

e-scider: A tool to retrieve, prioritize and analyze the articles from PubMed database

Sujit R. Tangadpalliwar¹, Rakesh Nimbalkar², Prabha Garg*³

¹ National Institute of Pharmaceutical Education and Research (NIPER), S.A.S nagar, Punjab, 160062, India. Phone no: +91 9872501289

² National Institute of Pharmaceutical Education and Research (NIPER), S.A.S nagar, Punjab, 160062, India. Phone no: +91 9872501289

³ National Institute of Pharmaceutical Education and Research (NIPER), S.A.S nagar, Punjab, 160062, India. Phone no: +91 9872880963

Abstract:

Count of literature in the biomedical databases is growing at a double-exponential pace; manual extraction of such relevant literature in a structured format is a highly laborious and time-consuming. Therefore, there is a tenacious need to assist the retrieval of scientific information with automated text mining tools. Here, we engineered “e-scider” (e-scientific data fetcher), a text-mining application to obtain the most relevant literature from PubMed database. e-scider provides an interactive user-friendly platform that enables users to retrieve and analyze information in various customized ways. As like PubMed, it assists the retrieval of article information like title, author(s) name, and abstract with highlighted query word. One of the most distinctive features of e-scider is to allow field-wise extraction of articles pertaining to various scopes of journals. Illustrating the current scenario of the field, it provides graphical representations of publication count in each journal, year and country for the given query. It also prioritizes the retrieved articles based on relevancy scores and categorized them into most, moderate and less relevant articles. Moreover, it enables users to download multiple full-text articles in a single platform for the given query. Thus, e-scider is an easy to use tool that aid in literature survey for the pre-stage researcher. The entire information was made available online where the user can access the data and downloads the results for future offline reading.

Availability: e-Scider and its manual can be freely accessible at <http://14.139.57.41/e-scider/>

Keywords: PubMed, data mining, e-scider, literature, extraction, retrieval, biomedical, informatics

Corresponding Author: Prabha Garg

INTRODUCTION

In this era of big data, massive amount of data is generated every day. Similarly, in the case of biomedical research, plethora of information rich literature is being published. Currently, MEDLINE (PubMed) is one of the major resources for life sciences, specifically biomedical articles, which contain more than 25 million records from approximately 5,600 worldwide journals [1]. Finding, categorization and extraction of relevant articles from such a big text-ocean is cumbersome, tedious, time consuming and require lots of manual intervention. Thus, this exponentially growing unstructured bibliomic data leads to the challenge for extraction of useful information in reasonable time with minimal manual efforts. Text mining, also known as Intelligent Text Analysis, uses tools and techniques of data mining and machine learning methods for the retrieval of relevant information and to recognize pattern from unstructured data [2,3]. However, manual handling of text mining operations on such huge literature is also sometimes laborious and obscure in the real life scenario.

MEDLINE gives records as an output based on user provided keyword(s). Although, PubMed provides straight forward and fast results for a query text [4], extensive manual interference is required to find out and categorize relevant records of user interest from the obtained records. Till now, various solutions in the form of computational tools have been published that further modifies, classifies or extrapolate the results of PubMed output. Smalheiser *et al.*, [5] developed a tool named 'Anne O'Tate' that displays and thus highlights the most important words found in titles and/ or in abstracts in PubMed search results according to their predefined categories. PubMed Assistant, a tool developed by Jing Ding *et.al.* [6], allows boolean query editing and enhances the efficiency of PubMed search by keyword highlighting. It also provides options like export to citation managers, clickable links to Google Scholar *etc.* PubFocus develop by Plikus *et al.*, [7] prioritizes the articles based on algorithm that considers journal impact factor, authors' contribution level, and other factors. It also performs extraction of biomedical terminology based on predefined dictionary. However, to our knowledge, no tool exists to date that can retrieve the articles for a given query based upon the scope and/ or field *viz.* biology, chemistry *etc.* Also, no tool highlights the current scenario of the topic by providing the pictorial depiction of trend of publications for the given keyword in each journal, year and country. Therefore, a new computational tool is needed to augment the efficacy for retrieval of relevant literature information and to provide a current trend of the research field.

