TIGER ANALYTICS

Winning with AI: Gaining the First Mover Advantage in KOL Identification

TABLE OF CONTENTS

Click on the tiles to go the respective page

INTRODUCTION

The pharmaceutical industry continues to expand its influence on global health, from controlling diseases to providing treatment options to better manage complex health conditions. It is now becoming increasingly obvious that overall global health can be improved not just by developing drugs and vaccines to manage health but also by ensuring that when treatment options are available, the information is relayed to the prescribers and patients in an unbiased manner so that there is increased adoption.

In the pharmaceutical industry, Key Opinion Leaders (KOLs), who are thought leaders and highly respected for their expertise in specific therapeutic areas, play a key role across the drug lifecycle ranging from drug discovery to analyzing treatment outcomes, to driving R&D efforts and helping with adoption.

Traditionally KOLs helped disseminate information through publications, speaker programs, conferences, seminars etc. However, in the current digital era, KOLs also share their opinions through blog posts, digital news articles, social media, or webinars.

This paper presents an approach to identify KOLs who have risen to prominence on the strength of their research and publications and therefore are sought after for their evidence-based opinions and recommendations.

This paper also presents an approach to KOL prediction that can help identify future KOLs by analyzing the research and publication activity of early-stage professionals. This will help pharmaceutical companies identify Heath Care Professionals (HCPs) with whom they can start building relationships at a very early stage of their careers so that these could potentially be leveraged when they become KOLs in the industry.

WHO ARE KOLS?

KOLs are thought leaders and are highly respected for their expertise in their specialization. KOLs are usually researchers in a particular therapeutic area and may be editors or contributors to key journals or may hold offices in professional associations and are frequent presenters at conferences.

KOLs usually share their opinions through different channels such as speaker programs conducted by pharmaceutical companies, in conferences organized by the government or professional associations, and through various other digital and offline channels.

Due to their strong pedigree and professional stature, KOLs exert a strong influence over other HCPs. Their assessment of various treatment options influences the prescribing patterns of HCPs and in many cases tends to influence patient behavior as well.

As KOLs rise in prominence over the course of their career, they influence increasingly larger audiences of peers and patients. For the pharmaceutical industry, KOLs can therefore have an impact on the entire drug lifecycle right from identifying unmet needs, to drug development, product launch, and market performance.

THE ROLE OF KOL AND WHY THEY ARE IMPORTANT

Across the lifecycle of a drug, KOLs can play various important roles that can have a significant bearing on how the need for a drug is established, how it is developed, how it is launched, and how it performs in the market.

1. **Discovery & Development Phases**

   KOLs can help companies prioritize their development focus by highlighting treatment issues and unmet needs in the therapeutic area.

2. **Clinical Trial Phase**

   KOLs may serve as investigators, which influences the perceptions of other healthcare professionals.

   KOLs may present findings from the clinical trials.

3. **Launch and Post-launch**

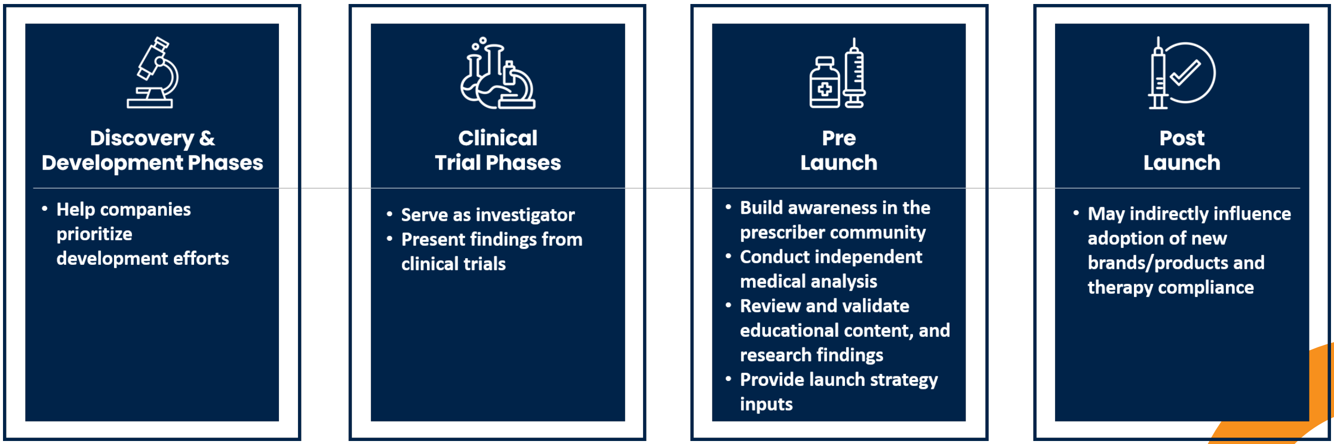
   KOLs may help build awareness in the prescriber community of the disease state, need for diagnosis, available treatment options, and recommend preferred options.

   KOLs may conduct independent medical analysis, and share evidence supporting the class of drugs.

   KOLs may review and validate educational content, and research findings before pharmaceutical companies publish them.

