COMMUNICATION SCIENCES & DISORDERS

| ISSN 2288-0917 (Online) | Commun Sci Disord 2019;24(1):234-243

Original Article

https://doi.org/10.12963/csd.19579

# Using Text Mining to Identify Trends in Oropharyngeal Dysphagia Research: A Proof of Concept

**Rahul Krishnamurthy, Radish Kumar Balasubramanium**

*Department of Audiology and Speech Language Pathology, Kasturba Medical College, Mangalore; and Manipal Academy of Higher Education, Manipal, Karnataka, India*

Correspondence: Rahul Krishnamurthy, MSc
Department of Audiology and Speech-Language Pathology, Kasturba Medical College, Light House Hill Road, Mangalore, India
Tel: +91-8971875636
Fax: +91-824-2445858
E-mail: rahul.k@manipal.edu, rahulknaidu@outlook.com

Received: January 4, 2019
Revised: February 9, 2019
Accepted: February 21, 2019

**Objectives:** Meta-research can provide valuable insights into patterns in research. Text mining as a specific method to carry out meta-research has been less explored by the speech-language pathologist community. The purpose of this article is to delineate the history of research trends in the area of oropharyngeal dysphagia and its evolution across the past five decades. It also aims to identify hidden patterns in the research field using a combination of text mining and bibliometric-scientometric techniques. **Methods:** We utilized quantitative and qualitative approaches through text mining techniques in the background of scientometric and bibliometric analyses. Abstracts of the articles were text mined from the Scopus database and were subjected to hierarchical cluster analysis and co-occurrence networks analyses. The frequency of published research across journals as well as amount of research articles published overtime were calculated. **Results:** A total of 1,526 articles were published in the area of oropharyngeal dysphagia across 60 journals in the Scopus database. The evolution of research themes has been described. **Conclusion:** The present study summarizes the research that has been carried out from 1970 till present in the area of oropharyngeal dysphagia using a text mining technique. Dysphagia research has evolved to be truly multi-disciplinary, as contribution from various professionals could be observed. The research area of oropharyngeal dysphagia continues to pose new challenges and offers wider prospects for further research.

**Keywords:** Dysphagia, Oropharyngeal dysphagia, Swallowing disorders, Meta research, Text mining

Science is no longer restricted to works of few intellectual academicians. Millions (co)author scientific papers and even more individuals participate in research. The result of this is research output that is massive but also fragmented and often non-transparent, which include several important works of research and equally irrelevant attempts at replication and reduplication. It is highly inefficient to leave research practices to serendipity, biasing influences, methodological illiteracy, and statistical innumeracy (Ioannidis, 2018). An understanding of patterns among existing research is necessary to avoid wasted effort, optimize resources and also to provide the right directions to prioritize research. A relatively new discipline, called the meta-research, aims to provide a bird's eye view at existing research by studying research itself. It is interdisciplinary and can benefit from better tools and methods in statistics and informatics.

The present authors are of the opinion that there exists a lot of hidden information in the scientific literature that cannot be studied from a purely statistical viewpoint. The technique of data mining attempts to bridge this gap and aims to unravel and interpret information inaccessible to statistical treatment, especially when

the magnitude of data is vast (Dörre, Gerstl, & Seiffert, 1999; Gonzalez, Tahsin, Goodale, Greene, & Greene, 2015). Witten, Frank, Hal, & Pal (2017) define data mining as a computational process of extracting new information from existing large amounts of data. Data mining is an umbrella term that refers to classification algorithms (such as decision trees, naïve Bayesian classification, and other classifiers), frequent pattern algorithms (association rule mining, sequential patterns mining, and others), clustering algorithms, graphs, and network algorithms (Che, Safran, & Peng, 2013; Healand, Khoshgoftaar, & Wald, 2014). Piatetsky-Shapiro (2012) opines that data mining includes text mining, image mining, web mining, predictive analytics, and big data techniques.

Text mining is a subfield of data mining that aims to extract valuable new information from existing sources (Feldman & Sanger, 2007). This technique extracts information from within those documents and combines the extracted pieces over the entire collection of existing information to uncover or derive new information. Thus, given as input a set of documents, text mining techniques seek to discover novel patterns, relationships and trends contained within the documents. Specifically, text mining, as an interdisciplinary approach, analyzes data in natural language text through the use of specific algorithms (Cohen & Hunter, 2008; Nie & Sun, 2017). The present study is a combination of text mining and bibliometric-scientometric techniques which would allow the identification of hidden patterns in research field.

