Deep Learning Method to Automatically Identify Reports of Scientifically Rigorous Clinical Research from the Biomedical Literature

*Guilherme Del Fiol¹, *Matthew Michelson², Alfonso Iorio³,⁴, Chris Cotoi⁵, R Brian Haynes³,⁴

* GDF and MM are first co-authors.
¹Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, USA.
²Evid Science, Los Angeles, California, USA.
³Department of Health Research Methods, Evidence, and Impact, McMaster University, Hamilton, ON, Canada.
⁴Department of Medicine, Faculty of Health Sciences, McMaster University, Hamilton, ON, Canada.
⁵Health Information Research Unit, McMaster University, Hamilton, ON, Canada.

Corresponding Author:
Guilherme Del Fiol
Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, USA.
421 Wakara Way, Suite 140, Salt Lake City, UT 84108-3514. Phone: +1-801-213-4129
Email: guilherme.delfiol@utah.edu.

Abstract

**Background:** A major barrier to the practice of evidence-based medicine is efficiently finding scientifically sound studies on a given clinical topic.

**Objective:** To investigate a deep learning approach to retrieve scientifically sound treatment studies from the biomedical literature.

**Methods:** We trained a Convolutional Neural Network using a noisy dataset of 403,216 PubMed citations with title and abstract as features. The deep learning model was compared with state-of-the-art search filters, i.e. PubMed’s Clinical Query Broad treatment filter, McMaster’s textword search strategy (no MeSH terms), and Clinical Query Balanced treatment filter. A previously annotated dataset (Clinical Hedges) was used as the gold standard.

**Results:** The deep learning model obtained significantly lower recall than the Clinical Queries Broad treatment filter (96.9% vs. 98.4%; p<0.001); and equivalent to McMaster’s textword search (96.9% vs. 97.1%; p=0.57) and Clinical Queries Balanced filter (96.9% vs. 97.0%; p=0.63). Deep learning obtained significantly higher precision than the Clinical Queries Broad filter (34.6% vs. 22.4%; p<0.001).
and McMaster’s textword search (34.6% vs. 11.8%; p<0.001); but significantly lower than the Clinical Queries Balanced filter (34.6% vs. 40.9%; p=<0.001).

**Conclusion**: Deep learning performed well compared to state-of-the-art search filters, especially when citations are not indexed. Unlike previous machine learning approaches, the proposed deep learning model does not require feature engineering, time-sensitive or proprietary features, such as MeSH terms and bibliometrics. Deep learning is a promising approach to identifying reports of scientifically rigorous clinical research. Further work is needed to optimize the deep learning model and to assess generalizability to other areas, such as diagnosis, etiology, and prognosis.

**Keywords**: Information retrieval; evidence-based medicine; deep learning; machine learning; literature databases.

**Introduction**

With roughly 95 clinical trials published per day, the biomedical literature is increasing at a very rapid pace, imposing significant challenge to the practice of evidence-based medicine. However, only 1% of studies in the biomedical literature meet minimum criteria for scientific quality[1] and most published research findings are eventually shown to be false.[2] As a result, a major barrier to the practice of evidence-based medicine is efficiently finding the relatively small number of scientifically sound studies on a given clinical topic. Systematic reviews and meta-analyses attempt to summarize the available evidence on a given clinical question aiming for near perfect recall. However, systematic reviews are often not available and become quickly outdated. Therefore, clinicians may benefit from access to the latest evidence from high quality clinical trials before they are included in systematic reviews.

For over two decades, the Clinical Query filters have been the state-of-the-art approach to retrieve scientifically sound clinical studies from the primary literature, both for the development of systematic reviews and point-of-care decision support. [3, 4] The Clinical Query filters consist of search strategies based on textwords and MeSH terms that have been developed and validated through a systematic approach. [5] Although initially developed in the 1990s, the Clinical Query filters have been updated over time. In addition, a more recent study has shown that the recall and precision of the filters developed in 2000 did not change significantly a decade later. [6] Clinical Query filters for several topics are available in PubMed and several other bibliographic biomedical databases, with focus on areas such as therapy, diagnosis, etiology, and prognosis; and tuned for precision or recall. A limitation of the Clinical Query filters is their dependency on MeSH terms, which are added to PubMed citations 23 to 177 days after an article is published according to a previous study[7] and 17 to 328 days according to our more recent analysis. In addition, there is room for improvement, especially in terms of retrieval precision.

