Optimizing Literature Search: TEMAS, A New Text-Mining Algorithm-Assisted Search Tool

Emmanuel Bonnet (✉ emmanuelbeaunez@gmail.com)

University of Montpellier: Universite de Montpellier

Daurès Jean-Pierre

University of Montpellier: Universite de Montpellier

Landais Paul

University of Montpellier: Universite de Montpellier

Methodology

Keywords: Bibliography, Literature search, Text Mining, recall rate, precision, web application

Posted Date: September 8th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-71504/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Optimizing Literature Search: TEMAS, A New Text-Mining Algorithm-Assisted

Search Tool

Bonnet Emmanuel*, Daurès Jean-Pierre, Landais Paul

Montpellier University, EA2415, Clinical Research University Institute, Montpellier, France

*Corresponding author: E. Bonnet

Institut Universitaire de Recherche Clinique, 34093 Montpellier CEDEX 5, FRANCE

Tel.: +33 4 11 75 98 39

E-mail: Emmanuel.bonnet@umontpellier.fr

ABSTRACT

Background: Literature search is challenging when thousands of articles are potentially involved. To facilitate literature search we created TEMAS a Text Mining Algorithm-assisted Search tool that we compared to a PubMed reference search (RS) in the context of etiological epidemiology.

Methods: The 4 steps of TEMAS are: 1) a classic PubMed global search 2) a first sort removing articles without abstracts or containing off-topic terms 3) a clustering step with a descending hierarchical classification regrouping articles in independent classes 4) a final sort extracting from the targeted class the abstracts containing the terms of interest, with a link to the corresponding PubMed articles. Validation was performed for risk factors of breast cancer. We estimated the precision and recall rate compared to RS. Average precision and discounted cumulative gain (DCG) were also computed to perform a ranking-based evaluation. We also compared TEMAS results with articles selected in two meta-analyses.

Results: For risk factors of breast cancer, breastfeeding, mammographic density, oral

contraceptive, and menarche were explored. TEMAS consistently increased precision vs RS (from 23% to 32%), with a recall rate from 95% to 97%, and divided the number of selected articles to read from 2.3 to 4.8 times. Mean average precision for 100 articles was 47.4% for TEMAS vs 20.9% for PubMed ranked by best match, and DCG showed a consistent improvement for TEMAS compared to PubMed best match.

Discussion: TEMAS divided the results of a literature search by 3.2, and improved the precision rate, the average precision, and the DCG compared to RS for epidemiological studies. Reducing the number of selected articles inevitably impacted the recall rate. However, it remained satisfactory and did not bias the corpus of information. Moreover, the recall rate was 100% for the two meta-analyses we analyzed, which suggests that the loss of recall rate observed above concerned articles not relevant enough to be included in the meta-analyses.

Conclusion: TEMAS provides a user-friendly interface for non-specialists of literature search confronted with thousands of articles and appeared useful for meta-analyses.

Keywords: Bibliography; Literature search; Text Mining; recall rate; precision; web application

CONTRIBUTIONS TO THE LITERATURE

- Literature search is challenging when thousands of articles are potentially involved. To facilitate it, we created TEMAS which was applied to etiological studies in epidemiology. It divided the results of a literature search by 3, improved the precision rate together with the average precision, and the discounted cumulative gain compared to the PubMed best match.
- TEMAS web application provides a user-friendly interface for non-specialists of

literature search, and does not require any specific knowledge.

- Moreover, TEMAS appears as a potentially useful tool for meta-analyses with a 100% recall rate.
- These findings contribute to make literature search simpler and more efficient.

INTRODUCTION

Literature search is an integral part of any research project. It requires using the correct syntax and the right engine, as well as appropriately selecting and managing information and documents. This last issue becomes complex when the search results bring thousands of articles.

The PubMed search engine indexes more than 30 million citations of the biomedical literature, from more than 5,200 worldwide journals and online books. Several communities have addressed the challenge of automated information retrieval in the literature¹⁻³. Several methods are provided to perform systematic reviews such as PRISMA⁴, MOOSE⁵, or COSMO-E⁶. These methods are very effective when the research question is clearly formulated, however, they afford little guidance when the informational need of the researcher is less circumscribed, for instance looking for risk factors or predictive factors. There are also many methods using machine learning algorithms⁷⁻⁹, or methods focused on specific topics¹⁰. However, these methods do not provide information for sorting without *a priori* a large number of articles.

