c) Top 21 journals with highest number of retracted publications. (d) Retraction rate per 10,000 publications for journals in (c).

I also compared the distribution of retracted literature over the journal impact factor (IF) for the top five countries. For retractions in journals with IF not available to an IF of 3, China dominates while at higher IFs, the USA dominates (Figure 4a). I examine this pattern more deeply in Figure 4b. As can be seen, India's number of retractions for journals with no IF is higher than its number of retractions in other categories. Moreover, it is also the highest among the five countries when considered as a percentage. Similarly, the highest percentage for journals with IF 0-5 is for China and for journals above 5 is for the USA. Additionally, Japan and Germany also have lower percentage of their retractions in low IF journals (no impact factor and journals with 0–5 impact factor) compared to India and China. This shows a slight skew of distribution of retracted papers for these three developed countries (USA,Japan,Germany) showing more of their retractions in higher impact factor journals compared to the two developing countries, China and India.

My analysis also empirically corroborates the inferences drawn by Fang et al. [1] that countries with long-standing research culture, e.g., U.S., Germany, and Japan, have more retractions in high-impact journals while countries still developing such culture, e.g., India and China have more retractions in lower-impact journals. They infer this by showing that fraud is more prevalent in higher IF journals and in countries with long-standing research culture while plagiarism and duplication is more prevalent in lower IF journal and in countries still developing such culture. The results above empirically highlight this fact.

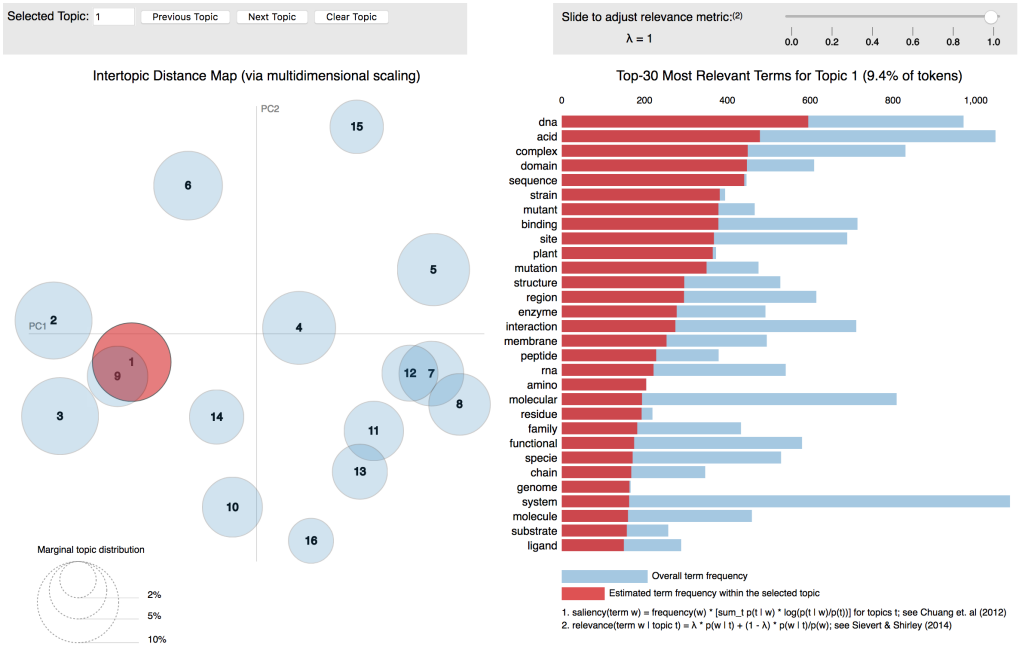


Figure 6: Visualization of topics for an LDA model with 16 topics.

3.3 Trend among journals

The 6,936 retracted papers analyzed here have been published in 2,102 different scientific journals. 54.4% of these journals have only one retracted paper published in them(Figure 5a). Highest retractions are coming from Journal of biological chemistry with 279 retractions, Plos One with 177 and Tumor Biology with 145 retractions (Figure 5c).

In addition, out of the top 21 journals with the highest retractions (I wanted to select top 20 journals but the 20th journal tied with the 21st), Diagnostic pathology followed by Tumour Biology has the highest retraction rate (number of retractions per total number of papers published by a journal) among these 21 journals with highest retractions (Figure 5d). Interestingly, Diagnostic pathology has 4 times less number of publications and nearly 4 times less retractions than Tumor Biology.