Previous studies investigated the use of machine learning approaches to automate the retrieval of scientifically sound studies.[8-10] Features used in those studies include bibliometrics (e.g., citation count, impact factor), words in the article title
and abstract, MeSH terms, UMLS concepts, and semantic predications. Although the results of machine learning studies were promising, they had important limitations that precluded wide adoption in practice, such as a requirement for significant feature engineering (e.g., UMLS concepts, bibliometrics), reliance on proprietary and time-dependent features (e.g., MeSH index, citation counts), and potential overfitting to a particular dataset.

In the present study, we investigated a deep learning approach for the retrieval of scientifically sound treatment studies from PubMed. To overcome limitations of previous methods, we focused on an approach that requires very little feature engineering and does not rely on proprietary or time-dependent features. We then compared the performance of a deep learning model with state-of-the-art PubMed search strategies against Clinical Hedges, a rigorous gold standard of over 50,000 studies that were systematically rated for scientific quality according to rigorous criteria.[5]

**Deep machine learning**
Recent advances in machine learning have led to dramatic improvements in the abilities of computers to mimic human activities. Many of these improvements leverage “deep learning,” which embody neural-networks with many nodes that are fully connected across layers of the network. In the context of supervised deep learning, which we utilize here, such a network is trained by providing many examples of the objective to classify, as well as many counter examples.

The main learning for a neural network involves “forward propagation” and “backward propagation.” During forward propagation, inputs are translated into features by transforming the inputs into real-valued vectors of fixed sizes. These vectors (e.g., “layers”) are combined with weights and passed through an activation function that summarizes the contribution of each feature of the vector and its weight. Layers are connected to one another such that the values from the activation function of the current layer become the inputs to the next layer. Therefore, the “forward propagation” starts with input and passes activation values from layer to layer until the final layer, which outputs some decision vector. In our case, this final output function is a sigmoid activation function, which can assign probability to class membership.

During “backward propagation” the final classification decision is compared with the known result from the training data, and errors are propagated backward through the network, from the output layer to the input layer. Each weight is updated according to its contribution to the erroneous (or correct) decision, via gradient descent.

To train a network, therefore, involves multiple passes of forward propagation followed by backward propagation. It is common to call each iteration over all the training data an “epoch”. The model generally stops this training process at a fixed number of epochs or when the metric of success appears to have reached some maximal value.
In this study, we utilize a particular deep-learning neural-network known as Convolutional Neural Network (CNN) to do our text classification, following the approach by Kim.[11] CNNs analyze text using sliding word windows of specified sizes. Each sliding word window generates a set of real-valued vectors. Generally, each word or even character is associated with a “word-embedding”, which is a low-dimension real-valued vector that represents the semantic space for the word.[12] Therefore, as each term is associated with a vector, each sliding word window then represents a matrix. Each sliding word window is then passed through an activation function, and a “max pooling” is applied such that only the maximum value is kept from the set of values produced by the activation function as applied to the window. That is, each window is associated with its single, maximal value outputted by the activation function. These maximal values are concatenated together to form their own vector representing the set of windows. This set of concatenated values forms the next layer, which is then passed to the final layer, which includes the decision-making activation function (such as Softmax, as described above).

An example of a CNN is shown in Figure 1 below. From left to right, we see one set of input words and their word embeddings, which forms the initial input matrix. This network uses two sets of sliding windows, one of size two and one of size three. These produce the convolutional layer, transforming the sliding window’s features into new feature values which are then pooled such that only the maximum value is kept (the “max pooling”). Finally, the max-pooled values are passed through the fully connected final (output) layer, which uses Softmax to assign a probability of class membership (shown as “yes” or “no” for binary class membership).

Figure 1 – Example of a Convolutional Neural Network
**Methods**

Overall, our approach consisted of training and testing deep learning models with a large and noisy dataset obtained automatically through PubMed searches based on the Clinical Query treatment filter; and evaluating the performance of the resulting model against the Clinical Hedges gold standard. Clinical Hedges is a database previously developed by the Hedges Group at McMaster University and that was used to develop and evaluate the Clinical Query filters.[5] The database has 50,594 articles published in 170 clinical journals. All articles were manually annotated by highly-calibrated information science experts according to type (e.g., etiology, prognosis, diagnosis, prevention, therapy, clinical prediction) and whether or not each study meets pre-specified methodologic criteria for scientifically sound research.

Specifically, the study method consisted of the following steps, which are described in more detail in the sections below: 1) preparation of a dataset for training the deep learning models; 2) training and tuning deep learning models; 3) comparison of the deep learning approach with state-of-the-art search filters and McMaster’s textword filter in terms of precision and recall; 4) analysis of deep learning performance in terms of precision at several levels of K.