So we set up a new tool allowing researchers to easily explore very large volumes of data, selecting relevant articles by a text mining approach. We assessed whether TEMAS, a text mining algorithm-assisted search, was more relevant than a classic PubMed search in the field of etiological epidemiology, which studies the causal factors of diseases. Most etiological studies are observational case/control studies or exposed/non exposed cohorts. The validation

of this tool was carried out by comparing TEMAS and a classic PubMed search for non-genetic risk factors of breast cancer and for meta-analyses including case-controls or cohorts studies.

MATERIALS AND METHODS

We developed an interactive web application using Shiny¹¹ to implement our new Text Mining Algorithm-assisted Search (TEMAS) tool. Shiny is a R software¹² package that enables building easily interactive web apps straight from R. Shiny, which combines the computational power of R with the interactivity of the modern web. We complied with MECIR¹³ guidelines for searching and selecting studies (1.5 and 1.6) available in Additional file 3.

This 4 step web application is available at <https://shiny.temas-bonnet.site>.

TEMAS Tool description

A chart describing the TEMAS tool with the proceeding steps appears in Figure 1.

TEMAS Step 1: Global Search

Our new tool allows choosing a date range for the study and entering the search keywords (labeled “Global Search”). An advanced search is possible since it respects the PubMed search syntax. The number of PubMed answers matching the search criteria is displayed.

All the abstracts are retrieved using the package rentrez¹⁴ of R software, which provides an interface to the NCBI’s EUtils API. It enables users to search databases like PubMed (which is, here, the default database), processing the results of those searches, and loading data into their R sessions. The result of this research can be downloaded as a file in CSV format.

We chose a time period from 2006 to 2017 to perform our documentary search¹⁵. We used an advanced search with the following syntax:

((risk_factors[MeSH Terms] OR

```
risk_factor[Title/Abstract] OR  
risk_factors[Title/Abstract]) AND  
(breast neoplasms[MeSH Terms] OR  
breast_cancer[Title/Abstract] OR  
breast_neoplasms[Title/Abstract])  
)
```

TEMAS Step 2: First sort

A first sort allows removing PubMed results without abstracts (letter to the editor, teaching case, erratum). A second pass can be performed to eliminate all articles that contain off-topic terms (selected by the user) of the search. The result of this step is downloaded as a file in TXT format that will be used at the next step.

TEMAS Step 3: Clustering exploration and extraction

Clustering exploration

The resulting abstracts are analyzed using the Rainette package¹⁶. This is a R package for multidimensional analysis of texts and questionnaires that enables statistical analyzes of large corpora of texts¹⁷. More specifically, it allows classifying abstracts by similarity using Reinert's classification¹⁸.

It is a descending hierarchical classification (DHC) carried out in two stages, which offers a global vision of the corpus explored. After a corpus partitioning, statistically independent classes of words are identified, which are characterized by specific correlated terms. This type of analysis enables grouping articles according to concepts. These classes are interpretable according to their profiles, which are characterized by specific correlated terms. The DHC summarizes this process by a dendrogram, which offers the possibility to choose a

clusterization including 2 to 6 classes. Once these classes characterized, the class of interest can be identified among the available classes. This is the class that contains the terms or concepts that best match the user's search.

Clustering extraction

Then the extraction step begins. At this stage, there are two distinct possibilities whether the search deals with a single term or multiple associated terms:

- Single term extraction: Selects the abstracts containing the “term of interest” within the retained class of interest mentioned above.
- Multiple terms extraction: First, target the most representative term, here called “main term”, and define the “complementary term”. Then define the maximum distance (expressed as a number of characters) allowed between “main” and “complementary” terms. The selected articles contain the “main term” and the associated “complementary term” within the defined maximum distance allowed (e.g. for “oral contraceptive”, the “main” term is “contraceptive” and the “complementary” term is “oral”).

These terms were called “extracted term(s)”.

TEMAS Step 4: Final sort

This last step enables carrying out the final sorting of the articles extracted from the former step. A default choice is offered, which only keeps articles the abstracts of which contain at least one of the following terms: Odds Ratio, Odds, OR, Relative Risk, RR, Hazards Ratio, or HR. Other relevant terms for the targeted search may be introduced also.