Dysphagia has evolved from a relatively young clinical practice area to that of an independent multidisciplinary field (Ciucci, Jones, Malandraki, & Hutcheson, 2016). As with any research and clinical domains, advancements in technology and a rapidly growing knowledge base has driven the field forward. There are national and international organizations, a dedicated interdisciplinary scientific journal and a few universities offering specialized degrees (Clavé & Shaker, 2015). As Ciucci et al. (2016) rightfully discuss the future trend that may influence the science and practice of dysphagia, the present article is an attempt to trace its changing trends. The purpose of this article is to delineate the history of research trends in the evaluation of oropharyngeal dysphagia and its evolution across the past five decades.

## METHOD

The present study can be described as an attempt towards 'research on research' in assessment/evaluation of oropharyngeal dysphagia. It utilizes quantitative and qualitative approaches through text mining techniques in the background of scientometric and bibliometric analyses.

### Source Selection and Search Strategy

The Scopus database was selected as the source for text mining as it includes only those journals that are peer-reviewed and follow stringent publishing standards. We utilized the Scopus database to search using the keywords 'dysphagia', 'oropharyngeal', and 'esophageal' (See Appendix 1 for detailed keyword list). The Boolean operations of 'AND', 'OR', and 'AND NOT' were used in combination with the keywords mentioned above. To not include dysphagias of esophageal origin, the boolean operation of 'NOT' was used. This enabled us to limit to only those articles foucsing on oropharyngeal dysphagia while excluding esophageal dysphagias. The above mentioned keywords were selected based on the experience of the second author and also considering the search fields of the Scopus database. Appropriate filter settings with respect to year of publication and article type were used by the authors (Appendix 1).

The present study aimed to investigate changes in the research trends over the past five decades. To determine the time periods, an initial pilot study considering a 10-year time frame was carried out. The results of this pilot study revealed a very few articles for a span of 10 years. Hence, a minimum time frame of three decades was fixed to provide better insights and also to get a broader perspective of research trends. Based on this the time frames between 1970 to the present were divided into two, time period 1 was considered to be between 1970 to 2000, whereas the years after 2000 to present were considered to be time period 2.

### Information Extraction

Search results from the Scopus databases were handled in two different ways. First, the information on indexation data was imported into excel sheet (.CSV format), and the text was transformed into columns. Only that information related to the title, journal,

year of publication and the author was retained. Second, time periods between 1970 to present was further divided into two for a better understanding of shifts in research paradigms if any existed. Time period 1 was considered to be between 1970 and 2000, whereas the years after 2000 to present were considered to be time period 2, hereafter referred to as first and second time periods, respectively. Data with respect to title and abstract from the time periods were imported into two separate texts (.txt) files.

### Handling of Data

The first aim of the study was to investigate the journal wise distribution of research across the first and second time periods. Frequency tables on published research on oropharyngeal dysphagia and journals were generated using SPSS version 23 (IBM, Armonk, NY, USA).

The second aim of the study was to investigate the evolution of research trends in oropharyngeal dysphagia across the first and the second time periods. For this purpose, a data mining approach was utilized to identify and compare the predominant research themes between the first and second time periods. The data mining was carried out using KH Coder version 3, which is an open source software for computer-assisted qualitative data analysis, mainly quantitative content analysis and text mining. Based on the search operations described in the earlier section, the title and abstract data retrieved from the Scopus database as text (.txt) files were fed into KH Coder for further analyses which have been described below.

Co-occurrence network for words was generated for both the time periods. This method of analysis provides a graphical representation of the association between the words through connected lines. Closely associated themes were color-coded with the size of each node representing the frequency of occurrence. Sentences were considered to be the unit of analyses, and the filter edge was set to 30 words.

Another method used was the hierarchical cluster analysis, represented in the form of a dendrogram that allows analyzing word combinations with similar appearance grouped into patterns. Both these methods allow transformation of data into a visual representation based on the nature of words.

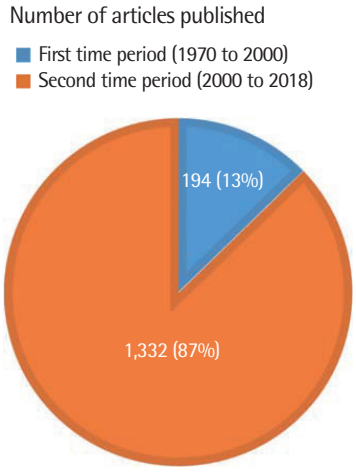


**Figure 1.** Graphical representation of articles published over the first and the second time period.

## RESULTS

The results have been presented in the following headings to provide better insights into the publication and research trends across the first and the second time periods.