**Preparation of training dataset**

The training/testing dataset consisted of 403,216 citations retrieved from PubMed using one search strategy aiming to retrieve scientifically sound studies (i.e. positive studies) and a second strategy to retrieve negative studies (Figure 2). The first strategy was based on the Clinical Queries treatment filter tuned for precision (“narrow” filter). In previous studies, this filter yielded 93% recall and 54% precision for scientifically sound treatment studies in the Clinical Hedges gold standard. Therefore, this search strategy was used as a surrogate for retrieving a large dataset of scientifically sound studies that are similar to the ones in the Clinical Hedges gold standard. The second strategy retrieved studies conducted in humans and that were not retrieved by the first search strategy. The strategies were limited to retrieve a maximum of 150,000 and 300,000 citations respectively to yield a dataset with one third positive and two thirds negative citations. Both strategies were limited to citations published between 2007 and 2017. Citations without an abstract were removed. The search strategies were executed with PubMed’s eUtils API. The resulting dataset contained 147,182 positive and 256,034 negative citations (Figure 2).

**Scientifically sound studies (positive citations)**

**Therapy/Narrow[filter] AND**

("2007/09/23"[PDAT]:"2017/09/19"[PDAT])

**Negative citations**


Figure 2 - Search strategies used to retrieve scientifically sound (positive) and negative citations.
**Training and tuning deep learning models**

Deep learning models were trained using 90% of the citations in the dataset, using the remaining 10% as a “development” set. As the training/development split was randomly generated, the development set maintained a similar proportion of positive to negative instances as the training set. To build model inputs, we concatenated the title with the abstract, removed stopwords, and kept the first 650 tokens of the remaining words.

Our model itself follows the convention of applying Convolutional Neural Networks for text classification. The first layer of our model applies character embedding to the words, so that words outside of the known vocabulary can be included for prediction. Then, the character embeddings are combined with word embeddings (built from the training data), to capture semantic similarity. This input is passed into our model, which contains two convolutional layers, one for sliding word windows of size two and one for word windows of size three. Each convolutional layer contains 512 filters associated with it. We apply a ReLU unit to the convolutional layers, and pass them through a max pooling procedure. The resulting max-pooled features are then concatenated into a single layer. The max-pooled layer is passed to the next layer which consists of 512 units, fully connected, and to which we apply a Softmax activation function to predict the probability of a citation belonging to either class. We then take the Argmax of the Softmax predictions as the predicted class. We ran this model with dropout regularization of 0.5 to prevent over-fitting, for 30 epochs. Hyper-parameters were chosen experimentally based on maximized precision on the training data.

![Figure 2 - Datasets used for training and testing the deep learning models.](image)
Comparison of the deep learning approach with state-of-the-art PubMed search strategies

We tested three hypotheses that reflect the requirements imposed by different information retrieval scenarios. The first scenario consists of search strategies to support the development of evidence-based syntheses, such as systematic reviews and clinical guidelines.[13] In this scenario, there is a requirement for near perfect recall. The hypothesis for this scenario is that the deep learning approach yields equivalent recall with higher precision for scientifically sound treatment studies compared with the PubMed Clinical Queries Broad filter, which has almost perfect recall (Figure 4).

The second scenario reflects the need to retrieve recent studies, such as in literature surveillance efforts to identify new evidence to update existing systematic reviews and clinical guidelines.[14-16] Since Clinical Query filters depend partially on MeSH terms and publication type, they are less effective for literature surveillance. Instead, search strategies based on terms in the citation title and abstract are preferred. The hypothesis for this scenario is that the deep learning approach yields equivalent recall but higher precision for scientifically sound treatment studies compared with a textword search strategy provided by the Clinical Hedges group from McMaster University (Figure 4).