TEMAS effectiveness analysis

Precision and recall rates

We studied the “relevance” of our information retrieval¹⁹. The information retrieval evaluation is based on relevance metrics, namely recall and precision that were estimated as follows^{20–22}.

$$\text{precision rate} = \frac{\text{number of relevant articles in the RS database (TEMAS, resp.)}}{\text{total number of articles in the RS database (TEMAS, resp.)}}$$

$$\text{recall rate} = \frac{\text{number of relevant articles in the TEMAS database}}{\text{number of relevant articles in the RS database}}$$

Precision is the fraction of relevant instances among the retrieved instances. Recall is the fraction of the total amount of relevant instances that were actually retrieved. Both precision and recall are therefore based on an understanding and measure of relevance.

The measurement of precision and recall rates requires a qualified individual to inspect the output from a search and to address the output into two groups of articles: relevant and not relevant.

The recall rate cannot be calculated for the RS. Indeed, all articles in the PubMed database should be read to calculate the RS recall rate. The recall rate of a PubMed search is not 100% since there may be relevant articles that are not extracted by the RS. Conversely, it was possible to calculate TEMAS recall rate relative to RS. Precision was calculated for both TEMAS and RS.

To assess TEMAS’ efficiency versus a classic PubMed search, we performed a reference search (RS) with the syntax below:

("Global Search" AND

("extracted word(s)"[MeSH Terms] OR

"extracted word(s)"[Title/Abstract])

)

This RS retrieved n_{RS} articles, which corresponds to the total number of RS articles.

In order to ascertain the content of each selected article, a stratified random sampling of 300 articles to read was performed; $\frac{n_{TEMAS}}{n_{RS}} \times 300$ articles in the n_{TEMAS} articles and $\frac{n_{RS} - n_{TEMAS}}{n_{RS}} \times 300$ articles in $n_{RS} - n_{TEMAS}$ articles.

A figure explaining the stratified random sampling is available in Additional file 1.

Two authors (EB and PL) independently read the 300 randomly selected articles for each risk factor to assess their relevance. The criterion of relevance was the presence, in the abstract, of an odds-ratio (OR), a relative risk (RR), or a hazards ratio (HR) for the “extracted word(s)”.

The final step in the comparison was to verify that a bias was not introduced in the representativeness of the articles selected by TEMAS in relation to the RS articles. We performed a Chi2 test (with simulated P-value for small numbers) to compare the distribution of OR (< 1 , $= 1$ or > 1) among the relevant articles retained by TEMAS and RS.

Average Precision and Discount Cumulative Gain (DCG)

To evaluate our method, we also calculated the average precision and the Discount Cumulative Gain (DCG)²³ to take into account the notion of ranking. We therefore retrieved the first 100 most recent TEMAS articles ranked by decreasing publication date, and the first 100 PubMed articles classified by "Best Match"²⁴, the relevance search for PubMed.

Precision at n ($P@n$) is the proportion of the top- n documents (the first n in the ranking) that are relevant. If $\{r_1, r_2, \dots, r_k\}$ are the relevant documents retrieved at rank n the first n in the ranking, then: $P@n = \frac{k}{n}$. The Average Precision (AP) is the mean of the precision scores

after each relevant document retrieved:

$$AP@n = \frac{\sum_{i=1}^k P@r_i}{k}$$

We computed $AP@n$ for $n \in \{25, 50, 75, 100\}$

The Mean Average Precision (MAP) is the arithmetic mean of the average precision values for an information retrieval system over a set of q queries. Let $AP_i@n$ be the average precision at n for query i :

$$MAP@n = \frac{1}{q} \sum_{i=1}^q AP_i@n$$

Discounted Cumulated Gain (DCG) assumes that the greater the ranked position of a relevant document, the less valuable it is for the searcher, because the less likely it is that the searcher will ever examine the document, and at least has to pay more effort to find it. DCG formalizes these assumptions by crediting a retrieval system for retrieving relevant documents by relevance, which is discounted by a factor dependent on the logarithm of the document's ranked position. Let r_i ($i \in \{1, \dots, n\}$) be the rating of article i ; $r_i = 0$ if the article is not relevant and $r_i = 1$ if the article is relevant.

$$DCG_n = r_1 + \sum_{i=2}^n \frac{r_i}{\log_2 i}$$

Meta-analysis

We tested whether TEMAS would be suitable to find all the articles selected by the researchers in meta-analyses. On the one hand, we performed a reference search following the queries described in the meta-analysis method to retrieve the articles. On the other hand, we followed TEMAS method from step 1 to step 3 using the same query strategy. If the relevance criterion in the meta-analysis was the presence of an OR, a RR, or a HR, we stopped TEMAS at step 3. We performed TEMAS step 4 otherwise.

We then retrieved all the articles used in two meta-analyses and compared these articles to those retrieved by the reference search and TEMAS.