   KOLs may provide inputs to the launch strategy such as the brand attributes to be highlighted, concerns to be addressed, etc.

   KOLs' association with a brand/product may indirectly influence patient behavior towards adoption of the new launch and drive higher therapy compliance.



Discovery & Development Phases

- Help companies prioritize development efforts

Clinical Trial Phases

- Serve as investigator
- Present findings from clinical trials

Pre Launch

- Build awareness in the prescriber community
- Conduct independent medical analysis
- Review and validate educational content, and research findings
- Provide launch strategy inputs

Post Launch

- May indirectly influence adoption of new brands/products and therapy compliance

Figure 1: Role of KOLs at various stages of the Drug Lifecycle

The KOL community comprises many thought leaders with varying clinical expertise and influence on their peers. The hierarchical structure of the KOLs in the below figure depicts the broad categories of KOLs. Over the course of their career, KOLs rise up the ranks as they gain more expertise in their field and consequently their sphere of influence over their peers increases as they move up the hierarchy.

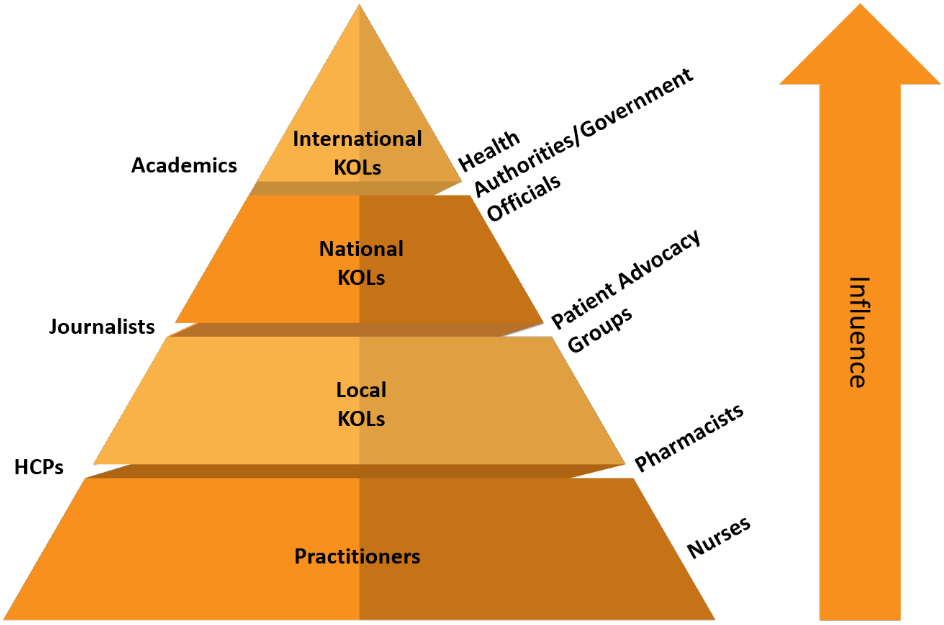


Figure 2: Hierarchical relationship of KOLs. People at the top of this hierarchy are usually more influential.

KOL IDENTIFICATION & ENGAGEMENT

Pharmaceutical companies generally have Medical Science Liaison (MSL) teams within the Medical Affairs group to build and nurture relationships with KOLs.

Traditionally, KOLs are identified based on their academic background, professional affiliations, research, publications, and participation in conferences and events, all of which are lagging indicators while identifying KOLs i.e., these approaches help identify KOLs after they have risen in prominence and established their influence in professional communities.

KOLs, especially the more prominent ones, are overwhelmed by MSLs who want to engage with them for professional networking. So, there is a growing need to identify KOLs early and start building relationships with emerging KOLs early in their career.

This paper presents a data-driven approach to identifying KOLs using PubMed data. Since PubMed data provides a historical snapshot of all research and publications for over fifty years, we can not only identify current KOLs but also map out pathways that the current KOLs traversed to get to their current professional standing.

In addition, this paper extends the identification approach to then predict who among the current early-stage professionals are displaying the right "signals" to potentially become future KOLs. This ability to predict future KOLs will help pharmaceutical companies to engage with potential future KOLs very early in their career and leverage their relationships to develop and launch successful products in the future.

PUBMED DATA

PubMed comprises more than 33 million citations for biomedical literature from Medline, life science journals, and online books. Citations on PubMed may include links to full-text content from PubMed Central and publisher web sites[1]. PubMed also contains abstracts of biomedical literature from several NLM (National Library of Medicine) literature resources. NLM produces an annual baseline set of PubMed citation records in XML format for download every year.

The figure below shows the count of citations on PubMed until Dec 2019.

In addition to this, daily update files are also produced everyday that include new, revised, and deleted citations. Those datasets are available for download or can be accessed through APIs[2].
The analysis presented in this paper uses citations data available on the PubMed website to identify KOLs who might be working on innovative breakthroughs in this industry.