Next, I analyzed these journals by their impact factor (IF). 456 journals with 1,030 retractions in our data were missing an IF in the IF list (Figure 5b). With 1,273 retractions, journals with impact factor of 3 had the most retractions. This was followed by 1,030 papers that had no impact factor, which is closely followed by papers with impact factor of 2 and 4. 55.3% of the retractions are from journals with no impact factor or below 3.5 impact factor.

3.4 Topic analysis

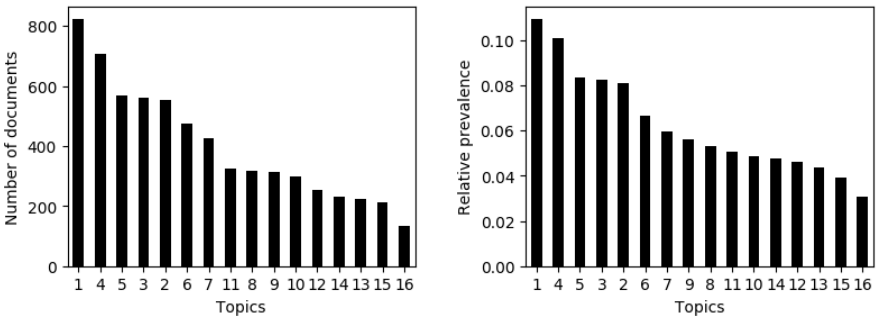
Visualization of results. Recall from the previous section that we selected an LDA model with 16 topics. Figure 6 presents a visualization of topics from this model. The figure is actually a screenshot of the interactive visualization prepared with LDAvis [25, 26]. On the left panel, topics are represented as circles where the center of a circle is determined by inter-topic distances. Thus, the closely related topics are spatially close and conversely, unrelated topics are spatially distant. Additionally, the area of a circle determines the prevalence of the topic. Thus, bigger circles represent more prevalent topics.

On the right panel, a bar chart presents key words for a topic selected on the left panel. The overlaid bars represent topic-specific frequency in red color and corpus wide frequency in grey. Upon hovering over a topic circle in the left panel, its key words are shown in the right panel and hovering over a word in the right panel shows its conditional distribution among topics on the left panel. Figure 6 is a screenshot with the first topic selected.

Topic interpretation. LDA provides topics in the form of co-occurring words. These topics need further interpretation. I analyzed these topics based on the words occurring within them as well as the abstracts that had high prevalence of the respective topics. Table 1 presents the topic analysis: the topic interpretation, the salient keywords representing the topic and remarks indicating how the abstracts are relevant to the interpreted topic.

Table 1: Topics for the constructed LDA model.

Num	Topic	Salient keywords	Remarks
1	Molecular Biology	dna, sequence, strain, domain, mutant, enzyme, mutation, acid, rna, amino	Majorly contains genetics, biochemistry, molecular cloning, sequencing and other techniques
2	Cell growth, proliferation and death	growth, apoptosis, proliferation, cell line, tumor, assay, migration, cycle, cell proliferation, western	Majorly focusses on mechanisms underlying cell growth, proliferation and cell death associated with different cancer types as well as various mechanism of action of inhibitory molecules (biological or chemical) for cancer cell proliferation
3	Signaling pathways	receptor, activation, kinase, phosphorylation, signaling, pathway, nuclear, akt, inhibitor, channel	
4	Interdisciplinary	temperature, nanoparticle, water, field, system, surface, material, electron, performance, paper	Many non-biology related topics such as physics and chemistry.
5	Public health	health, child, care, age, quality, index, body, prevalence, score, risk	Includes epidemiology and health policy and management
6	Cancer	tumor, carcinoma, tissue, metastasis, survival, mirna, hcc (hepatocellular carcinoma), node, nsclc (non-small cell lung cancer), breast	Diagnosis, prognosis, and other characterization of various cancers
7	Circulatory system	heart, cardiac, artery, coronary, myocardial, ventricular, pulmonary, graft, transplantation, infarction	Focuses on surgery and diagnosis related to circulatory system complications
8	Analgesics, antiemetics, anaesthetics	pain, surgery, postoperative, placebo, granisteron, anesthesia, propofol, dose, vomiting, efficacy	Majorly focusses on their usage
9	Immunology	immune, cytokine, inflammatory, macrophage, differentiation, inflammation, intestinal, arthritis, dc (dendritic cell), t cell, cd34	
10	Metabolism	insulin, glucose, endothelial, oxygen, diabetic, vascular, stress, nitric, reactive oxygen specie, antioxidant	Emphasizes on glucose metabolism and oxidative stress
11	Brain and Kidney	brain, renal, injury, pressure, cognitive, cerebral, airway, disorder, neuron, dysfunction	Disorders and complications of brain and kidney and their treatments
12	Bone health	bone, fracture, hip, vitamin, lesion, density, mineral, implant, calcium	
13	Blood	blood, virus, serum, infection, platelet, hes (hydroxy ethyl starch), plasma, syndrome, concentration, albumin	Focusses on intravascular volume therapy
14	Therapeutics and diagnostics	drug, agent, compound, resistance, therapy, extract, inhibitor, natural, delivery, derivative, combination, toxicity	Discusses synthesis and characterization of drugs that include absorption/transport, pharmacological evaluation
15	Risk factors	confidence interval, risk, interval, polymorphism, meta analysis, association, odd ratio, genotype, population, trial, database	Majorly contains meta-analysis
16	Neuro-muscular system	stem, stimulation, nerve, muscle, spinal, cord, pdi (transdiaphragmatic pressure), fatigue, stimulus, contractility	