Material studied

We performed these effectiveness analyses for several searches:

- Precision and recall, average precision and DCG were calculated for four non-genetic risks factors of breast cancer: breastfeeding, menarche, oral contraceptive, and breast density;
- Comparisons with two meta-analyses were performed. The first on physical activity, the second for age at menarche and at menopause as well, as risk factors for breast cancer.

RESULTS

TEMAS search

TEMAS step 1: Global Search

With this Global Search (GS) we obtained a set of 15,582 articles.

TEMAS step 2: First sort

We excluded articles without abstract, and we defined the following terms as off topic since they were downstream the screening procedure:

- “mastectom” for mastectomy, mastectomies...
- “metasta” for metastatic, metastasis...
- “chemot” for chemotherapy, chemotherapies
- “surger” for surgery and surgeries
- “surgic” for surgical

Among the articles extracted in Step 1, we excluded 954 articles without abstract and 4,582 off-topic articles. We thus obtained a database of 10,046 classifiable articles.

TEMAS step 3: Clustering exploration

Clustering:

As a result of the classification, 99.9% of the abstracts were classified in five disjoint classes as shown on Figure 2.

Class 1 gathered genetic terms such as: polymorphism, gene, allele, genotype, DNA or genetic.

Class 2 included risk factors such as menopause, BMI, age, hormone, parity for the most relevant terms. It also comprised “statistical” terms such as: ratio, confidence interval (CI), hazards ratio (HR), relative risk (RR), logistic, Cox. In this class we therefore expected to find articles relative to risk factors corresponding to our criterion of relevance, which was the presence in the abstract of an odds-ratio (OR), a relative risk (RR), or a Hazards Ratio (HR) for the “risk factor studied”.

Class 3 gathered all breast tumor detection methods with terms such as MRI, ultrasound, mammogram, image, or biopsy.

Class 4 was related to public health aspects of screening. Indeed, the selected terms were: screen, program, access, public, social or communication. Thus this class referred to a systemic approach and management of breast cancer screening rather than a risk factors search.

Class 5 was focused on treatments as suggested by the first terms of the class: therapy, treatment or treat. This class was therefore of less interest because it referred to conditions that were downstream the screening of breast cancer.

Therefore Class 2 appeared to be the most appropriate class of interest because it contained information related to risk factors.

We illustrate single, multiple terms extraction or both as follows.

Single term extraction:

- Menarche: the extraction term was “menarch” and extraction led to 246 abstracts.

Multiple term extraction is illustrated by the two following examples:

- Oral contraceptive: the main term retained was “contracept” and the complementary term was “oral” with a maximum distance of 25 characters. This extraction led to 144 articles.
- Breast/mammographic density:
 - i. Breast density: the main term was “dens” and the complementary term was “breast” with a maximum distance of 10 characters. This extraction led to 204 articles;
 - ii. Mammographic density: the main term was “dens” and the complementary term was “mammo” with a maximum distance of 25 characters. This extraction led to 316 articles;
 - iii. After removing duplicates, this extraction led to 368 abstracts.

Both single and multiple terms extraction:

- Breastfeed/Breast-feed:
 - i. Breastfeed single term extraction: extraction term was “breastf” and extraction led to 162 abstracts;
 - ii. Breast-feed multiple term extraction: main term was “feed” and complementary term was “breast” with a maximum distance of 10 characters. This extraction led to 176 abstracts;
 - iii. After removing duplicates, this extraction led to 189 abstracts.

TEMAS step 4: Final sort

We chose the default choice which only kept articles the abstracts of which contain at least one of the following terms: Odds Ratio, Odds, OR, Relative Risk, RR, Hazards Ratio or HR.

All the complete search queries are available in Additional file 2.

Thus the final TEMAS and RS databases for the 4 risk factors are shown in Table 1

TEMAS effectiveness analysis

Precision and recall rates

The calculation method and results for precision and recall are displayed in Figures 3 to 6.

For “breastfeeding” we performed a stratified random sampling of 300 articles as follows: 130 out of 142 in the TEMAS set, and 170 out of 186 ($n_{RS} - n_{TEMAS}$) articles for the RS set. This gave a representative sample of 300 articles. The RS returned 23% of relevant articles. On the other hand, the precision of TEMAS was 51%. Thus, TEMAS effectively concentrated the relevant articles by dividing by 2.3 the number of articles to be read and by increasing by 28% ($p < 0.0001$) the precision rate. Moreover, the recall rate of TEMAS was 97%. The precision and recall rates for the other risk factors studied are displayed in Table 2.

