TECHNICAL REPORT 10 March 2021
DOI: TODO

# A Simple Machine Learning Framework to Aid Citation Screening in Systematic Reviews and Meta-Analyses of Aging and Longevity Research Studies

Marko Lalović*

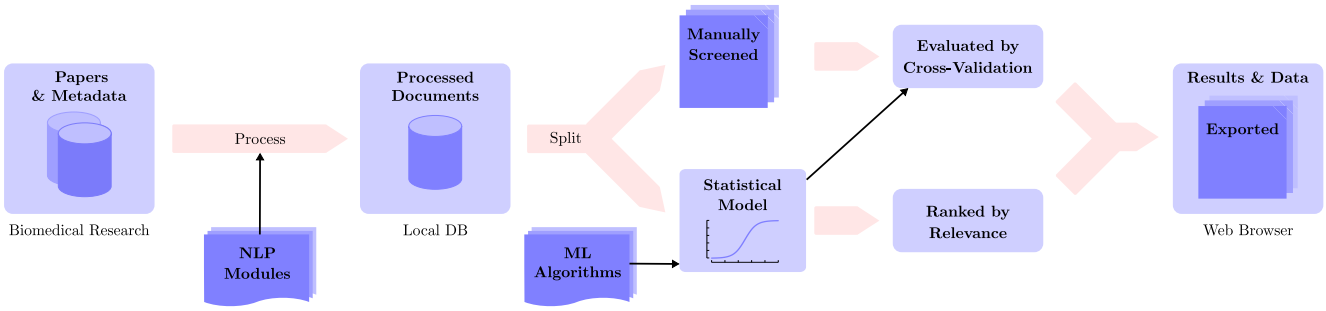



Figure 1: General overview of proposed framework.

## Abstract

*This report describes the initial results of a research project to develop a machine learning framework to semi-automate citation screening in systematic reviews and meta-analyses. It was developed and evaluated in context of aging and longevity research studies. We conducted experiments on dataset related to a specific risk-benefit analysis. Empirical results show that by using the proposed system, reviewers save an average of around 60% of screening work compared to unaided screening to identify 95% of relevant papers. The logistic regression model used in our framework is one of the simplest machine learning models and very efficient to train. The resulting model is used for the update of the analysis, thereby offering an opportunity to reduce the reviewers workload. In addition, we use the results from cross-validation procedure to help identify potential screening errors made. We outline the framework and how it can be used in similar screening practices. The software is open sourced with dataset freely available*[1].

***Author Keywords:*** Automated screening, Literature prioritization, Machine learning, Text mining, Software



*mailto:marko@markolalovic.com
[1] https://github.com/markolalovic/longevity-research-screening/
[2] https://creativecommons.org/licenses/by/4.0/

## 1. Introduction

A systematic review typically addresses a specific clinical question by collecting and analyzing data from all the relevant and unbiased set of studies. Citation screening is the first yet tedious task of narrowing down the large set of citations retrieved via a broad database query to those relevant for the review. Machine learning (ML) refers to algorithms that can learn from and make predictions on data by building statistical models.

**Related Work.** The work in (semi)-automation of citation screening is active and diverse. Wallace et al. [2] developed a semi-automated citation screening algorithm for systematic reviews of biomedical literature. Bannach-Brown et al. [3] described their approaches to aid citation screening for a systematic review of preclinical animal studies. Howard et al. [4] deployed a general software system that automate the required methodologies called "SWIFT-Review". Przybyła et.al [5] introduced a web-based software system called "RobotAnalyst". O'Mara-Eves et al. [1] performed a systematic review of current approaches. They concluded, on one side, that the use of these tools to automatically eliminate studies is promising but should be used cautiously, since the reviews are at risk of limiting their review to such a degree that the validity of their findings is questionable. On the other side, using them in order to prioritize the order in which papers are screened should be considered safe and ready for use in actual reviews. However, the use of these tools varies greatly across disciplines. To date, no use of any tools related to automating (or semi-automating) the screening process of systematic reviews or