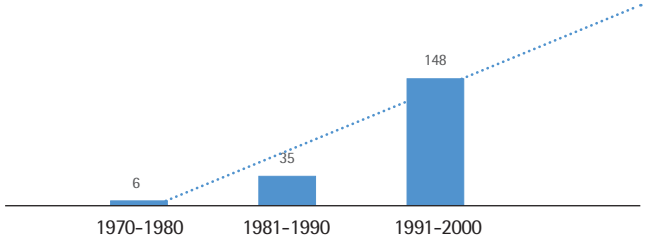


**Figure 2.** Distribution of published research on oropharyngeal dysphagia over the first time period with linear forecast trend line. Number of publications per publication decade.

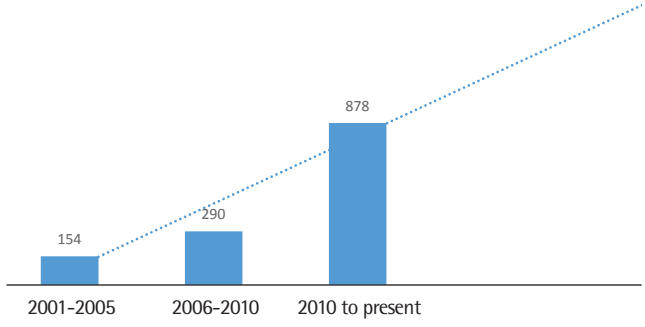


**Figure 3.** Distribution of published research on oropharyngeal dysphagia over the second time period with linear forecast trend line. Number of publications across time periods.


https://doi.org/10.12963/csd.19579

## Amount of Published Research

From 1970 until present, a total of 1,526 articles were published in the Scopus database for the above mentioned search. Among the 1,526 articles, only 194 research articles (13%) were published between 1970 and 2000, and the remaining 1,332 research articles (87%) were published during the second time period. A graphical representation of publishing distribution over the first and the second time period has been depicted in Figure 1. A decade wise distribution of number of articles published in first and second time period have been depicted in Figures 2 and 3, respectively.

**Table 1.** Showing top five journals on oropharyngeal dysphagia during the first time period

| Rank | Journal | Number of publication (%) |
|---|---|---|
| 1 | Dysphagia | 31 (15.97) |
| 2 | Laryngosocpe | 10 (5.14) |
| 3 | Journal of Speech and Hearing Research | 6 (3.09) |
| 4 | Otolaryngology Head and Neck Surgery | 5 (2.57) |
| 5 | Archives of Otolaryngology Head and Neck Surgery | 4 (2.06) |

**Table 2.** Showing top five journals on oropharyngeal dysphagia during the second time period

| Rank | Journal | Number of publication (%) |
|---|---|---|
| 1 | Dysphagia | 176 (13.21) |
| 2 | Head and Neck | 52 (3.9) |
| 3 | Laryngoscope | 52 (3.9) |
| 4 | International Journal of Radiation Oncology Biology Physics | 39 (2.92) |
| 5 | European Archives of OtoRhinoLaryngology | 23 (1.72) |