For “menarche” TEMAS divided by 2.8 the number of relevant articles to be read and increased by 23% ($p < 0.0001$) the precision rate.

For “oral contraceptive” TEMAS divided by 2.7 the number of articles to be read, and increased by 32% ($p < 0.0001$) the precision rate.

For “breast density” TEMAS concentrated the relevant articles dividing by 4.8 the number of articles to be read, and increased by 26% ($p < 0.0001$) the precision rate.

For the four risk factors studied, the slight loss in the recall rate did not affect the representativeness of TEMAS with a distribution of ORs, shown in Table 3, not significantly different from RS ($p = 1$).

Average Precision and Discount Cumulative Gain (DCG)

The results for AP and DCG are displayed in Tables 4 and 5. For all the risk factors studied the average precision and the DCG are higher with TEMAS than with a reference search on PubMed. *MAP@25* is 60.16% for TEMAS and 24.22% for PubMed, and *MAP@100* is still better with TEMAS (48.28%) vs PubMed (21.5%).

Meta-analysis

We focused on two meta-analyses to test the ability of TEMAS to retrieve all the articles selected by the researchers to perform their meta-analysis.

The first one was focused on physical activity and the risk of breast cancer²⁵ and based on prospective studies. The authors selected 31 articles to perform their meta-analysis; we will call these articles “relevant articles”. We performed a PubMed reference search with the search terms indicated in the meta-analysis that led to 552 articles. Among these articles, there were 30/31 relevant articles. The missing article was either found by the authors in another database or through another search not detailed in the article methodology. TEMAS (whose method applied to this performance test for meta-analysis is detailed in Additional file 2) led to a total of 145 articles among which we also found the 30 relevant articles. The recall rate for TEMAS with this meta-analysis was therefore 100%, while dividing the number of articles to be read by 3.8.

The second meta-analysis²⁶ was focused on age at menarche and at menopause as risk factors for breast cancer. The articles included in the meta-analysis were divided into 2 groups, cohort studies and case-control studies. There were 31 relevant articles available on PubMed. Among these 31 articles, 26 were found by the reference search; other articles came from other sources, as indicated in the methodology. The recall rate for TEMAS (we stopped TEMAS in step 3 because there was no condition on the presence of OR or RR in the meta-

analysis methodology) was 100% and divided the number of articles to read by 2.56 (5431 articles to read with TEMAS vs 13885 articles to read with a PubMed reference search).

For case-control studies, TEMAS recall rate was also 100%, all 52 relevant articles have been found; while dividing the number of articles to read by 2.92.

For each of the three meta-analyses tested, TEMAS recall rate was 100%. It therefore appears that TEMAS can help reducing the time required to search the literature for articles on PubMed, in order to perform a meta-analysis, by dividing between 2.5 to 3.8 times the number of articles to be read, while maintaining a recall rate of 100%.

DISCUSSION

Our Text-Mining Algorithm-assisted Search enabled making a global search on all the risk factors of a given disease by performing once a single classification. Then we could extract from this classification the articles concerning the risk factors of interest. For each risk factor studied the number of articles to read was divided by 3.2 on average (2.3 to 4.8) compared to RS. Thus, if we considered the five risk factors studied, RS lead to 2607 articles to read when it was 706 for TEMAS. Moreover, we improved the precision rate by 23% to 32%, according to the risk factor studied. Whatever the bibliographic search, TEMAS always improved the precision rate even if RS provided poor precision. TEMAS can be used to look for risk factors whatever the studied conditions, and appeared of interest for literature search for epidemiological studies. In this context we retained the most used metrics, namely precision and recall, average precision and Discounted Cumulative Gain to quantify the performance of TEMAS information retrieval.

The workload dedicated to reading all the articles selected by a classic search becomes so important that new approaches are needed. Several approaches are based on text-mining using machine learning methods⁷⁻⁹. However, they need to be trained on a carefully

constructed training data set representative of the topic of interest. Of note, it is necessary to train these algorithms for each targeted topic. Moreover, it requires specific skills to implement these algorithms. For instance, text mining tools have been implemented to extract information about microbial biodiversity in food¹⁰. However, this search cannot be extended to another subject without re-calibration. Conversely, TEMAS does not require a training set. It uses a hierarchical classification procedure based on co-occurrences on terms appearing in the abstracts. This framework enabled to develop different types of searches without a training step. Furthermore, this method can be implemented by non-specialists thanks to the R Shiny web app that helped providing a user-friendly interface available at <https://shiny.temas-bonnet.site>.