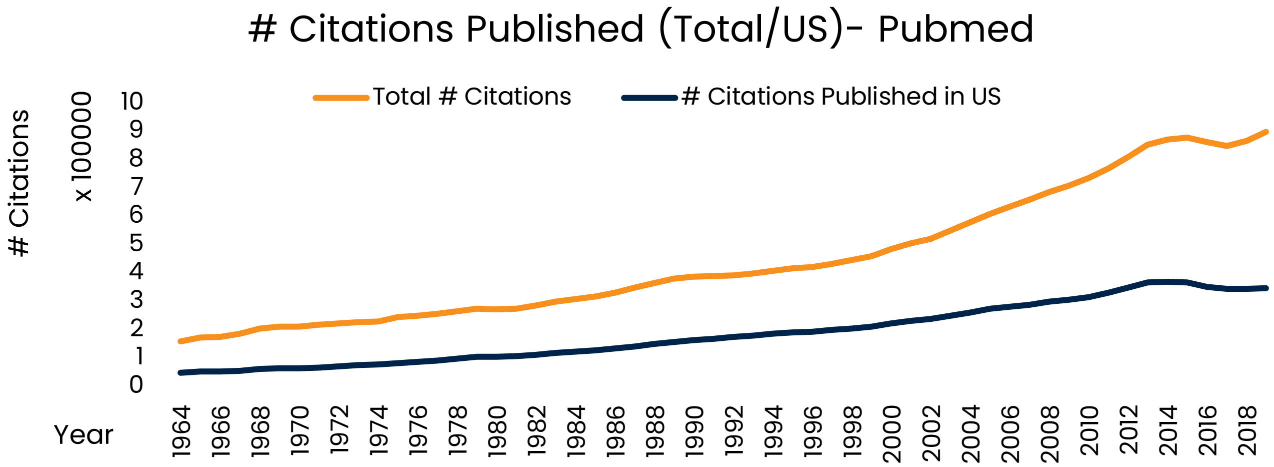


Figure 3 Published Citations by Year[4]

As the above chart indicates, the number of articles published on PubMed has increased over time with over 800K articles published each year since 2012.

This data can be very useful for analyzing various medical domain trends. The analysis presented in this paper uses the citation data to identify key influencers based on their research.

ANALYSIS

Citation data (size>35 GB) downloaded from the PubMed website[1] has more than 1100 XML files. Most of those XMLs contain metadata for more than 100K articles. This data captures over 45 attributes[5] for each published article. Some of those attributes have additional sub-attributes. e.g., 'keywordlist' could be an attribute, and then 'keywords' within this attribute could be sub-attributes. For our analysis, we used the following attributes:

- PubMed Unique identifiers
- Creation/ Completion/ Revision Dates
- Article details
  - Title
  - Keywords
  - Abstract
  - Publication year
- Citation data
- Author(s) details (Last Name, First Name)

Due to the large size of the dataset, the initial data processing attempts took multiple days. We implemented a stepwise data processing approach to process the data within a reasonable period. A few intermediate data summaries were initially created and later used for run-time interpretations.

Since the data did not have unique identifiers for individual authors, identification of the right author was not straightforward. We had to utilize fuzzy name-matching techniques over the authors' given names, the content of the articles, and the affiliations of the authors to avoid duplication of author names.

For this paper, we limited our analysis to the ~9MM electronic articles that were published after 2010. When analyzed, it appeared that many of those articles (~68%) were cited a very limited number of times (<=5). The distribution of articles and the frequency with which the articles were cited has been shown below.