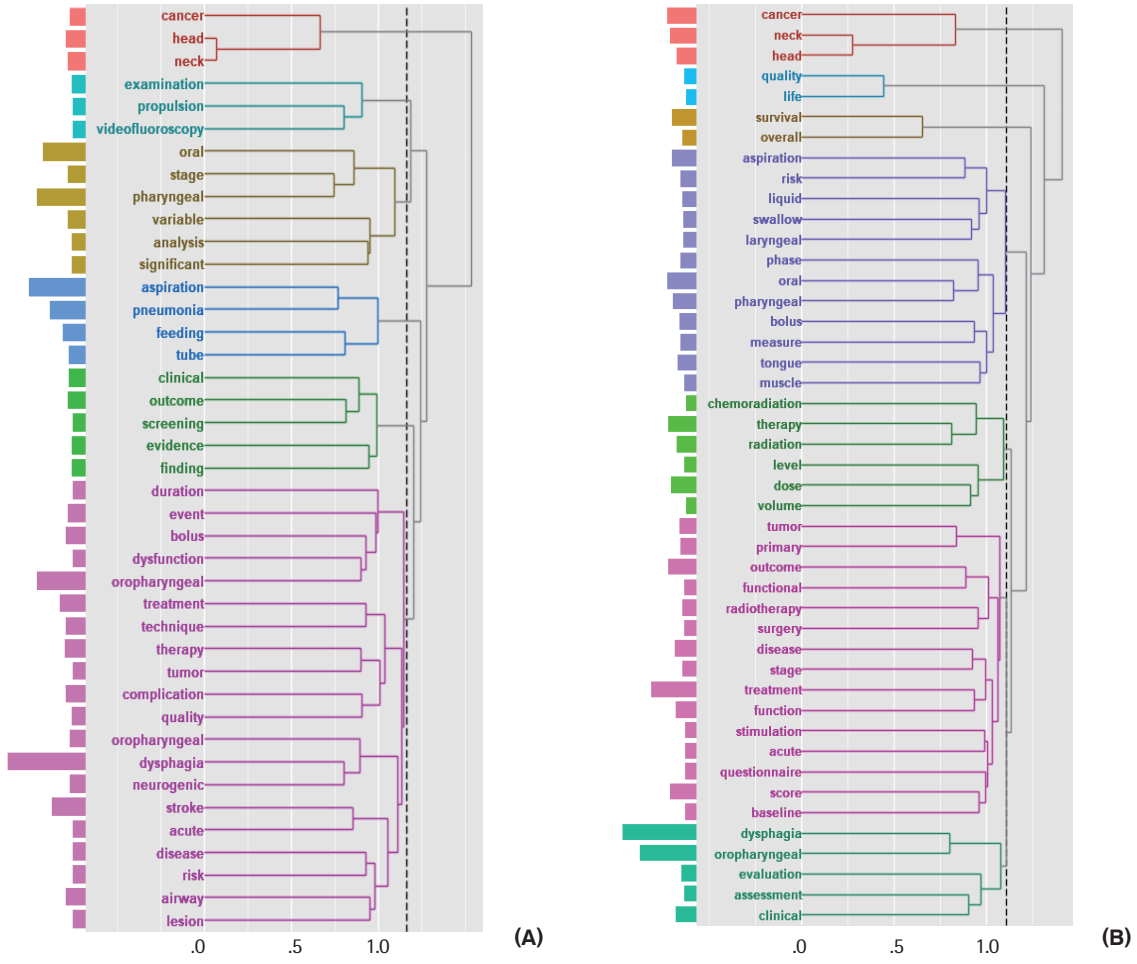


**Figure 4.** Hierarchical cluster analysis for title and abstracts during the first time period (A) and for the second time period (B).

## Journals That Published Most Papers

During the first time period, more than 60 journals published research studies on oropharyngeal dysphagia. The journals were ranked based on the number of research articles published, and the top five research journals have been presented in Table 1.

During the second time period, 92 journals published research studies on oropharyngeal dysphagia. These journals were ranked taking the number of research articles published, and the top five research journals have been presented in Table 2.

## Most Frequently Researched Themes

### During the first time period

To identify the most frequently researched themes hierarchical cluster analysis was carried out. The dendrograms generated for the first time period revealed seven different groups and the bars on the left-hand side of the dendrogram represent the term frequency of each word (Figure 4). When these seven groups were visually inspected the highest frequency was observed for the term 'dysphagia,' hence this was considered to be the first group along with the associated words which are grouped under the same color code. The term dysphagia was frequently associated with 'neurogenic'. Moreover, these two words are in close association with the term 'oropharyngeal'. Thus oropharyngeal dysphagias of neurological origin can be considered the first theme. Other words from this cluster of 'treatment', 'technique', and 'therapy' suggests that dysphagia intervention may be the sub-theme researched during this time period. Trends in the management of oropharyngeal dysphagia is a different research question by itself and are beyond the scope of the present article.