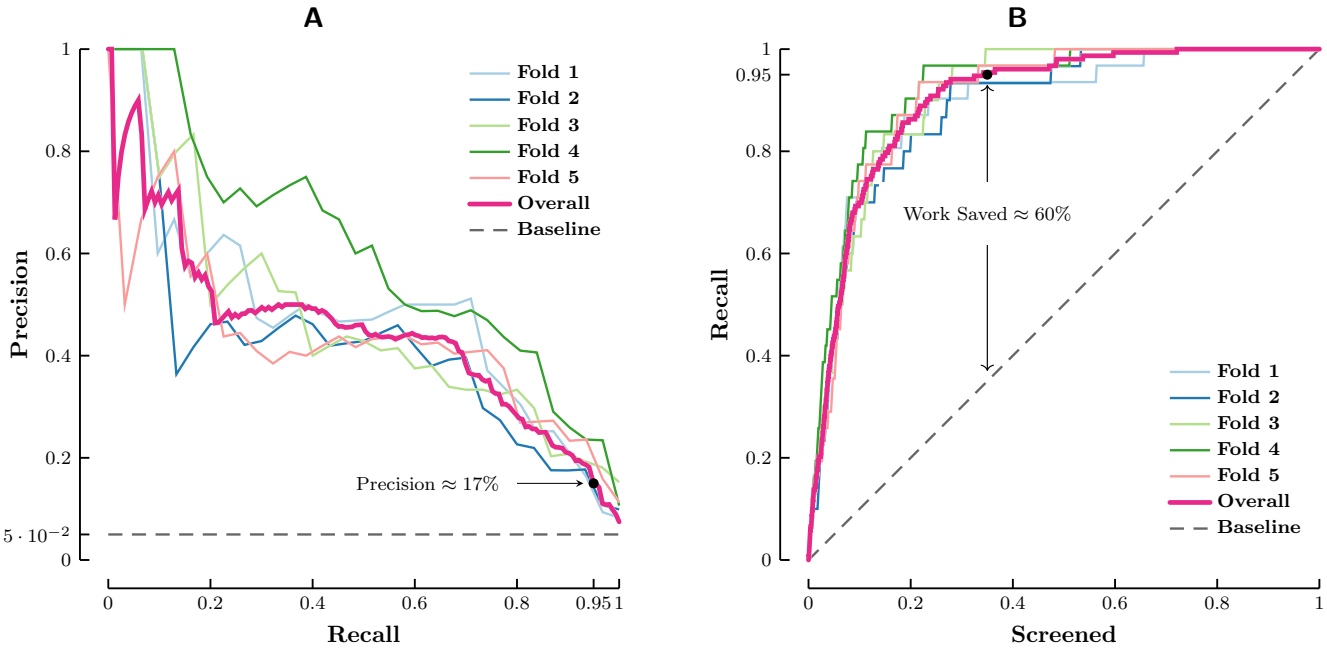



Figure 2: Visualized results of 5-fold cross-validation for D&Q Analysis.

meta-analyses of aging and longevity research was reported. More developmental work is needed and more validation of these methods needs to be performed before all systematic review teams can take advantage of these tools.

The aim of this work is to develop an easy to use tool that can be used in the screening stage of systematic reviews or meta-analyses of aging and longevity research studies to reduce the reviewers workload (and screening error). Here, we present a simple ML framework that enables us to connect the data and models, evaluate the performance and deploy and interpret the results in the form of interactive tables accessible from a browser.

**General Overview**. Query databases of biomedical research publications using a provided list of search terms. Process documents (consisting of metadata, title and abstract) using natural language processing tools and save the results into a local database. Split the documents in the following two sets. First set consists of *labeled documents*, meaning that each document is manually screened and assigned a label (1 = relevant, 0 = not relevant) based on reviewers decision. The set of labeled documents is used to train a statistical model. The fitted model estimates the probability called *relevance score*, that a given document is relevant. Apply the model to the second set of unlabeled documents to estimate their relevance scores and rank them according to their estimated relevance scores. In addition, using the cross-validation procedure, assign relevance scores to all labeled documents to help identify any false labeled documents. Export the results in the form of interactive tables to a web application. These steps are visualized in Figure 1.