Considering the fact that effective literature search is a founding stone of any research field, here we present 'e-Scider', a searching tool for the effective and customized retrieval of the biomedical research information. e-Scider is an easy to use, user-friendly interface to MEDLINE, developed using client side and server side languages like: HTML, PHP, CSS, JavaScript. It sends query to MEDLINE via NCBI Entrez Utilities ESearch service employing python script. All articles related information like title, authors name, affiliation, abstract, source, publication date, type of article (research/review), PubMed central identifier *etc.* can be retrieved efficiently in structured manner and export to excel sheet for further offline reading. Alongside displaying output in the form of records similar to PubMed, e-Scider comprises of additional salient characteristics like (i) It gives publication trend throughout years with graphical representation (ii) It gives the opportunity for retrieval of articles according to selected research field (iii) It facilitates the retrieval of freely available articles in pdf format in single click for the given keyword (iv) Based upon the relevance score, it classifies output into three categories *i.e.* most, moderate and least relevant (v) User has option for retrieval of most cited or recent articles

(vi) It provides other customized options like retrieval of articles in between years and from specific journals.

INFORMATION EXTRACTION

e-Scider runs on Python scripts at the backend for undertaking multiple tasks. User enters query to the e-scider interface with the help of python scripts it forwards this request to the MEDLINE server, retrieves the results *via* EFetch service and process them according to the user requirement again using scripts.

ARCHITECTURE AND IMPLEMENTATION

Graphical user interface of e-scider provides four different browsing options (Fig. 1).



Fig 1: Browse page Screen

1.1. The first part *i.e.* upper horizontal input field of e-scider allows for the entry of keyword(s) to PubMed and gives number of records for given query on 'search' button click. Additionally, it provides the statistics demonstrating publication trends throughout the journals, years and countries with graphical representation for the given query on 'evaluate' button click. For example, for the keyword 'zika virus' e-scider gives publication scenario with respect to journal, years and countries. By analyzing figure 2a it is found that highest articles *i.e.* 51 out of 1131 for 'zika virus' are published in the LANCET journal subsequently to the BMJ with 45 hits. This graphical information helps researcher to find journals that accepts or focus work related to 'zika virus' for publication. Similarly, in the case of fig 2b

and 2c most of the articles for 'zika virus' are publish in the year 2016 with 1007 publications and in United States with 407 publications respectively. This information depicts the publication pattern and thus provides the research scenario for 'zika virus' with respect to year and country.

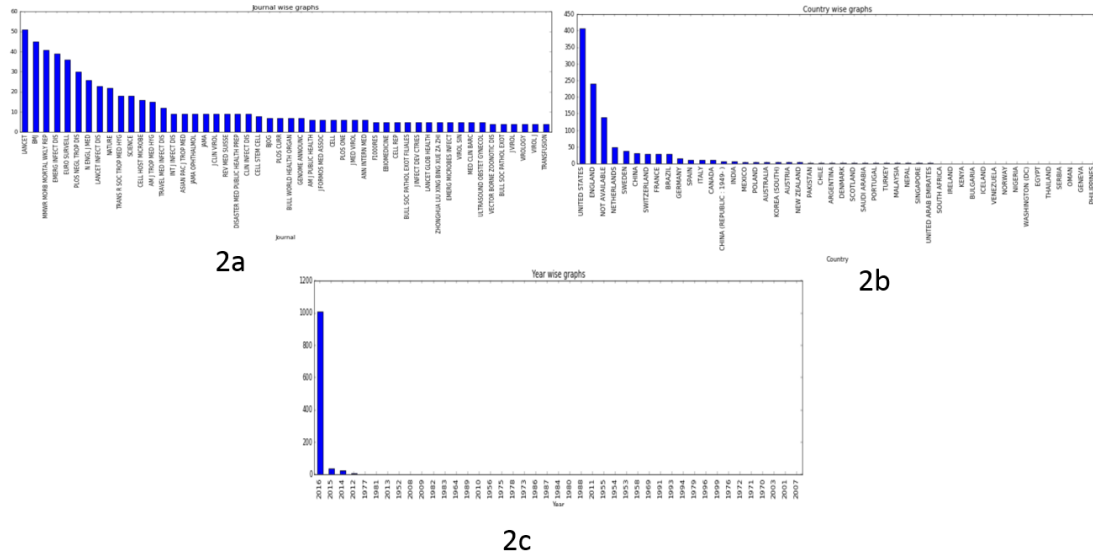


Fig 2: Graphical output of e-Scider for the 'zika virus' keyword

1.2. The second browsing option, named as 'Algorithm panel', needs two separate text files containing keywords against which articles are required to fetch as inputs. It implements data mining algorithm in order to prioritize the retrieved articles in most, moderate and less read categories (Fig 3 explains the algorithm flow).