Our method has some limits. At the classification stage, the choice of the optimal number of classes might seem complex. Indeed, there are no prescriptive thresholding rules. In our classification procedures, we kept 5 classes for breast cancer since it enabled obtaining enough disjointed classes and an interesting clustering for our searches. The optimal number of classes is different from a search to another.

For multi terms searches, we selected a distance of 10 to 25 characters between the main and complementary terms depending on the risk factor studied and on the retained proximity between the retained terms. These thresholds had been tested and appeared satisfactory without omitting relevant articles on one side, and on the other side limiting irrelevant information.

Reducing the number of articles to read inevitably impacted the recall rate. However, the recall rate of TEMAS compared to RS remained satisfactory, ranging from 95% to 97% for the risk factors of breast cancer. It meant only one to three missing articles per search. In addition, we ranked the relevant articles according to the value of the OR (< 1 , $= 1$, or > 1) for the risk factor studied. We did not observe any significant difference in the distribution of ORs between RS and TEMAS relevant articles. Even if we did not get a recall rate of 100%,

the loss in recall rate did not bias the overall information given the above-mentioned distribution of the OR values.

We also checked whether TEMAS would achieve a 100% recall on articles selected in meta-analyses. We tested two meta-analyses, the recall rate was 100%. The loss of recall rate that we observed for the risk factors analyses might have targeted articles not relevant enough to be included in a meta-analysis. Of note, this PubMed-based search did not explore either other databases or other sources (grey literature, theses,...)

Future work will test whether our approach applies to other medical literature databases.

CONCLUSION

Faced with the growing number of articles indexed in PubMed, TEMAS offers an effective way to better target relevant articles in the literature. TEMAS was applied to etiological studies in epidemiology. It divided the number of articles to read by 3, while increasing search accuracy by at least 23% with a satisfactory recall rate. It also improved the average precision and the discounted cumulative gain compared to PubMed best match. Moreover, TEMAS appears as a potentially useful tool for meta-analytical search with a 100% recall rate. In addition TEMAS tool is accessible online by everyone thanks to the web application that we have developed.

LIST OF ABBREVIATIONS

- AP: Average Precision
- DCG: Discounted Cumulative Gain
- HR: Hazards Ratio
- MAP : Mean Average Precision
- OR: Odd Ratio
- RS : Reference Search
- TEMAS: Text-Mining Algorithm-Assisted Search

DECLARATIONS

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author (EB) on request.

Competing interests

The authors declare no competing interests

Funding

This work was funded by the University of Montpellier and Nîmes University Hospital which supported EB PhD thesis.

Authors' contributions

EB, PL and JPD conceived the project; EB designed the web application; EB collected the material; EB and PL read the selected articles; EB, PL and JPD participated in results interpretation, wrote and revised the manuscript.

Acknowledgements

We acknowledge Montpellier University and Nîmes University Hospital for enabling to develop this research.

REFERENCES

1. Fleuren WWM, Alkema W. Application of text mining in the biomedical domain.

- Methods*. 2015;74:97-106. doi:10.1016/j.ymeth.2015.01.015
2. Song M. TAKES: Two-step Approach for Knowledge Extraction in Biomedical Digital Libraries. *J Inf Sci Theory Pract*. 2014;2(1):6-21. doi:10.1633/jistap.2014.2.1.1
 3. Huang CC, Lu Z. Community challenges in biomedical text mining over 10 years: Success, failure and the future. *Brief Bioinform*. 2016;17(1):132-144. doi:10.1093/bib/bbv024
 4. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *J Clin Epidemiol*. 2009;62(10):1006-1012. doi:10.1016/j.jclinepi.2009.06.005
 5. Stroup DF, Berlin JA, Morton SC, et al. Meta-analysis Of Observational Studies in Epidemiology A Proposal for Reporting. *JAMA*. 2000;283:2831-2836. doi:10.1001/jama.283.15.2008
 6. Dekkers OM, Vandenbroucke JP, Cevallos M, Renehan AG, Altman DG, Egger M. COSMOS-E: Guidance on conducting systematic reviews and meta-analyses of observational studies of etiology. *PLoS Med*. 2019;16(2):e1002742. doi:10.1371/journal.pmed.1002742
 7. Simon C, Davidsen K, Hansen C, Seymour E, Barnkob MB, Olsen LR. BioReader: A text mining tool for performing classification of biomedical literature. *BMC Bioinformatics*. 2019;19. doi:10.1186/s12859-019-2607-x
 8. Ševa J, Wiegandt DL, Götze J, et al. VIST - a Variant-Information Search Tool for precision oncology. *BMC Bioinformatics*. 2019;20(1). doi:10.1186/s12859-019-2958-3
 9. Lin RTK, Dai HJ, Bow YY, Chiu JL Te, Tsai RTH. Using conditional random fields for result identification in biomedical abstracts. *Integr Comput Aided Eng*. 2009;16(4):339-352. doi:10.3233/ICA-2009-0321