The second group was identified by the next highest frequency of word occurrence. It was found that the term 'aspiration' was the second most frequently occurring word and was in close associa-

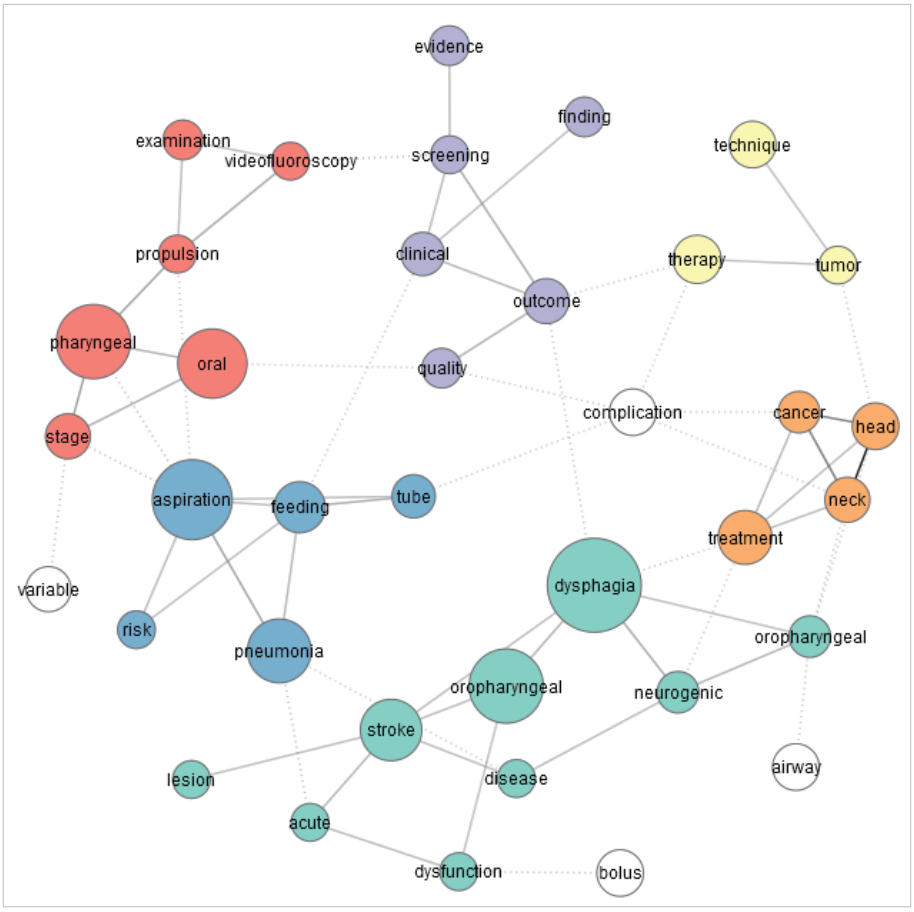


**Figure 5.** Co-occurrence network analysis for title and abstracts during the first time period.

 https://doi.org/10.12963/csd.19579

tion with the term 'pneumonia'. Hence 'aspiration pneumonia' is considered the second theme. The third cluster shows the highly frequent word 'pharyngeal', and 'oral' are in close association with the word 'stage' and 'variable' hence the third theme. The fourth cluster comprises of 'outcome', 'clinical' which are in close association with the terms 'evidence' and 'screening'. This association of word combinations gives an impression that 'evidence-based practice' concerning evaluation may be the fourth theme. The fifth cluster shows common appearing terms to be 'head', 'neck' which are related to 'cancer', hence the fifth theme can be considered as 'head neck cancers'. The sixth cluster consists of 'videofluoroscopy' and 'examination' itself.

Co-occurrence networks showed 37 most frequent words for the first time period (Figure 5). These words were grouped into color-coded clusters with the connecting line representing the association between them. The first cluster (represented in light green) is identified as the most prominent node, representing the highest frequency words 'dysphagia' and 'oropharyngeal'. Other terms 'stroke', 'lesion', 'acute', and 'neurogenic' reveal a strong association with the main terms 'oropharyngeal dysphagia'. The blue community can be considered the second one, as it has a dual connection to the first community. The terms 'aspiration', 'pneumonia', 'feeding', and 'tubes' are very close to each other. The third community (red) is identified as the third and includes the terms 'pharyngeal', 'oral', 'videofluoroscopy', and 'examination'. A closer inspection of the association between the terms in the red community reveals that the central theme of this cluster may be evaluation of oral and pharyngeal stages of dysphagia using videofluoroscopy.