(a) (b)

Figure 7: Ranking of topics according to (a) the number of documents where they have highest probability compared to the other topics; and (b) relative prevalence.

Analysis. Note that some topics are more closely related than others. For example topic 1 and topic 9, which are about molecular biology and immunology respectively, show an overlap in Figure 6. The topics have some common terms such as complex, region, peptide, interaction, membrane. Also, molecular biology is a broad topic encompassing biomolecules that are functional in various aspects of cell biology, including in immunology, making the overlap intuitive. Similarly, we find topic 1 closely related to signaling pathways (topic 3) and cell growth, proliferation and death (topic 2) and therefore, these topics appear close to each other in Figure 6. For another example, topic 15 (risk factors), which frequently concerns genetic epidemiology, is related to public health (topic 5), which contains epidemiology, and the two topics are not far from each other. Additionally, topic 7 (circulatory system) overlaps with topic 12 (bone health) as well as with topic 8 (analgesics, anti-emetics, anaesthetics), all related to human body and health. Similarly, we observe an overlap between topic 11 (brain and kidney) and topic 13 (blood).

Recall that a single document, which is an abstract in our case, can be made up of multiple topics, but the prevalence (represented as a probability) of these topics in a document can vary. This leads to two ways of ranking the topics:

- By number of documents where the given topic is the most prevalent (Figure 7a).
- By the total prevalence across all documents (Figure 7b). (Note that each document is equally weighted in this ranking because the prevalence of all topics in a document sums to 1.0.)

Topic 1 (Molecular biology) is the top topic according to both the ranking criteria. This is intuitive because the topic is quite vast, encompassing genetics and biochemistry. Topic 4 ranks second by both rankings. Further study reveals that while this topic contains interdisciplinary articles such as on biophysics and computational biology, many of the articles are not related to biology at all. Upon checking 100 randomly selected abstracts of the 705 comprising this topic, I found 49 to be non-biology. Extrapolating this to the rest of 705 abstracts, this suggests that about 5.4% (345 out of 6417) abstracts can be non-biology related. The reason why we see such research articles in our dataset is because the journals where these articles are published are either broad scientific journals (such as Plos One, PNAS, Nature and Science) or generic chemistry or physics or computer science (such as The Journal of Chemical Physics, Langmuir, Neural Networks) that sometimes publish articles relevant to life sciences and biomedical sciences. In either case, PubMed indexes the entire journals irrespective of the article subject areas.

Interestingly, topic 2 (cell growth, proliferation and death) is highly related to cancer. Among randomly selected 100 abstracts from 555 abstracts that have topic 2 as their most prevalent topic, 77 were related to cancer. If we extrapolate this percentage of abstracts related to cancer, then 427 abstracts would be cancer-centric. Adding this number to the number of abstracts from cancer (topic 6), i.e., 474, the total abstracts that are related to cancer would be 901. This would make cancer the most prevalent topic in all the documents.

It can also be observed that some topics are very broad and some noise is creeping in the topics. Noise in LDA models is not unheard of and is especially an artifact of small data as is the case in this study, which consisted of only abstracts from only about six thousand articles. It has been suggested that in small datasets like this one, by using full text (which this study did not have access to) instead of abstracts will largely alleviate the above two mentioned problems of noise and broad topics [27]. Notwithstanding, we are able to see highly interpretable topics as well in our study.

Another limitation of our topic analysis is that we have visibility into only the absolute number of retractions appearing in each topic. Understanding the retraction rate for each topic would be more informative. However, that

analysis would need getting the number of publications in each topic, which is complicated due to each topic consisting of several key words, each having varying probabilities of appearing in different topics. Thus, such analysis would require a more sophisticated study design and is left as future work.