We conducted experiments on dataset related to Dasatinib and Quercetin Senolytic Therapy Risk-Benefit Analysis (D&Q Analysis). The analysis is part of "Rejuvenation Now" non-profit initiative that: "seeks to continuously identify potential rejuvenation therapies and systematically evaluate their risks, benefits, and associated therapeutic protocols to create transparency" published by Forever Healthy Foundation[3]. Here, the split was based on publication date of review. We used labeled documents from existing systematic review. The resulting model was then applied to new documents retrieved in subsequent search to be used for an update of the original review. In our approach we focused on maximizing the recall. *Recall* (also known as *sensitivity*) is the fraction of relevant documents that are also retrieved as relevant by the system. For a formal definition of recall, other key measures and curves, see Section 2.5.

**Results.** The following values are estimates based on 153 labeled documents using 5-fold cross-validation procedure. For more details, see Section 2.6. The empirical results show that the proposed system can identify 95% of relevant documents with precision 17%; see Figure 2A. This means that only around 17% of identified documents are actually relevant. The performance of the system is still reasonable, since the reviewers have to screen only around 35% of retrieved documents on average. This saves them on average

[3] https://brain.forever-healthy.org/display/EN/

| Topic | Terms |
|---|---|
| 1. | trial, clinic, efficaci, report, safeti, assess, evalu, show, advers, ... |
| 2. | week, placebo, treatment, group, extract, symptom, score, patient, hypericum, ... |
| 3. | dasatinib, patient, case, treatment, report, chronic, leukemia, therapi, myeloid, ... |
| 4. | cvd, risk, cardiovascular, prevent, factor, profil, diseas, lipid, import, ... |
| 5. | flavonoid, anthocyanin, individu, adult, genistein, dietari, isoflavon, flavanon, intak, ... |
| 6. | sunitinib, sorafenib, target, imatinib, includ, cancer, erlotinib, anticanc, malign, ... |
| 7. | bone, marrow, chromosom, abnorm, deriv, prognost, signific, aberr, delet, ... |
| 8. | lifespan, elegan, longev, stress, effect, life, span, extend, increas, ... |
| 9. | muscl, smooth, vsmc, skelet, aortic, havsmc, resist, accordingli, folfiri, ... |
| 10. | quercetin, effect, cell, concentr, human, dose, depend, studi, increas, ... |

Table 1: Terms of the first 10 extracted topics.

around 60% of work comparing to screening documents in random order where they would need to screen 95% of documents on average to achieve the desired 95% recall; see Figure 2B.

## 2. Methods

We describe the methodology and how it was applied to Dasatinib and Quercetin Senolytic Therapy Risk-Benefit Analysis (D&Q Analysis).

### 2.1. Data

We first query the publicly available PubMed (MEDLINE) database using pymed API. This can be easily expanded to other sources, for example to Cochrane Library. The search terms for D&Q Analysis were devised by Forever Healthy foundation. Additionally, we scrape some data directly from websites of journals or clinical trials (ClinicalTrials.gov) using chromedriver in Python. The retrieved data is saved into local MySQL database. For each article we save its url (where it was retrieved from), publication date, publication types, title and abstract. For papers that were manually screened, we also save the assigned label (1 = relevant, 0 = not relevant).

We retrieved 2837 potentially relevant papers for D&Q Analysis. Of these, only 153 papers were labeled as relevant. The prevalence of relevant class in this dataset is approximately 5%. This means that this dataset is highly imbalanced and requires more complex statistical modeling and evaluation.

### 2.2. Pre-Processing

For the purposes of statistical modeling, the retrieved data is processed in the following way. The words in title and abstract are filtered based on part-of-speech tags to remove low information words such as english stop words. The remaining words are then reduced to their base forms using lematization to improve the statistical modeling performance. Both of this text processing steps are done using spaCy Python library. Lastly, we append the prefix `title` to words in title to separate them from words in abstract. Combined title, abstract and publication types of each paper is called *document*.

### 2.3. Feature Extraction

Each document $d$, is represented by a vector $X^d$. Components of $X^d$ are called *features* and denoted by $X^d_*$ where * stands for the name of the feature. First, two sets of simple binary features are created. First set of binary features is based on the provided list of search terms to construct a query (e.g.: `dasatinib`, `senolytic`, `senescent`, ...). For each search term $st$, the value of $X^d_{st}$ is 1, if $st$ appears in the document $d$, otherwise the value is 0. Second set of binary features is based on possible publication types (e.g.: `case report`, `clinical trial`, `review`, ...). For each publication type $pt$, the value of $X^d_{pt}$ is 1, if $pt$ appears in the document $d$, otherwise the value is 0.