Hypothesis #1 (evidence synthesis - maximize recall) - Clinical Query Treatment Broad filter

\[
((\text{clinical}[\text{Title/Abstract}] \text{ AND } \text{trial}[\text{Title/Abstract}]) \text{ OR clinical trials as topic[MeSH Terms]} \text{ OR clinical trial[Publication Type]} \text{ OR random*[Title/Abstract]} \text{ OR random allocation[MeSH Terms]} \text{ OR therapeutic use[MeSH Subheading]})
\]

Hypothesis #2 (literature surveillance - no MeSH terms, maximize recall) - McMaster's textword search

\[
\text{hasabstract[text]} \text{ AND (effect*[Title/Abstract]} \text{ OR control[Title/Abstract]} \text{ OR controlled[title/abstract] OR random*[Title/Abstract]} \text{ NOT (qualitative[title/abstract] OR retrospective[title/abstract]} \text{ OR mice[title/abstract]} \text{ OR rat[title/abstract]} \text{ OR rats[title/abstract]} \text{ OR editorial[pt]} \text{ OR letter[pt]})
\]

Hypothesis #3 (patient care decision making - balance between precision and recall) - McMaster's Balanced Clinical Query filter

\[
((\text{clinical}[\text{Title/Abstract}] \text{ AND } \text{trial}[\text{Title/Abstract}]) \text{ OR random*[Title/Abstract]})
\]

Figure 4 – Search strategies used to retrieve scientifically sound treatment studies in comparison with the deep learning model.
The third scenario represents clinicians searching the literature for evidence to meet clinicians’ information needs that are raised in the care of a specific patient.[17] In this scenario, trading a small loss in recall for substantial gains in precision is acceptable. We hypothesize that the deep learning approach yields equivalent recall but higher precision for scientifically sound treatment studies compared with McMaster’s Balanced Clinical Query filter, which uses a combination of textwords, MeSH terms, and publication types (Figure 4).

The Clinical Hedges gold standard was used to test the three hypotheses. For positive citations, we retrieved from the Clinical Hedges database 1,524 original, scientifically sound studies with focus on treatment. For negative citations, we retrieved 29,144 treatment studies from Clinical Hedges that were not in the positive set. For statistical analysis, we split the resulting dataset into 20 random subsamples, stratified to ensure a balanced ratio of positive and negative citations in each subsample. Measures of precision, recall, and F-measure were obtained for the four approaches on each of the 20 subsamples (Figure 3). Last, we ranked the output of the deep learning model according to its probability score and obtained measures of precision at several levels of top K citations (10, 20, 50, 100, 200, 300 and 500).

Figure 3 – Evaluation method, including the datasets and hypotheses tested in three experiments.
Statistical Analysis
Classification performance was measured according to the average precision and recall across 20 data samples. We used the paired Student t-test to test the significance of the differences in recall and precision between the two approaches in each experiment, with the significance level set at 0.05.

Results
Table 1 shows the results of the comparisons for Hypothesis #1 (evidence synthesis - maximize recall). The Clinical Queries Broad filter had statistically significantly higher recall than the deep learning model (98.4% vs. 96.9%; p=0.002), although the difference was small (-1.6%) and likely marginal in practice depending on the use case. The deep learning model had significantly higher precision than the Clinical Queries Broad filter, with a +12.2% absolute difference (34.6% vs. 22.4%; p<0.001).

Table 1 – Average recall, precision, and F-measure of the deep learning model and Clinical Query Broad filter according to the Clinical Hedges gold standard (N=20).

<table>
<thead>
<tr>
<th></th>
<th>Deep Learning</th>
<th>CQ Broad</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>96.9%</td>
<td>98.4%</td>
<td>p&lt;0.001</td>
</tr>
<tr>
<td>Precision</td>
<td>34.6%</td>
<td>22.4%</td>
<td>p&lt;0.001</td>
</tr>
<tr>
<td>F-Measure</td>
<td>51.0%</td>
<td>36.5%</td>
<td>p&lt;0.001</td>
</tr>
</tbody>
</table>

*CQ = PubMed Clinical Query Treatment filter

Table 2 shows the results of the comparisons for Hypothesis #2 (Literature surveillance – no MeSH terms). The deep learning model was equivalent to McMaster’s textword search in terms of recall (97.1% vs. 96.9%; p=0.57); and had significantly higher precision than the textword search (34.6% vs. 28.5%; p<0.001).

Table 2 – Average recall, precision, and F-measure of the deep learning model and McMaster’s textword search according to the Clinical Hedges gold standard (N=20).

<table>
<thead>
<tr>
<th></th>
<th>Deep Learning</th>
<th>Textword Search</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>96.9%</td>
<td>97.1%</td>
<td>0.57</td>
</tr>
<tr>
<td>Precision</td>
<td>34.6%</td>
<td>11.8%</td>
<td>p&lt;0.001</td>
</tr>
<tr>
<td>F-Measure</td>
<td>51.0%</td>
<td>21.0%</td>
<td>p&lt;0.001</td>
</tr>
</tbody>
</table>

Table 3 shows the results of the comparisons for Hypothesis #3 (Patient care decision support – balance between recall and precision). Compared with the McMaster Balanced treatment filter, the deep learning model had similar recall (96.9% vs. 97.0%; p=0.63), but lower precision (34.6% vs. 40.9%; p<0.001) (Table 3).