The purple community can be considered as the fourth, with terms like 'evidence', 'clinical', 'outcome' indicating that the common theme may be 'evidence-based practice'. The fifth community (orange) shows common appearing terms to be 'head' and 'neck',

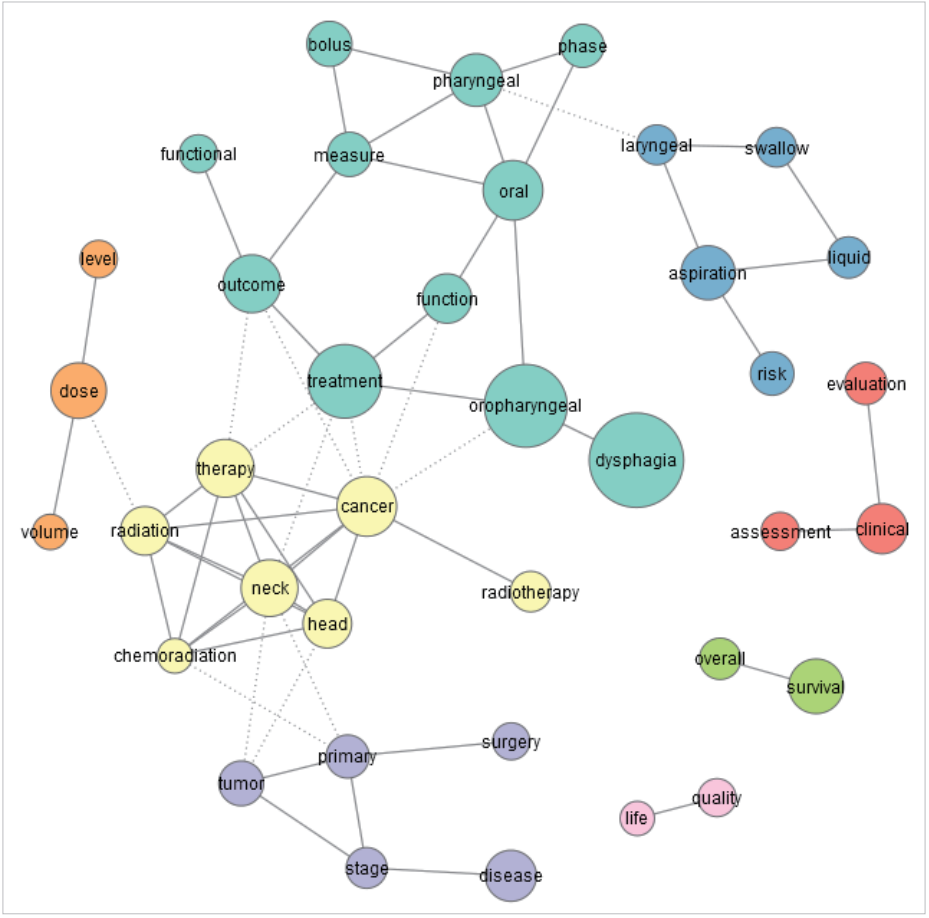


**Figure 6.** Co-occurrence network analysis for title and abstracts during the second time period.

which are related to 'cancer' and 'treatment'.

Further, this community is near the term 'oropharyngeal dysphagia', hence the fifth theme maybe treatment of oropharyngeal dysphagia associated with head and neck cancers. The sixth community (yellow) consists of terms 'therapy' and 'techniques' itself, which may indicate the theme of research to be 'intervention'.

### During the second time period

Hierarchical analysis of title and abstracts from the year 2000 to present (Figure 6) revealed that the highest frequency of occurrence for the term 'dysphagia', was in close association with the term 'oropharyngeal'. The next closet words sets were 'evaluation', 'assessment', and 'clinical', hence it can be considered the first group. Further, it appears that the first major theme of research in the second time period was the clinical assessment of oropharyngeal dysphagias. The second frequent word appears to be 'treatment', which is in association with terms like 'stage' and 'function'. Hence the second theme of research in this time period could be dysphagia treatment outcomes in various stages of cancer.

The third cluster shows the highly frequent word 'oral', and 'pharyngeal' which are in close association with the words 'bolus', 'measures', and 'swallow'. This clustering of word combinations gives us an impression that the third research theme may be 'Effect of bolus characteristics on the oral and pharyngeal phase of the swallow'. The fourth cluster comprises of 'therapy' and 'radiation', which are in close association with the terms 'chemoradiation' and 'level'. This combination of words is in direct association with the terms 'oropharyngeal' Hence the fourth theme can be identified as oropharyngeal dysphagias in patients undergoing chemotherapy and radiotherapy. The fifth cluster shows common appearing terms to be 'head' and 'neck', which are related to 'cancer'; hence the fifth theme can be considered as 'head and neck cancers'. The sixth cluster consists of 'quality' and 'life', which are in close proximity to the cluster 'head neck cancer'. Hence the sixth theme can be identified as the quality of life measures in individuals with head and neck cancers.