The next set of features is based on bag-of-words representation of $n$-*grams*. These are contiguous sequences of $n$ words called *terms* (e.g.: `chronic myeloid`, `adverse event`, `tyrosine kinase`, ...). Here, we allow for 2-grams (pairs of words). From this representation, Term Frequency - Inverse Document Frequency (TF-IDF) matrix is constructed. The matrix consists of term scores. The terms which help to distinguish between documents have higher scores. For each term $t$, the value of $X^d_t$ is TF-IDF score of the term $t$ in document $d$:

$$X^d_t = \text{TF-IDF}_{t,d}.$$

For the last set of features, the latent Dirichlet allocation (LDA) topic model is constructed using MALLET library. See Table 1 for examples of extracted topics. For each topic $u$, the value of $X^d_u$ is the probability that document $d$ belongs to the topic $u$:

$$X^d_u = \Pr(\text{document } d \text{ belongs to the topic } u).$$

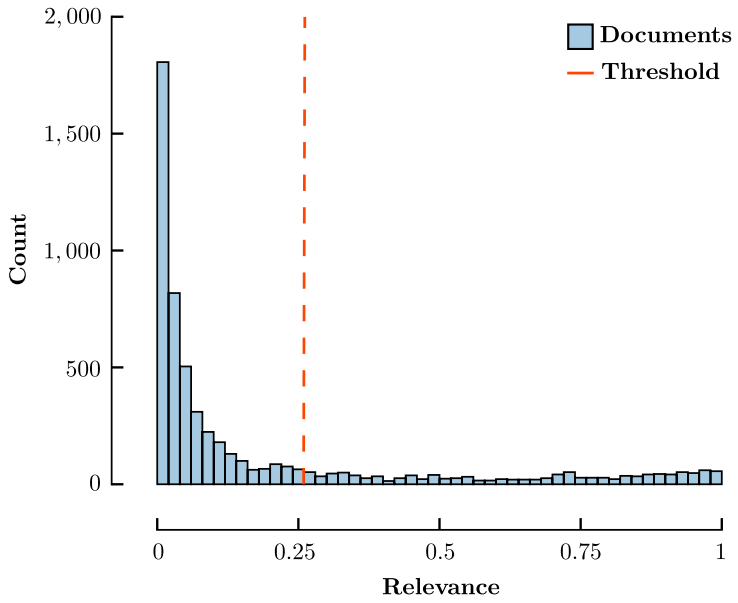



Figure 3: The selected cut-off threshold for D&Q Analysis was 0.26 where the binary classifier achieved 95% recall.

Finally, the features are standardized to have zero mean and unit variance:

$$X^d := \frac{X^d - \overline{X^d}}{\sigma_{X^d}}.$$

## 2.4. Model

Given the labeled documents, we build a logistic regression model from scikit-learn library in Python. The logistic regression model used in our framework is one of the simplest machine learning models and very efficient to train. The model estimates the conditional probability, called *relevance score*, that a given document $d$ is relevant given feature vector $X^d$:

$$\Pr(d \text{ is relevant} | X^d)$$

To avoid over-fitting, we use L1-regularization parameter $C$. For tuning parameter $C$ we place more emphasis on recall by using recall scorer. Model is fitted using Liblinear solver with balanced class weights to improve the recall. For D&Q Analysis the value of L1-regularization parameter $C$ was selected to be around $0.02$. The support, or the number of features with non-zero weights in the model, was around 28. These are average values across the folds when using cross-validation procedure explained next.

## 2.5. Performance Evaluation

Evaluation is done using the standard testing procedure called *k-fold cross-validation* for $k = 5$. This involves randomly partitioning the set of documents into 5 complementary subsets called *folds* of nearly equal size. Stratified sampling is used for partitioning to ensure that the folds have similar proportion of relevant documents. Then over 5 iterations, the model is fitted on 4 folds, called *train set*, and validated on the remaining fold, called *test set*. Each fold is used once as a test set. Thus, every document is part of the test set once and is assigned the relevance score. The distribution of relevance scores for all the documents are shown in Figure 3.