Table 3 – Average recall, precision, and F-measure of the deep learning approach and McMaster’s Balanced Treatment filter according to the Clinical Hedges gold standard (N=20).

<table>
<thead>
<tr>
<th></th>
<th>Deep Learning</th>
<th>McMaster’s CQ Balanced filter</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>----------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td>Recall</td>
<td>96.9%</td>
<td>97.0%</td>
<td>0.63</td>
</tr>
<tr>
<td>Precision</td>
<td>34.6%</td>
<td>40.9%</td>
<td>P&lt;0.001</td>
</tr>
<tr>
<td>F-Measure</td>
<td>51.0%</td>
<td>57.5%</td>
<td>P&lt;0.001</td>
</tr>
</tbody>
</table>

**Precision at K**

The precision at K curve for the ranked output of the deep learning model shows that precision ranged from 75.5% to 61% among the top 10 to top 100 citations and only decreased substantially after the top 200, 300, and 500 citations (Figure 4).

![Figure 4](image)

**Discussion**

To our knowledge, this is the first study to investigate the use of deep learning techniques to identify reports of scientifically sound studies in the biomedical literature in three different information seeking scenarios. The deep learning approach performed reasonably well compared with state-of-the-art search filters, especially for literature surveillance. For evidence synthesis, the deep learning approach had slightly lower recall (-1.6%), but significantly higher precision (+12.2%) than the PubMed Clinical Query Broad treatment filter. For literature surveillance, the deep learning approach had equivalent recall to McMaster’s textword filter, but significantly higher precision (+22.2%). For patient care decision-making, the deep learning model had similar recall, but lower precision (-6.3%) than McMaster’s Balanced filter. Strengths of the study method include the
use of a very large training set, comparison with state-of-the-art search strategies, and evaluation with a rigorous gold standard, completely independent from the training set.

The proposed deep learning approach has three main potential benefits compared with previous approaches. First, unlike previous machine learning approaches, which depend on features that are not always openly and contemporaneously available (e.g., MeSH terms, citation counts, journal impact factors), the proposed deep learning approach only uses citation title and abstract, which are available as soon as citations are entered in PubMed. Although full-text articles could be added as features as an attempt to improve performance, obtaining access to the full-text of all articles indexed in PubMed is impracticable since most journals do not provide open access to full-text. To assess the potential duration of delays for literature surveillance strategies based on MeSH filters, we determined the time between the date of creation of the article record in PubMed (CRDT) and the date of posting of Medical subject headings (MHDA) for 107 journals (55,237 articles) in the McMaster PLUS database, from which the Clinical Hedges database was derived. The mean delay in MeSH indexing per journal was 162 days (95% confidence interval (CI) 157 to 167), with a range of 17 to 328 days. Indexing intervals for journals were inversely correlated with journal impact factors (for 2016), but the correlation was relatively weak (-0.38; CI -0.199 to -0.517). As a second benefit, the deep learning model provides a ranked output with 70% or higher precision among the top 50 citations. This feature could be particularly useful for clinicians in busy clinical settings who are less likely to look beyond the top 20 citations that are displayed in PubMed searches.[17, 18] In addition, citation ranking could help with systematic review development, since front-loading “eligible” citations can be used to help train and calibrate citation screeners and prioritize work.[19] Third, the deep learning model obtained reasonable performance despite being trained on a noisy dataset (an estimate of roughly 50% of the positive cases were false-positives). This finding confirms the robustness of the deep learning approach, which is known to be resilient to noisy training data.[20]