Explanation of algorithm steps with example:

Input files: File_1 contains broader category keywords like cancer, liver, malaria, skin *etc.* and File_2 contains keyword like carcinoma, apoptosis, hepatocytes *etc.*

- i) e-Scider picks first keyword (*eg.* Cancer) from file_1 and retrieves all articles which contain cancer anywhere in the article.
- ii) Then it picks first, second and so on... (carcinoma, apoptosis, hepatocytes) keywords one by one from file_2 and search in title and abstract portion of articles retrieved in step i), if found, then prints all information about this article.
- iii) After that, it picks next keyword (*eg.* Liver) from file_1 and again do the same process mentioned in step ii.
- iv) Step ii and iii repeats till all the keywords used.
- v) Finally, it assigns the relevancy score according to the occurrence of the keyword in title and/or abstract. Thus, based on cutoff, highest score articles are marked as most relevant articles, while least score articles are assigned as less relevant articles and remaining as moderate relevant articles.

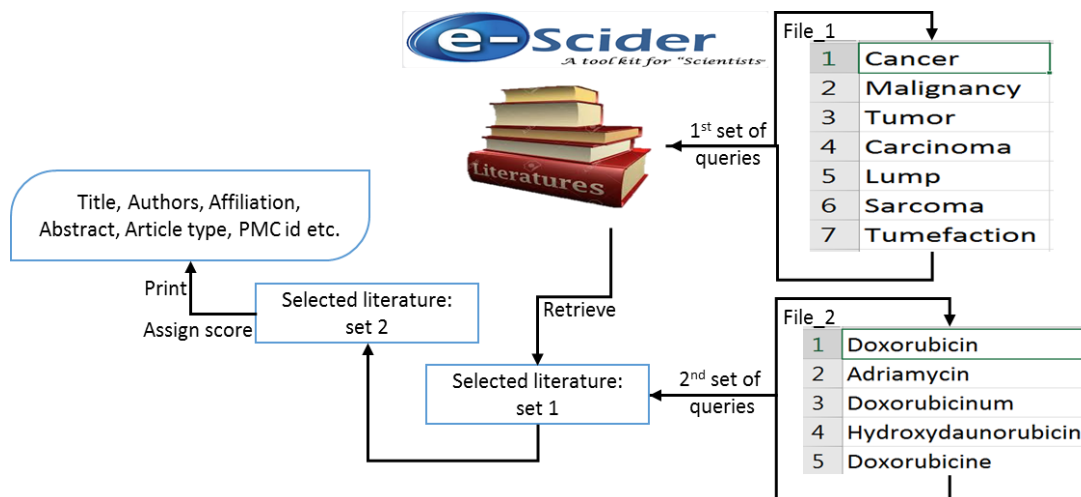



Fig 3: Figure showing flow of algorithm

1.3. The third panel named 'Paid articles' gives the result exactly similar to the PubMed output with extra feature enabling users to export result in excel sheet for offline reading. Unlike PubMed it provides 3 additional options to user for the retrieval of articles *viz.* a) Category wise: Output produced by PubMed needs manual intervention to categorize them into various research fields. One of the unique features of e-scider is fetching of articles according to user selected/defined category. For this purpose, user has to select field(s) of study from the given list and enter query word; system retrieves records for the same. For example, if user wants to retrieve only chemistry and/or computational chemistry field articles for the keyword 'cancer', it retrieves all the articles only related to user selected field. b) Journal and Year wise: Retrieves the articles from given journal for the year for the user-defined keyword. c) In between years: Used, when, focusing search on the subset of articles published from 1990 to 1994 (when the field was beginning to expand). Fig. 4 shows result page.