These relevance scores are then compared to human assigned labels in the following way. Documents with relevance score higher than selected cut-off threshold are classified as relevant. All other documents are classified as not relevant. The result is a binary classifier. For D&Q Analysis the selected threshold was 0.26 where the binary classifier achieved 95% recall. For a particular document, the four possible outcomes of comparing the classification result with human assigned label are defined in the following table:

| | | **Classified as relevant** | |
|---|---|---|---|
| | | **Yes** | **No** |
| **Labeled as relevant** | **Yes** | *True positive* | *False negative* |
| | **No** | *False positive* | *True negative* |

| Fold | Recall | Precision | PR-AUC | WSS@R |
|---|---|---|---|---|
| 1 | 0.94 | 0.13 | 0.54 | 0.53 |
| 2 | 0.90 | 0.16 | 0.33 | 0.61 |
| 3 | 1.00 | 0.14 | 0.48 | 0.63 |
| 4 | 0.94 | 0.19 | 0.54 | 0.67 |
| 5 | 0.97 | 0.20 | 0.43 | 0.71 |
| Mean | 0.95 (0.03) | 0.17 (0.03) | 0.46 (0.08) | 0.63 (0.06) |

Table 2: Summarized results of 5-fold cross-validation for D&Q Analysis.

The number of all true positives out of all the documents is denoted by $TP$. Similarly for other outcomes, the numbers of outcomes out of all the documents are denoted by $FN, FP$ and $TN$. Performance was assessed using four statistical measures of binary classifier performance. *Precision* $P$ is the fraction of documents labeled as relevant among documents classified as relevant:

$$P = \frac{TP}{TP + FP}$$

*Recall* $R$ (also known as *sensitivity*) is the fraction of documents labeled as relevant that were also classified as relevant:

$$R = \frac{TP}{TP + FN}$$

There is an inverse relationship between precision and recall. Depending on the selected threshold, it is possible to increase one at the cost of reducing the other. For example, by selecting the threshold of 0, we classify all documents as relevant and achieve maximum recall of 1 at the cost of precision, which is the proportion of documents labeled as relevant among all the documents (which is approximately 0.05 in our case). *Precision-recall curve* illustrates this trade off by plotting precision and recall as a function of selected cut-off threshold. *PR-AUC* is the area under the precision-recall curve. It is a performance measure used to compare classifiers or estimate the average precision of a classifier.

We can also order the documents based on their relevance scores and screen the documents in this order. We can visualize the ranking performance by plotting recall as a function of proportion of documents screened. *Work Saved over Sampling* $WSS@R$ is the reduction of documents that need to be screened compared to a random ordering of the documents to achieve a level of recall $R$:

$$WSS@R = \frac{TN + FN}{N} - (1 - R)$$

Here $N$ denotes the number of all documents:

$$N = TP + FN + FP + TN$$

## 2.6. Evaluation results

The results for each fold of 5-fold cross-validation for D&Q Analysis are summarized in Table 2 and visualized in Figure 2. For selected cut-off threshold of 0.26 the estimated average performance across all 5 folds (mean $\pm$ sd) was:

$$P = 0.17 \pm 0.03$$
$$R = 0.95 \pm 0.03$$
$$WSS@R = 0.63 \pm 0.06$$

We get very similar estimates from combined confusion matrix across all folds since we are using stratified sampling. We can interpret this results in the following way. In a set of new unlabeled documents that are assigned the relevance score of more than 0.26, we can expect the following:

- around 17% of documents in this set are relevant;
- around 95% of all the relevant documents are in this set;
- around 60% of screening work is saved if screening only this set comparing to random order screening.