References

[1] Fang, F. C., Steen, R. G. & Casadevall, A. Misconduct accounts for the majority of retracted scientific publications. *Proceedings of the National Academy of Sciences* **109**, 17028–17033 (2012).

[2] Wang, T., Xing, Q.-R., Wang, H. & Chen, W. Retracted publications in the biomedical literature from open access journals. *Science and engineering ethics* **25**, 855–868 (2019).

[3] Steen, R. G. Retractions in the scientific literature: is the incidence of research fraud increasing? *Journal of medical ethics* **37**, 249–253 (2011).

[4] Nath, S. B., Marcus, S. C. & Druss, B. G. Retractions in the research literature: misconduct or mistakes? *Medical Journal of Australia* **185**, 152–154 (2006).

[5] Steen, R. G. Retractions in the scientific literature: do authors deliberately commit research fraud? *Journal of medical ethics* **37**, 113–117 (2011).

[6] Wager, E. & Williams, P. Why and how do journals retract articles? an analysis of medline retractions 1988–2008. *Journal of medical ethics* **37**, 567–570 (2011).

[7] Steen, R. G., Casadevall, A. & Fang, F. C. Why has the number of scientific retractions increased? *PloS one* **8**, e68397 (2013).

[8] Moylan, E. C. & Kowalczuk, M. K. Why articles are retracted: a retrospective cross-sectional study of retraction notices at biomed central. *BMJ open* **6**, e012047 (2016).

[9] Bozzo, A., Bali, K., Evaniew, N. & Ghert, M. Retractions in cancer research: a systematic survey. *Research integrity and peer review* **2**, 5 (2017).

[10] King, E. G. *et al.* Analysis of retracted articles in the surgical literature. *The American Journal of Surgery* **216**, 851–855 (2018).

[11] Clarivate analytics, 2018 journal impact factor, journal citation reports (2019).

[12] Medical subject headings. URL `https://www.nlm.nih.gov/mesh/meshhome.html`.

[13] Lowe, H. J. & Barnett, G. O. Understanding and using the medical subject headings (mesh) vocabulary to perform literature searches. *Jama* **271**, 1103–1108 (1994).

[14] Coletti, M. H. & Bleich, H. L. Medical subject headings used to search the biomedical literature. *Journal of the American Medical Informatics Association* **8**, 317–323 (2001).

[15] Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent dirichlet allocation. *Journal of machine Learning research* **3**, 993–1022 (2003).

[16] Wang, H. *et al.* Finding complex biological relationships in recent pubmed articles using bio-lda. *PloS one* **6** (2011).

[17] Wu, Y., Liu, M., Zheng, W. J., Zhao, Z. & Xu, H. Ranking gene-drug relationships in biomedical literature using latent dirichlet allocation. In *Biocomputing 2012*, 422–433 (World Scientific, 2012).

[18] Zheng, B., McLean, D. C. & Lu, X. Identifying biological concepts from a protein-related corpus with a probabilistic topic model. *BMC bioinformatics* **7**, 58 (2006).

[19] Pratanwanich, N. & Lio, P. Exploring the complexity of pathway–drug relationships using latent dirichlet allocation. *Computational biology and chemistry* **53**, 144–152 (2014).

[20] Neumann, M., King, D., Beltagy, I. & Ammar, W. Scispacy: Fast and robust models for biomedical natural language processing (2019). `arXiv:1902.07669`.

[21] spacy. URL `https://github.com/explosion/spaCy`. Industrial-strength Natural Language Processing (NLP) with Python and Cython.

[22] Bird, S., Klein, E. & Loper, E. *Natural language processing with Python: analyzing text with the natural language toolkit* (" O'Reilly Media, Inc.", 2009).

[23] Řehůřek, R. & Sojka, P. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50 (ELRA, Valletta, Malta, 2010). Http://is.muni.cz/publication/884893/en.

[24] Masoomi, R. & Amanollahi, A. Why iranian biomedical articles are retracted? *The Journal of Medical Education and Development* **13**, 87–100 (2018).

[25] Sievert, C. & Shirley, K. Ldavis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, 63–70 (2014).

[26] Mabey, B. pyldavis. URL `https://github.com/bmabey/pyLDAvis`. Python library for interactive topic model visualization.

[27] Syed, S. & Spruit, M. Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. In *2017 IEEE International conference on data science and advanced analytics (DSAA)*, 165–174 (IEEE, 2017).