Article Citations – Frequency Distribution

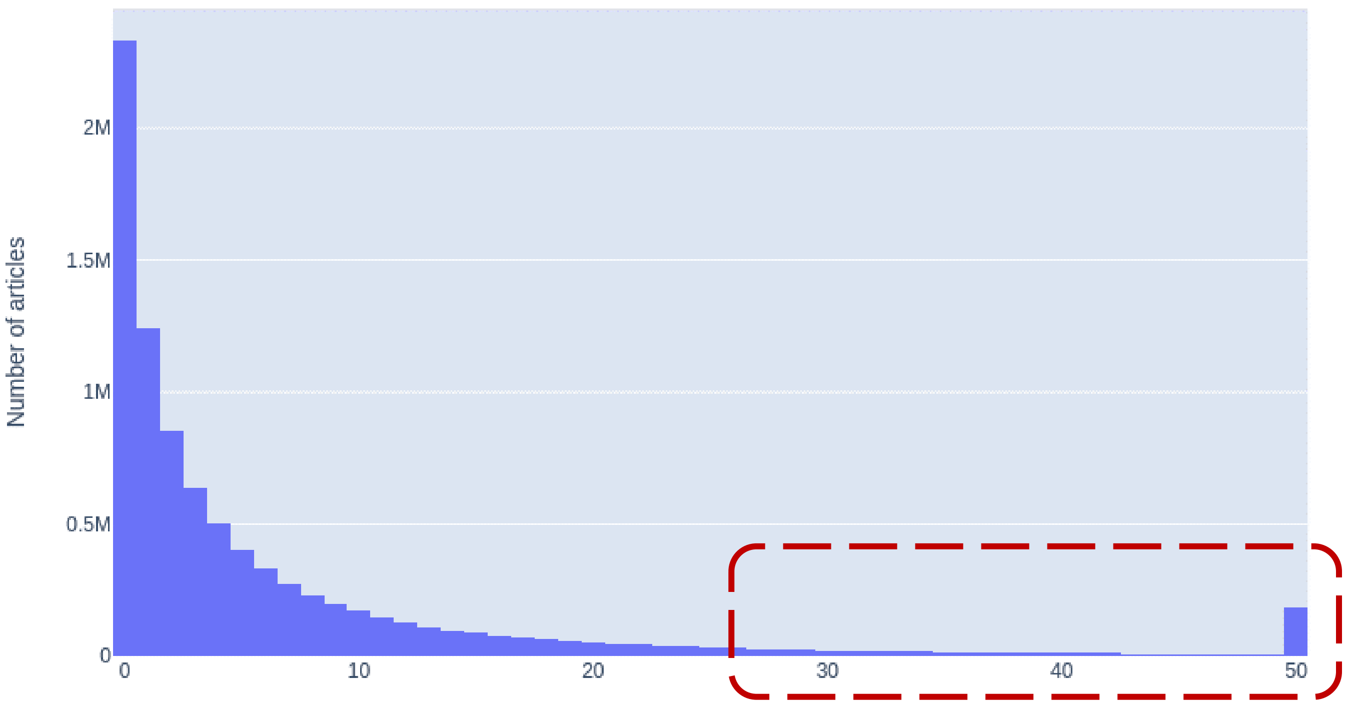
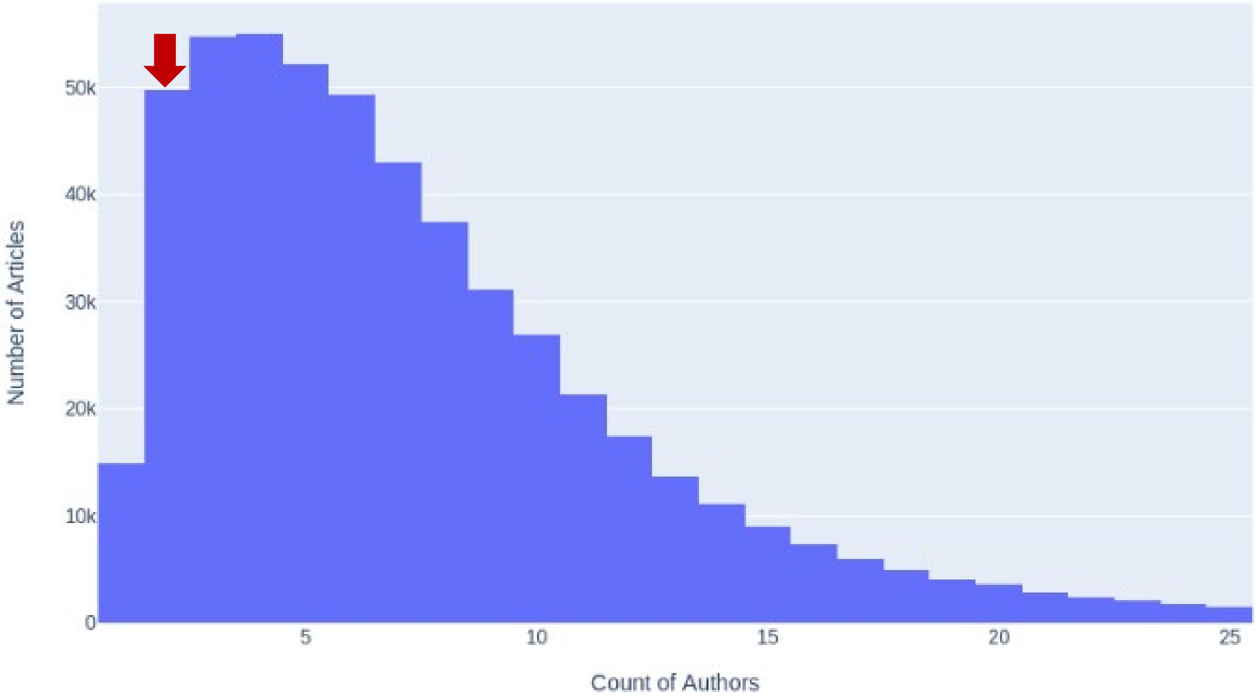


Figure 4: Number of articles Vs. number of times the articles were cited. The red box indicates the articles that were cited more than 25 times and were considered for the analysis. The rest of the articles were considered non-influential

To limit our analysis to a reasonably sized dataset containing the most influential articles, we excluded all articles with less than 25 citations.

This resulted in ~540K articles authored by 1.7MM unique authors. When this data was further analyzed it resulted in the following author frequency distribution.