## 2.7. Export

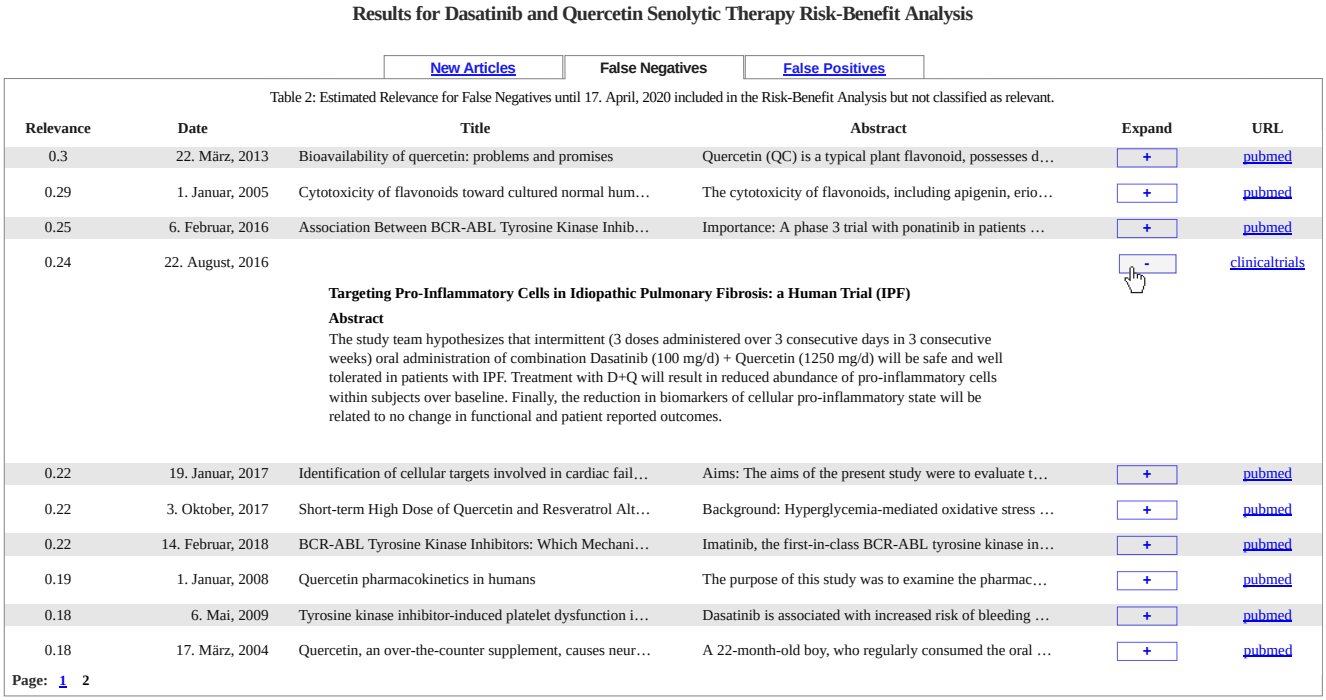
We exported 267 retrieved documents from new papers published after the publication of D&Q Analysis (April 17, 2020) sorted by relevance scores that were assigned by the final model trained on all the labeled documents to be used for the update of analysis. In addition, we exported all false negatives and false positives from labeled documents sorted by assigned relevance scores to identify potential screening errors made in the original review.

The result are 3 interactive tables of exported documents, shown in Figure 4 and accessible online[4]. The user can expand each document to see the title and abstract of the corresponding paper or follow the provided url.

## 2.8. Conclusion

In this work, we have introduced a simple machine learning framework that can be used in the screening stage of systematic reviews or meta-analyses. We applied it to a

[4] https://markolalovic.com/longevity-research-screening/



**Results for Dasatinib and Quercetin Senolytic Therapy Risk-Benefit Analysis**

New Articles | False Negatives | False Positives

Table 2: Estimated Relevance for False Negatives until 17. April, 2020 included in the Risk-Benefit Analysis but not classified as relevant.

| Relevance | Date | Title | Abstract | Expand | URL |
|---|---|---|---|---|---|
| 0.3 | 22. März, 2013 | Bioavailability of quercetin: problems and promises | Quercetin (QC) is a typical plant flavonoid, possesses d… | + | pubmed |
| 0.29 | 1. Januar, 2005 | Cytotoxicity of flavonoids toward cultured normal hum… | The cytotoxicity of flavonoids, including apigenin, erio… | + | pubmed |
| 0.25 | 6. Februar, 2016 | Association Between BCR-ABL Tyrosine Kinase Inhib… | Importance: A phase 3 trial with ponatinib in patients … | + | pubmed |
| 0.24 | 22. August, 2016 | | | - | clinicaltrials |