1.4. The fourth option in browsing page is 'Freely available articles panel'. It is used to fetch multiple full-text articles in pdf format for the user given keyword in a single click downloadable in zip folder. The search can be further customized with the options like; 1) date & relevance wise: where user has to enter keyword, number of articles to be required and select retrieval method date wise or relevance wise, 2) In between years: helps to retrieve articles publish in between years (e.g. from 2012 to 2016)



[Home](#)
[Browse](#)
[Developed By](#)
[Contact Us](#)

Title: Exon sequencing and association analysis of EPHX1 genetic variants with maintenance warfarin dose in a multiethnic Asian population.
Authors: ['Chan, Sze Ling', 'Thalamuthu, Anbupalam', 'Goh, Boon Cher', 'Chia, Kee Seng', 'Chuah, Benjamin', 'Wong, Andrea', 'Lee, Soo Chin']
Affiliation: Department of Epidemiology and Public Health, National University of Singapore, Singapore.
Abstract: BACKGROUND AND OBJECTIVES: Warfarin inhibits vitamin K epoxide reductase, of which microsomal epoxide hydrolase is a putative member. Several studies have found signals of association with warfarin maintenance dose in the EPHX1 gene. The aim of this study was to determine the effects of EPHX1 variants on warfarin maintenance dose in a multiethnic Asian population. METHODS: We sequenced the exons of EPHX1 using PCR and direct sequencing in 279 patients consisting of three major ethnic groups receiving maintenance warfarin with a stable international normalized ratio. The effects of EPHX1 variants were assessed using multiple linear regression. RESULTS: An association between an intronic SNP rs1877724 and warfarin maintenance dose was found, with homozygous variant carriers requiring approximately 0.5 mg/day lower than wild type and heterozygotes after adjustment for covariates. However, its contribution is small, explaining only an additional 0.8% of the dose variability. Rare variants were pooled but there was no association between their presence and warfarin maintenance dose. However, the presence of noncoding rare SNPs was significantly associated with warfarin maintenance dose. CONCLUSION: Despite a significant finding in rs1877724, which concurs with an earlier study, overall, genetic variants in EPHX1 do not have a clinically significant impact on warfarin dose requirements in our population.
Source: Pharmacogenetics and genomics
Journal Identifier: 101231005
Article Identifier: Not Available
Publication History: Not Available
Publication Type: ['Journal Article', 'Research Support, Non-U.S. Gov*']

[Evaluate](#)
[Export Results](#)

Fig 4: Result panel for the given query.

APPLICATIONS

e-scider provides several extraction options. It gives the list of journals with number of publications in it for user given keyword(s). It helps user to analyze trend of particular research topic in the past. Assist pre-doctoral and pre-stage researcher while choosing topic according to trend. Allows download of few thousands of free-text pdf articles on single click.

CONCLUSION

e-scider is easy-to-use application, mostly useful for researchers to find publications trend of their topic. As it is web application; thus, no restriction for installation on various platform as it runs using browser. It allows retrieval of articles according to the journal scope/category etc. It offers one-stop facility for literature retrieval, prioritization, categorization and exportation into local computer.

REFERENCES

1. https://www.nlm.nih.gov/pubs/factsheets/dif_med_pub.html
2. Cano, C, Monaghan, T, Blanco, A, Wall, D.P. Peshkin, L, Collaborative text annotation resource for disease centered relation extraction from biomedical text (2009) Journal of Biomedical Informatics, Vol. 42, pp.966-977.
3. Landge MA, Rajeswari K., A Survey on Chemical Text Mining Techniques for Identifying Relationship Network between Drug Disease Genes and Molecules (2016) International Journal of Computer Applications, Vol. 146, pp.5-9.

4. Frisch,M, Klocke,B, Haltmeier,M, Frech K, LitInspector: literature and signal transduction pathway mining in PubMed abstracts (2009) *Nucleic Acids Research*, Vol. 37, pp.135–140.
5. Smalheiser NR, Zhou W, Torvik VI, Anne O'Tate: A tool to support user-driven summarization drill-down and browsing of PubMed search results (2008) *Journal of Biomedical Discovery and Collaboration*, Vol. 3, pp.2
6. Ding J, Hughes LM, Berleant D, Fulmer AW, Wurtele ES, PubMed Assistant: a biologist-friendly interface for enhanced PubMed search (2006) *Bioinformatics* Vol. 22, pp.378-80
7. Plikus MV, Zhang Z, Chuong C, PubFocus: semantic MEDLINE/PubMed citations analytics through integration of controlled biomedical dictionaries and ranking algorithm (2006) *BMC Bioinformatics*, Vol. 7, pp.424