Number of Articles Vs. Number of Authors (for articles with <=25 Authors)- Total Articles (~524K)



Number of Articles Vs. Number of Authors (for articles with >25 Authors)- Total Articles (~13K)

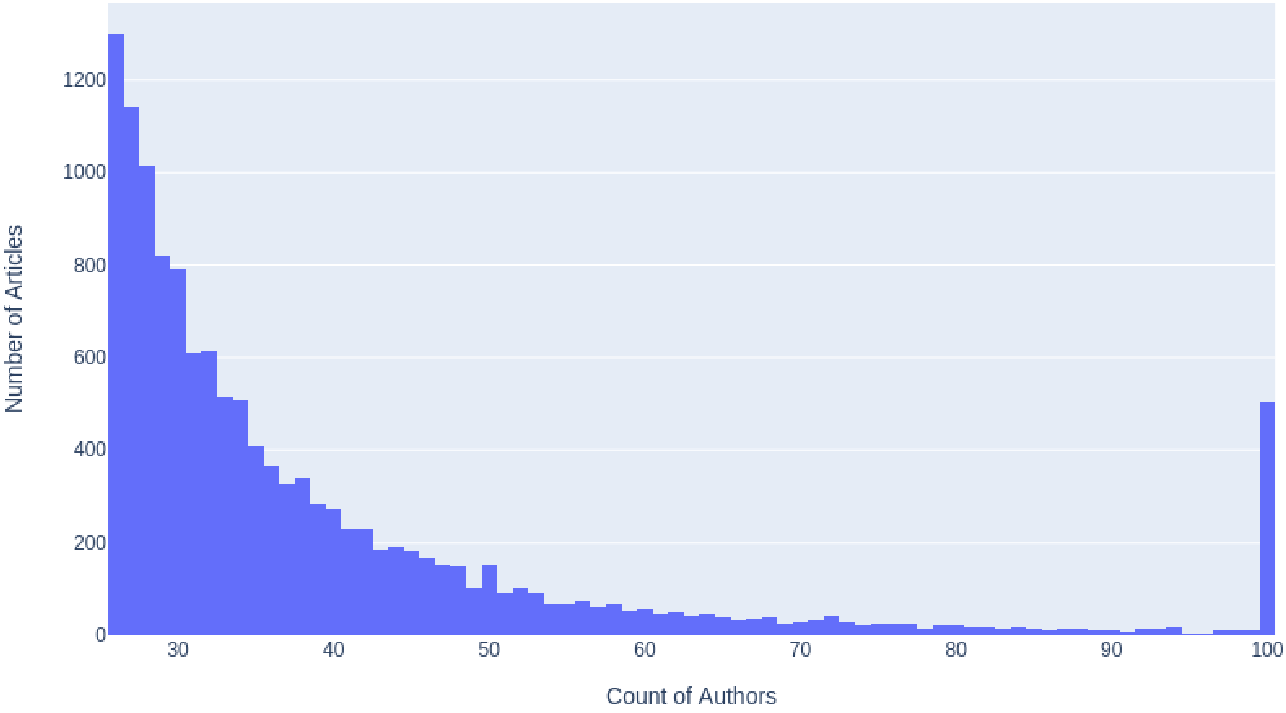


Figure 5: The above charts shows a frequency chart of number authors and number of articles. The red arrow on the above chart indicates that ~50K articles were written by 2 authors

Further, it was observed that in the selected articles that are cited more than 25 times, ~97% of articles had less than 25 authors, and more than 72% of articles had less than ten authors. In addition, there were a few articles (~500) that had more than 100 contributors.

For this analysis, the maximum score of such articles (with more than 100 authors) was set at a lower threshold value to give more weightage to the articles with a smaller number of authors.

KOL IDENTIFICATION

1. Abstract, titles, and author names were used for this analysis. Abstract and title fields are cleaned, tokenized, and processed for integration with the Information Retrieval system.

2. Each article is scored based on the number of times it was cited. For example, for the first run, the score was calculated as the number of times the article was cited.

3. Assigning scores to an article simply based on the number of times cited, would disproportionately bias the model towards older articles that have been available for a longer time. To normalize scores by adjusting for the "age" of the article, an in-year percentile score was assigned to each article.. Within articles published each year, the top 2% articles were assigned a score of 3, the top 90-98% was assigned 2.

   The next tier of articles 70-90% category were assigned a score of 1, articles in the 30-70% tier were assigned 0.5, and then all the other articles were assigned 0.3 scores

4. The score calculated in step #3 is used to calculate a new score. A new score for an article is defined as the sum of scores for all the articles that cited this article.

5. At this stage another module is triggered to adjust the score of articles with more than 100 authors.

6. A loss function was defined as the root mean squared difference of score post and pre updates. Steps #3-5 are recursively performed until the loss function results in a value below the threshold $2*10^{-5}$.