**Targeting Pro-Inflammatory Cells in Idiopathic Pulmonary Fibrosis: a Human Trial (IPF)**

**Abstract**

The study team hypothesizes that intermittent (3 doses administered over 3 consecutive days in 3 consecutive weeks) oral administration of combination Dasatinib (100 mg/d) + Quercetin (1250 mg/d) will be safe and well tolerated in patients with IPF. Treatment with D+Q will result in reduced abundance of pro-inflammatory cells within subjects over baseline. Finally, the reduction in biomarkers of cellular pro-inflammatory state will be related to no change in functional and patient reported outcomes.

| | | | | | |
|---|---|---|---|---|---|
| 0.22 | 19. Januar, 2017 | Identification of cellular targets involved in cardiac fail… | Aims: The aims of the present study were to evaluate t… | + | pubmed |
| 0.22 | 3. Oktober, 2017 | Short-term High Dose of Quercetin and Resveratrol Alt… | Background: Hyperglycemia-mediated oxidative stress … | + | pubmed |
| 0.22 | 14. Februar, 2018 | BCR-ABL Tyrosine Kinase Inhibitors: Which Mechani… | Imatinib, the first-in-class BCR-ABL tyrosine kinase in… | + | pubmed |
| 0.19 | 1. Januar, 2008 | Quercetin pharmacokinetics in humans | The purpose of this study was to examine the pharmac… | + | pubmed |
| 0.18 | 6. Mai, 2009 | Tyrosine kinase inhibitor-induced platelet dysfunction i… | Dasatinib is associated with increased risk of bleeding … | + | pubmed |
| 0.18 | 17. März, 2004 | Quercetin, an over-the-counter supplement, causes neur… | A 22-month-old boy, who regularly consumed the oral … | + | pubmed |

**Page:** 1 2

**Last updated:** Do 25. Feb 11:38:01 CET 2021

Figure 4: Interactive tables of exported documents for D&Q Analysis.

particular dataset related to a specific risk-benefit analysis published by Forever Healthy Foundation. The empirical results show that by using the proposed framework, we are able to identify 95% of relevant documents from a set of all documents retrieved via a broad database query. We estimated that reviewers can save an average of around 60% of screening work comparing to unaided screening to achieve this result. The proposed system can already be used to prioritize the order in which papers are screened. Or as a "second screener" based on exported documents that the system falsely classified as relevant, to help identify potential screening errors made.

More development is needed and validation on different datasets needs to be performed before using this framework to automatically remove studies for a review. The features should reflect what reviewers are looking for in a specific systematic review when deciding which paper is relevant. For example, they may be looking for effect size or try to detect large deviations in effect size. More collaboration with systematic review teams is needed to create such features. In the future, we would like to extend this work and evaluate it on different datasets.

## References


[1] A. O'Mara-Eves, J. Thomas, J. McNaught, M. Miwa, and S. Ananiadou, "Using text mining for study identification in systematic reviews: a systematic review of current approaches." https://doi.org/10.1186/2046-4053-4-5, 2015.

[2] B. C. Wallace, T. A. Trikalinos, J. Lau, C. Brodley, and C. H. Schmid, "Semi-automated screening of biomedical citations for systematic reviews." https://doi.org/10.1186/1471-2105-11-55, 2010.

[3] A. Bannach-Brown, P. Przybyła, J. Thomas, A. S. C. Rice, S. Ananiadou, J. Liao, and M. R. Macleod, "Machine learning algorithms for systematic review: reducing workload in a preclinical review of animal studies and reducing human screening error." https://doi.org/10.1186/s13643-019-0942-7, 2019.

[4] B. E. Howard, J. Phillips, K. Miller, A. Tandon, D. Mav, M. R. Shah, S. Holmgren, K. E. Pelch, V. Walker, A. A. Rooney, M. Macleod, R. R. Shah, and K. Thayer, "Swift-review: a text-mining workbench for systematic review." https://doi.org/10.1186/s13643-016-0263-z, 2016.

[5] P. Przybyła, A. J. Brockmeier, G. Kontonatsios, M. Pogam, J. McNaught, E. Elm, K. Nolan, and S. Ananiadou, "Prioritising references for systematic reviews with robotanalyst: A user study." https://doi.org/10.1002/jrsm.1311, 2018.