Co-occurrence networks (Figure 4) showed 37 most frequent words for the first time period. The first cluster (represented in light green) is identified as the most prominent node, representing the highest frequency words 'dysphagia' and 'oropharyngeal'. Other terms 'treatment', 'outcome', 'measures', and 'functional' reveal a strong association with the main terms 'oropharyngeal dysphagia'. The yellow community can be considered the second one as it has a dual connection to the first community. The terms 'head-neck', 'cancer', and 'chemoradiation', 'therapy' are very close to each other. The blue community is identified as the third and includes the terms 'liquid', and 'laryngeal', 'aspiration', and 'risk'. A closer inspection of the association between the terms in the blue community reveals that the central theme of this cluster may be the influence of bolus characteristics in the prevention of aspiration risk.

The red community can be considered as the fourth with terms like 'clinical', 'assessment' indicating that the common theme may be 'clinical assessment', and is in close proximity to the blue community. This gives us an impression that red community's theme maybe clinical assessments to identify aspiration risks. A small group of literature represented in the light green can be identified as the fifth theme with the terms 'quality' and 'life'. During the second time, period research themes like clinical assessment, dysphagia in head and neck cancers and quality of life measures have been predominant.

## DISCUSSION & CONCLUSION

An inspection of the amount of research published on the evaluation of oropharyngeal dysphagia reveals a definite increasing trend since 1970 till present. This shows the increased attention received by the diagnostic area of oropharyngeal dysphagia over the last five decades. During the first time period, 'dysphagia' journal published the most research articles, which accounted for 15.97% of total publications. Even for the second time period, 'dysphagia' continues to be the top journal publishing the most research articles which accounted for a total of 13.21%, and this can be attributed to the multidisciplinary nature of dysphagia journal.

The six major research themes identified through text mining analyses fall within the prospective areas suggested by Ciucci et al. (2016). Emerging areas such as e-health and tele-health are yet to receive the necessary attention from the dysphagia research community. The interdisciplinary nature of dysphagia research as emphasized by Clavé & Shaker (2015) is very much evident from the

 https://doi.org/10.12963/csd.19579

bibliometric-scientometric distributions of research studies. These considerable interdisciplinary effort across basic, clinical, and translational sciences has facilitated knowledge and skills from different disciplines to come together to provide necessary perspectives for the understanding of dysphagia.

The present study is a preliminary attempt at summarizing all of the research that has been carried out from 1970 to the present. The authors limited the search source only to Scopus database, and further studies can include other databases such as PubMed and Web of Science. Even though the present authors have rigorously and carefully examined the articles before their inclusion in the study, a selection bias may still persist. Some of the reported topics can appear out of its original context and may induce interpretation errors. Further research should consider stringent selection criteria using the same methodology. Even though publication count is one of the most used indicators, it can be subjected to criticism because it reveals only the sheer quantity, and not the quality of the publications. The present authors intentionally limited the number of keywords, so the interpretation could be made. Future researchers can use different combination of keywords, other selection methods and another type of methodological approach; so the validity and precision can be achieved. It would be relevant to reproduce this work of research in every decade to trace the history and trends.

## REFERENCES

Che, D., Safran, M., & Peng, Z. (2013). From big data to big data mining: challenges, issues, and opportunities. In B. Hong et al. (Eds.) *Database systems for advanced applications* (pp. 1-15). Heidelberg: Springer.

Ciucci, M. R., Jones, C. A., Malandraki, G. A., & Hutcheson, K. A. (2016). Dysphagia practice in 2035: Beyond fluorography, thickener, and electrical stimulation. *Seminars in Speech and Language, 37*(3), 201-218.

Clavé, P., & Shaker, R. (2015). Dysphagia: current reality and scope of the problem. *Nature Reviews Gastroenterology & Hepatology, 12*(5), 259-270.

Cohen, K. B., & Hunter, L. (2008). Getting started in text mining. *PLoS Computational Biology, 4,* e20.

Dörre, J., Gerstl, P., & Seiffert, R. (1999). Text mining: finding nuggets in mountains of textual data. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* San Diego, CA, 398-401.