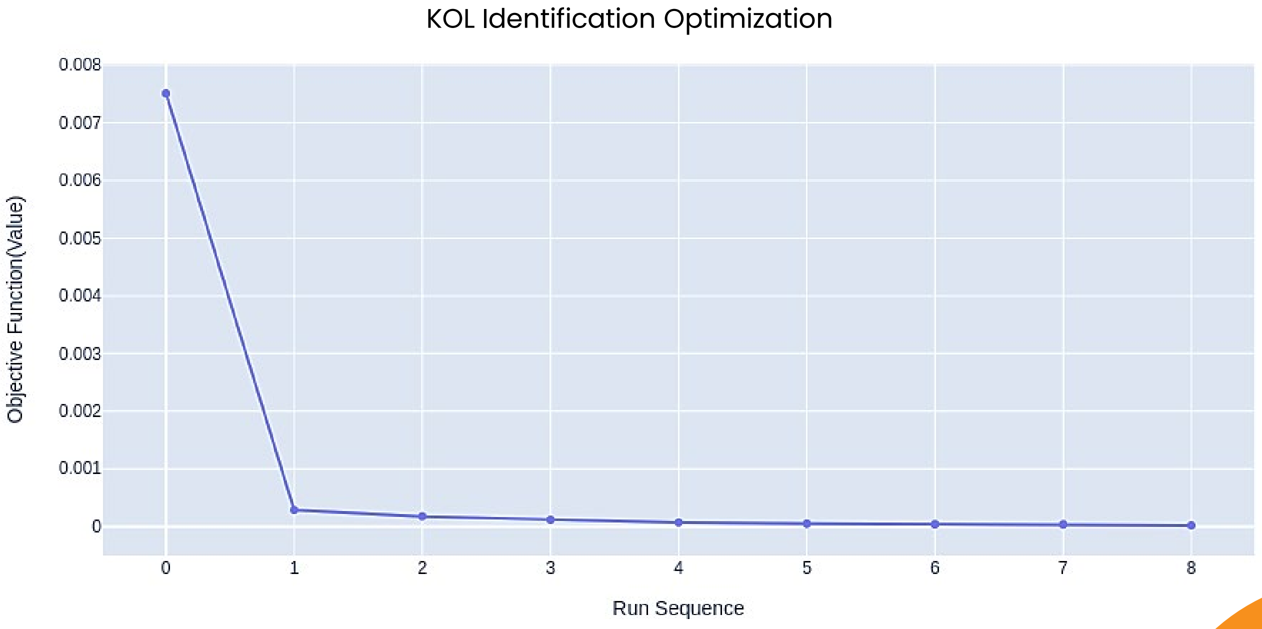


Figure 7: KOL Identification Optimization- Converged after 8th iteration

7. The scores calculated in the last run are stored in a final article level score table.

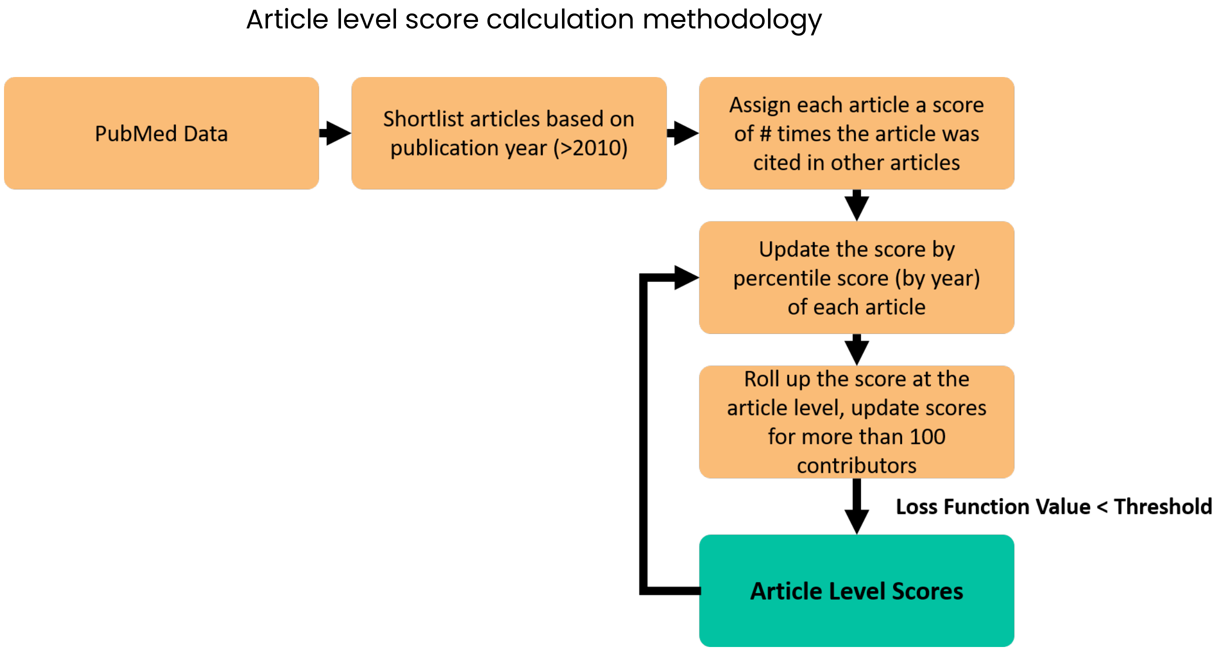

Article level score calculation methodology
PubMed Data
Shortlist articles based on publication year (>2010)
Assign each article a score of # times the article was cited in other articles
Update the score by percentile score (by year) of each article
Roll up the score at the article level, update scores for more than 100 contributors
Loss Function Value < Threshold
Article Level Scores

Figure 8: Article Level Score Calculation Methodology

8. An Information Retrieval (IR) system was set up to find the relevant articles based on the search query. To run the KOL identification analysis, users can write a search query such as "Oncology Lung Cancer" to run the KOL identification analysis. For all the relevant articles thus captured, the score is pulled from the table created in #7. Those scores are rolled up at the author level.