10. Chaix E, Deléger L, Bossy R, Nédellec C. Text mining tools for extracting information about microbial biodiversity in food. *Food Microbiol.* 2019;81:63-75.
doi:10.1016/j.fm.2018.04.011
11. Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie JM. shiny: Web Application Framework for R. 2019.
12. R Core Team. R: A Language and Environment for Statistical Computing. 2019.
<https://www.r-project.org/>.
13. Higgins JPT, Lasserson T, Tovey D, Thomas J. Methodological Expectations of Cochrane Intervention Reviews (MECIR) Standards for the conduct and reporting of new Cochrane Intervention Reviews , reporting of protocols and the planning , conduct and reporting of updates Flemyng and Rachel Churchill V. *Cochrane Methods.* 2020;(March).
14. Winter DJ. rentrez: An R package for the NCBI eUtils API. *R J.* 2017;9(2):520-526.
doi:10.32614/rj-2017-058
15. Baumann N. How to use the medical subject headings (MeSH). *Int J Clin Pract.* 2016;70(2):171-174. doi:10.1111/ijcp.12767
16. Barnier J. rainette: The Reinert Method for Textual Data Clustering. 2020.
<https://cran.r-project.org/package=rainette>.
17. Ratinaud P, Marchand P. Application de la méthode ALCESTE aux “gros” corpus et stabilité des “mondes lexicaux”: analyse du “CableGate” avec IRAMUTEQ [Application of the ALCESTE method to “big” corpora and stability of “lexical worlds”: analysis of “CableGate” with IRAMUTEQ]. *Actes des 11èmes Journées Int d’Analyse des Données Textuelles.* 2012:835–844. <http://lexicometrica.univ-paris3.fr/jadt/jadt2012/Communications/Ratinaud, Pierre et al. - Application de la methode Alceste.pdf>. Accessed July 15, 2019.

18. Reinert M. Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte. *Cah l'analyse des données*. 1983;3(2):187-198.
http://www.numdam.org/article/CAD_1983__8_2_187_0.pdf. Accessed July 15, 2019.
19. Kagolovsky Y, Möhr JR. A new approach to the concept of “relevance” in information retrieval (IR). In: IOS Press, ed. *Studies in Health Technology and Informatics*. Vol 84. ; 2001:348-352. doi:10.3233/978-1-60750-928-8-348
20. Sampath Kumar B, Pavithra S. Evaluating the searching capabilities of search engines and metasearch engines: a comparative study. *Ann Libr Inf Stud*. 2010;57(2):87-97.
<https://pdfs.semanticscholar.org/cff6/e1fa3ab2b406b0b68bd9ee85ce26703cb020.pdf>. Accessed July 30, 2019.
21. Marchionini G. Evaluating digital libraries: A longitudinal and multifaceted view. *Libr Trends*. 2000;49(2):304-333. <https://core.ac.uk/download/pdf/4817672.pdf>. Accessed July 30, 2019.
22. Harter SP, Hert CA. Evaluation of Information Retrieval Systems: Approaches, Issues, and Methods. *Annu Rev Inf Sci Technol*. 1997;32:3-94.
<https://eric.ed.gov/?id=EJ565471>. Accessed July 30, 2019.
23. Dupret G, Piwowarski B. Model based comparison of discounted cumulative gain and average precision. In: *Journal of Discrete Algorithms*. Vol 18. Elsevier; 2013:49-62.
doi:10.1016/j.jda.2012.10.002
24. Fiorini N, Canese K, Starchenko G, et al. Best Match: New relevance search for PubMed. *PLoS Biol*. 2018;16(8). doi:10.1371/journal.pbio.2005343
25. Wu Y, Zhang D, Kang S. Physical activity and risk of breast cancer: A meta-analysis of prospective studies. *Breast Cancer Res Treat*. 2013;137(3):869-882.
doi:10.1007/s10549-012-2396-7