Feldman, R., & Sanger, J. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data.* New York, NY: Cambridge University Press.

Gonzalez, G. H., Tahsin, T., Goodale, B. C., Greene, A. C., & Greene, C. S. (2015). Recent advances and emerging applications in text and data mining for biomedical discovery. *Briefings in Bioinformatics, 17*(1), 33-42.

Herland, M., Khoshgoftaar, T. M., & Wald, R. (2014). A review of data mining using big data in health informatics. *Journal of Big Data, 1,* 2.

Ioannidis, J. P. (2018). Meta-research: why research on research matters. *PLoS Biology, 16*(3), e2005468.

Nie, B., & Sun, S. (2017). Using text mining techniques to identify research trends: A case study of design research. *Applied Sciences, 7*(4), 401.

Piatetsky-Shapiro, G. (2012). The journey of knowledge discovery. In M. Gaber (Ed.), *Journeys to data mining* (pp. 173-196). Heidelberg: Springer.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2017). *Data mining: practical machine learning tools and techniques* (4th ed.). Cambridge, MA: Morgan Kaufmann.

**Appendix 1.** Search keywords in Scopus database

| | |
|---|---|
| Scopus: (TITLE-ABS-KEY (swallowing AND difficulties) OR TITLE-ABS-KEY (oropharyngeal AND dysphagia) AND TITLE-ABS-KEY (oropharyngeal AND dysphagia AND assessment OR evaluation) OR TITLE-ABS-KEY (oropharyngeal AND dysphagia AND management)) | |
| List of stop words used | % |
| | study |
| | age |
| | group |
| | case |
| | result |
| | day |
| | months |
| | method |
| | conclusion |
| | year |
| | first |
| | other |
| | use |
| | significant |
| | common |
| | life |
| | level |
| List of disciplines excluded | Environmental science |
| | Veterinary |
| | Chemistry |
| | Dentistry |
| | Pharmacology, Toxicology and Pharmaceutics |


https://doi.org/10.12963/csd.19579

## 국문초록

### 구인두 연하장애 연구동향을 파악하기 위한 텍스트 마이닝 활용: 개념 증명

Rahul Krishnamurthy · Radish Kumar Balasubramanium

Kasturba Medical College and Manipal Academy of Higher Education, India

배경 및 목적: 메타연구는 연구패턴을 파악하는 데 귀중한 통찰력을 제공해 준다. 메타연구를 수행하기 위해 특정적으로 사용되는 텍스트 마이닝은 SLP 사회에서는 사용된 바가 거의 없다. 본 연구는 구인두 연하장애 영역의 연구동향 역사와 과거 50년간의 변천과정을 파악하기 위해 시행되었다. 또한 텍스트 마이닝과 계량서지-과학계량적(bibliometric-scientometric) 기법을 함께 사용하여 연구분야에 숨겨진 패턴을 찾아내고자 하였다. 방법: 과학계량 및 계량서지 분석을 기초로 하고 텍스트 마이닝 기법을 활용하여 양적, 질적 접근법을 사용하였다. Scopus 데이터베이스로부터 논문초록의 텍스트 마이닝을 한 후 위계적 군집분석과 동시발생 네트워크 분석을 시행하였다. 모든 시간을 통틀어 출간된 연구논문의 총계뿐 아니라 모든 저널을 통틀어 출간된 연구의 빈도도 함께 계산하였다. 결과: Scopus 데이터베이스에서 구인두 연하장애 영역의 논문은 60개 저널을 통틀어 총 1,526개 논문이 출간되었다. 연구주제의 발전상황을 함께 설명하였다. 논의 및 결론: 본 연구는 1970년대부터 현재까지 구인두 연하장애 영역에서 수행된 연구를 텍스트 마이닝 기법을 활용하여 정리하였다. 연하장애 연구는 다양한 전문가의 기여에 의해 진실로 다영역적 차원에서 발전해 왔다. 구인두 연하장애의 연구영역은 새로운 도전을 하면서 지속되고 있고 추후 연구를 위한 더 광범위한 비전을 제공하고 있다.

핵심어: 연하장애, 구인두 연하장애, 삼킴장애, 메타연구, 텍스트 마이닝

ORCID

**Rahul Krishnamurthy** (http://orcid.org/0000-0003-1736-1737); **Radish Kumar Balasubramanium** (http://orcid.org/0000-0001-6485-4644)