For articles with multiple authors, each the author was assigned the same score (i.e. same as the article score.)

9. This resulted in author-level scores. Authors with the highest scores are identified as KOLs. When validated for a few cases, it is observed that people with top scores have a good presence on the internet such as Twitter and Wikipedia or hold good academic positions.

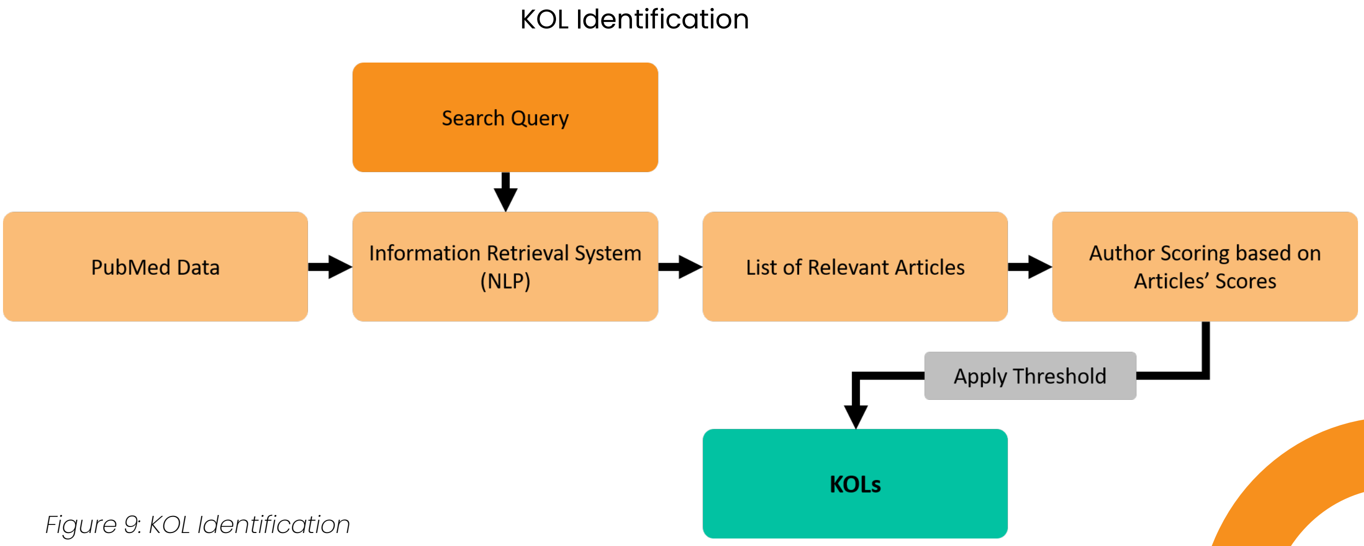

KOL Identification
Search Query
PubMed Data
Information Retrieval System (NLP)
List of Relevant Articles
Author Scoring based on Articles' Scores
Apply Threshold
KOLs

Figure 9: KOL Identification

KOL PREDICTION

There are a few researchers who perform highly reviewed and cited research from the early stage of their careers. They continue to perform similar research and keep publishing their work as their career progresses, and because of the quality of their work, their following and the number of their affiliations increase with time. Some of those individuals have the potential to become KOLs in near future. In this section, a methodology to identify potential future-KOLs at an early stage of their career is presented.

1. One of the assumptions for KOL prediction is that authors who have just started writing (in the last 4-5 years) and whose articles are already being cited in other very influential articles are going to become KOLs in the near future.

2. Article level score calculations are performed using the steps (1-7) mentioned in the previous (KOL Identification) section. The outcome table also captures information about the articles such as title, abstract, etc.

3. The table is filtered based on the first article year of authors. For this analysis, 2016 was selected as the cut-off year, i.e., all the authors who published their first article after 2016 were considered.

4. An Information Retrieval system is used for capturing and processing the inputs to identify the relevant articles. User text inputs are processed against abstract and title of the article. From this outcome table a crosstab is created to capture scores at the author name and year level.

5. From this list authors are selected based on the below-mentioned criteria

   A. Average score /article

   B. Average # KOLs' citations/article (i.e., on average how many KOLs are citing their papers)

   C. # Influential articles

   D. Minimum score for any year (except the last 2 years) i.e., from 2016-2019 for the data that is pulled in Jan'22. This would ensure that the authors are consistently writing influential articles and, the removal of the last two years' data ensures that authors are not penalized because of the lesser 'age' of their articles.

6. Authors with the highest scores and those passing all the above criteria could be potential KOLs in the future.

VALIDATION- KOL PREDICTION

Two different analyses were performed to validate this approach

A. A dataset from 2005-2015 was processed using the above steps and KOLs were predicted for a few search queries. The top 10 potential KOLs per search query were identified.

B. Another dataset from 2010-2021 was processed and was used to identify KOLs for the same search queries as A, using the approach mentioned in the previous section. The top 40 KOLs were selected using this approach.