**Comparison with Prior Work**

Previous work applied deep learning to classification tasks in the biomedical informatics domain. Lee classified sentences as belonging to papers that would be included in a systematic review, or not.[21] However, because they did not employ a large-scale training procedure, as we devised here, their results were poor. Also, it is not clear whether the author focuses solely on sentence classification, or document classification as in our work. Hughes, et al. applied CNNs to classify sentences according to one of 26 categories, such as “Brain” or “Cancer”, using a similar approach (though a different training procedure) to a different problem.[22] Both Nguyen, et al.[23] and Che, et al.[24] utilized CNNs to predict risk outcomes, such as hospital readmission, using electronic health record data as input. As with Hughes et al., although applied to different problems, the latter studies demonstrate precedent for using CNN in biomedical text classification.
A polynomial Support Vector Machine (SVM) classifier based on MeSH terms, publication type, and title/abstract words obtained a recall of 96% and precision of 18% against a gold standard of internal medicine articles included in the ACP Journal Club.[9] A different study compared Clinical Query filters, machine learning, and algorithms based on citation count and the PageRank algorithm using a gold standard of important literature on common problems in surgical oncology.[10] The PageRank algorithm obtained a precision at the top 10, 20, 50, and 100 citations of 7.8%, 13.0%, 19.9%, and 26.3% respectively. Overall precision and recall were not reported. More recently, a study by Kilicoglu et al. investigated a set of classifiers using features such as MeSH terms, title/abstract words, UMLS concepts, and semantic predications. A Naïve Bayes classifier with these features obtained a recall and precision of 91.4% and 52.5% for treatment studies in the Clinical Hedges database.[8] As discussed above, those previous approaches rely on substantial feature engineering and/or proprietary and time-sensitive features, compromising the use of those approaches in real-time information retrieval systems. In a recent study investigating an approach similar to ours, Marshall et al. developed CNN and support vector machine classifiers based on article title and abstract to identify reports of randomized controlled trials (RCTs).[25] The best classifier obtained a recall of 98.5% and precision of 21%. Although the authors also evaluated their classifiers against the Clinical Hedges database, the results cannot be directly compared with our study because their goal was to identify RCTs vs. scientifically sound studies (not all RCTs are scientifically sound and not all scientifically sound studies are RCTs). Another difference was that Marshall et al. used a training set derived from RCTs identified in Cochrane systematic reviews while we used a dataset obtained using the Clinical Queries Treatment Narrow filter.

**Limitations**

Our study has four important limitations. First, although we focused on deep learning models and optimization strategies that were most likely to produce the best results, we have not exhausted all deep learning optimization possibilities. For instance, new work on Recurrent Neural Networks (RNNs) may prove more accurate in document classification tasks.[26, 27] We chose to focus our efforts on CNNs because they run more efficiently, given the large scale of our text data, but there is a valid investigation into understanding the trade-offs between speed and accuracy by comparing these methods. We also did not exhaustively search the hyper-parameter space for our CNN. Many of our choices were empirical, as this is the first study, and further efforts might leverage more systematic approaches to hyper-parameter tuning.[28] Second, our approach is meant to be “end-to-end,” i.e. text simply enters our pipeline and is classified. This approach is preferable because it does not require significant feature engineering or time-dependent features such as MeSH terms. However, further studies can explore adding richer features into our model to improve performance. For example, since the McMaster’s textword filter has equivalent recall as, but lower precision than the Clinical Query filters, it is possible that MeSH-based features could improve the precision of our deep learning approach. Third, we have made comparisons with only one textword filter and no other machine learning approaches, since we did not have access to those machine
learning classifiers. Comparisons with two of the three previous machine learning approaches are indirect, since those studies did not use Clinical Hedges as a gold standard. Last, we focused on identifying “treatment” studies; further work is needed to verify if our approach generalizes to other areas, such as diagnosis, etiology, and prognosis.

Conclusion

We compared deep learning with state-of-the-art search filters to identify reports of scientifically sound studies in the biomedical literature. Overall, the resulting deep learning model compared well with other approaches, especially in scenarios involving recent citations prior to MeSH indexing. Advantages of the deep learning approach include low feature engineering requirements, no dependency on proprietary and time-sensitive features, and use of a very large training set. Future work is needed to investigate further optimization opportunities and to adapt the deep learning approach to other clinical areas. Deep learning is a promising approach to identifying scientifically sound studies from the biomedical literature and warrants further investigation as a potential alternative for or supplement to current search filters.

Acknowledgements

GDF was funded by National Library of Medicine grant 1R01LM011416 and National Cancer Institute grant 1U24CA204800. Further, this material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) Program Office under Contract No. W31P4Q-17-C-0103. The authors wish to acknowledge InferLink Corporation for their advice and support. The Clinical Hedges database was created with grants from the US National Library of Medicine and the Canadian Institutes of Health Research.

Conflicts of Interest

MM is the Chief Scientist of InferLink Corp. and CEO of Evid Science, Inc. both of which could benefit from using the above approach as a feature within existing or new medical literature analysis products. GDF, AI, CC, and RBH have no competing interests to declare.

References

4. Wilczynski NL, McKibbon KA, Haynes RB. Enhancing retrieval of best evidence for health care from bibliographic databases: calibration of the hand search