26. Hamajima N, Hirose K, Tajima K, et al. Menarche, menopause, and breast cancer risk: Individual participant meta-analysis, including 118 964 women with breast cancer from 117 epidemiological studies. *Lancet Oncol.* 2012;13(11):1141-1151.
doi:10.1016/S1470-2045(12)70425-4

LIST OF FIGURES

Figure 1: Flowchart describing the proceeding steps of TEMAS

Figure 2: TEMAS dendrogram for breast cancer risk factors

Figure 3: RS and TEMAS precision and recall rates for breastfeeding

Figure 4: RS and TEMAS precision and recall rates for menarche

Figure 5: RS and TEMAS precision and recall rates for oral contraceptive

Figure 6: RS and TEMAS precision and recall rates for mammographic density

TABLES

Table 1: retained articles for TEMAS and RS

Databases	n_{TEMAS}	n_{RS}
Risk factors		
Breastfeeding	142	328
Menarche	157	446
Oral contraceptive	112	306
Breast/mammographic density	174	836

Table 2: Recall and precision rates for the risk factors studied comparing TEMAS to the Reference Search

Risk factor of breast cancer	RS Precision*	TEMAS Precision*	TEMAS recall rate relative to RS
breastfeeding	23 %	51 %	97 %
menarche	13 %	36 %	95 %
oral contraceptive	20 %	52 %	95 %
breast density	8 %	33 %	95 %

*Relevance criterion: presence in the abstract of an odds ratio (OR), a hazards ratio (HR), or a relative risk (RR) for the risk factor studied.

Table 3: Comparison of the distribution of the ORs for relevant articles between TEMAS and RS.

Risk factor	Breastfeeding		Menarche		Oral Contraceptive		Mammographic density		
Databases	RS*	TEMAS	RS	TEMAS	RS	TEMAS	RS	TEMAS	
OR	<1	82%	82%	5%	5%	4.9%	5.3%	0%	0%
	=1	16%	17%	30%	29%	26%	25%	9%	10%
	>1	2%	2%	65%	66%	69%	70%	91%	90%
P-value	<i>p=1</i>		<i>p=1</i>		<i>p=1</i>		<i>p=1</i>		

*RS: Reference Search

Table 4: Average precision for TEMAS and PubMed Best Match.

Breast cancer risk factors	N	TEMAS (%)	PubMed Best Match (%)
Breastfeeding	25	73.3	47.7
	50	69.9	40.1
	75	66.6	36.9
	100	63.6	35.0

Age at menarche	25	67.8	25.0
	50	58.6	22.3
	75	54.0	22.6
	100	52.2	22.3
Oral contraceptive use	25	70.9	11.6
	50	59.5	15.9
	75	57.2	17.5
	100	56.1	18.7
Breast density	25	57.2	18.5
	50	50.9	18.0
	75	45.6	19.3
	100	42.8	19.3

AP (%) calculated for the first 25, 50, 75 and 100 articles read (N) of the breast cancer risk factors studied.

Table 5: Discounted Cumulative Gain for TEMAS and PubMed Best Match.

Breast cancer risk factors	N	TEMAS	PubMed Best Match
Breastfeeding	25	6.5	3.2
	50	8.5	4.4
	75	10.4	5.6
	100	12.4	6.8
Age at menarche	25	5.0	1.5
	50	7.1	2.6
	75	9.1	3.5
	100	10.7	4.4
Oral contraceptive use	25	5.0	0.8
	50	7.9	1.9
	75	10.3	3.0
	100	12.4	4.0
Breast density	25	3.6	1.5
	50	4.8	2.0
	75	6.5	3.2
	100	7.8	3.7

DCG calculated for the first 25, 50, 75 or 100 articles read (N) of the breast cancer risk factors studied

Figures

Figure 1

Flowchart describing the proceeding steps of TEMAS

Figure 2

TEMAS dendrogram for breast cancer risk factors

Figure 3

RS and TEMAS precision and recall rates for breastfeeding

Figure 4

RS and TEMAS precision and recall rates for menarche

Figure 5

RS and TEMAS precision and recall rates for oral contraceptive

Figure 6

RS and TEMAS precision and recall rates for mammographic density

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Additionalfile3.docx](#)
- [Additionalfile3.docx](#)
- [Additionalfile2.pdf](#)
- [Additionalfile2.pdf](#)
- [Additionalfile1.pdf](#)
- [Additionalfile1.pdf](#)