For most of the cases, 5 or more of the identified potential KOLs in A were present in the outcome received in B. On average 65%+ of the identified KOLs (A) appeared in the actual KOL list (B).

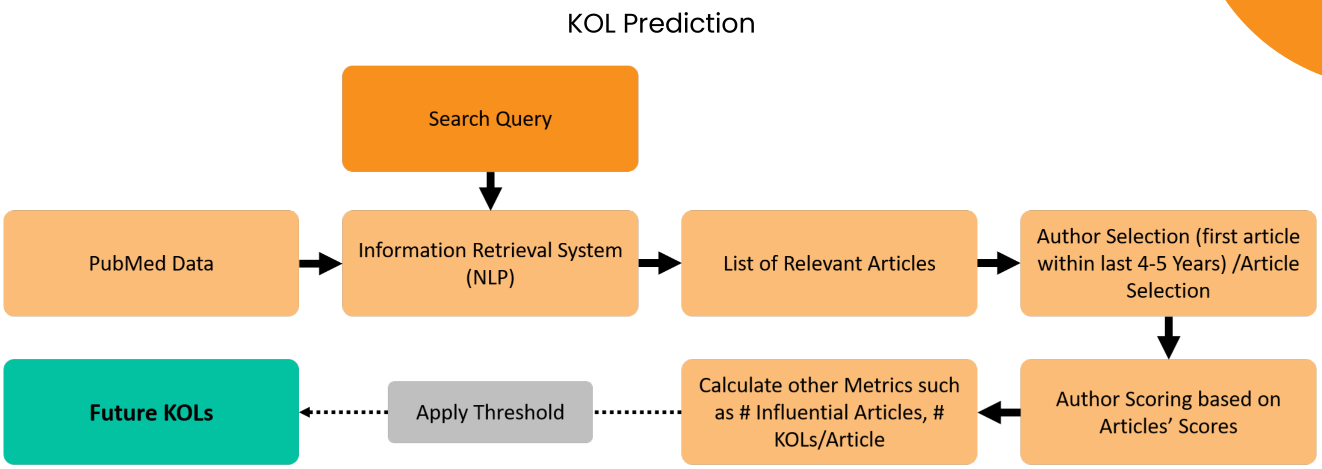

KOL Prediction
Search Query
PubMed Data
Information Retrieval System (NLP)
List of Relevant Articles
Author Selection (first article within last 4-5 Years) /Article Selection
Author Scoring based on Articles' Scores
Calculate other Metrics such as # Influential Articles, # KOLs/Article
Apply Threshold
Future KOLs

Figure 10: KOL Prediction

SUMMARY

The paper presents an approach to how AI can leverage publicly available data to identify current KOLs and predict future KOLs. Users need to provide inputs such as thresholds, scores, article tier information, etc. for KOL identification and KOL prediction analysis. This paper presents an approach and recommends parameter values that worked well in our validation experiments. But users can try different combinations to fine-tune performance to their specific needs. Such as, if a user wants to add more data and consider the last 7 years instead of 5 years for KOL prediction then the input parameters can be updated accordingly. This approach can help organizations leverage AI to gain the first-mover advantage in identifying KOLs.

REFERENCES

1. https://pubmed.ncbi.nlm.nih.gov/

2. https://pubmed.ncbi.nlm.nih.gov/download/

3. https://commetric.com/2018/11/23/identifying-key-opinion-leaders-in-pharma-through-influencer-network-analysis-ina/

4. MEDLINE® Citation Counts by Year of Publication<br /> (as of January 2021)* (nih.gov)

5. https://www.nlm.nih.gov/bsd/licensee/elements_descriptions.html#:~:text=This%20element%20has%20five%20attributes,and%20VersionDate%20as%20described%20below.

ABOUT THE AUTHORS

Asif Ghatala
VP & Head of Life Sciences

As the Head of our Life Sciences Practice, Asif helps Pharmaceutical and Bio-tech companies re-orient to an increasingly data-filled digital world where analytics is the cornerstone of effective decision-making. He leads a practice that designs and develops high performance data ecosystems; and combines it with artificial intelligence to equip businesses with real-time insights for decision-making.

Nitin Khandelwal
Associate Director, Data Science

Nitin is supporting large multinational pharma clients in strategy and Data Science projects across different markets. He has experience with a wide range of global projects across the entire product life cycle. Nitin has used traditional as well as most recent Machine Learning techniques to solve highly complicated predictive analytics problems. His areas of expertise include Machine Learning, Deep Learning (NLP, Computer Vision), Optimization, Forecasting, etc.

ABOUT TIGER ANALYTICS

Tiger Analytics, based in the Bay Area, is pioneering what AI (Data Science, Data Engineering and Application Engineering) can do to solve some of the toughest problems faced by global organizations and at scale. Our 2000+ skilled data scientists, data engineers and business consultants develop bespoke, OpenIP solutions powered by data and technology for 50+ Fortune 1000 companies. We have offices in multiple cities across the US, Canada, UK, India, and Singapore, and a substantial remote global workforce.