Bi-directional Relevance Matching between Medical Corpora

Jingnan Yang, Justin Ward, Erfaneh Gharavi, Jennifer Dawson, Rafael Alvarado
jy4fch@virginia.edu, jw8vw@virginia.edu, eg8qe@virginia.edu, JDawson@cochrane.org, rca2t@virginia.edu,

Abstract – **Readily available, trustworthy, and usable medical information is vital to promoting global health. Cochrane is a non-profit medical organization that conducts and publishes systematic reviews of medical research findings. Over 3000 Cochrane Reviews are presently used as evidence in Wikipedia articles. Currently, Cochrane's researchers manually search Wikipedia pages related to medicine in order to identify Wikipedia articles that can be improved with Cochrane evidence. Our aim is to streamline this process by applying existing document similarity and information retrieval methods to automatically link Wikipedia articles and Cochrane Reviews. Potential challenges to this project include document length and the specificity of the corpora. These challenges distinguish this problem from ordinary document representation and retrieval problems. For our methodology, we worked with data from 7400 Cochrane Reviews, ranging from one to several pages in length, and 33,000 Wikipedia articles categorized as medical. We explored different methods of document vectorization including TFIDF, LDA, LSA, word2Vec, and doc2Vec. For every document in both corpora, their similarity to each document in the opposing set was calculated using established vector similarity metrics such as cosine similarity and KL-divergence. Labeled data for this unsupervised task was not available. Models were evaluated by comparing the results to two standards: (1) Cochrane Reviews currently cited in Wikipedia articles and (2) a data set provided by a medical expert that indicates which Cochrane Reviews could be considered for specific Wikipedia articles. Our system performs best using TFIDF document representation and cosine similarity.**

Index Terms – Document Similarity, Cochrane, Medical Document Analysis, Automated Citation Recommendation.

Introduction

Medical journals provide information that directly informs consumer health decisions. By having this information accessible, more people are able to learn and make use of information that is valuable for general health and well-being. One major drawback of this research area is that, due to the in-depth details covered and context specific language used, it can be a challenge for a person without background expertise to gain practical insight from the literature. This is limiting to both the spread and general use of the research.

One approach to sharing information to a general audience is through online encyclopedias such as Wikipedia. Wikipedia is among the most popular sources of medical information [1]. Depending on the topic, Wikipedia's medical articles have the potential to be exposed to and consumed by millions of people from around the world. There are over 35,000 articles that pertain to human health on Wikipedia. Numerous medical experts volunteer to keep these articles up to date, however, ensuring that all evidence shared in these articles is accurate and unbiased can be a challenging task.

Cochrane is a non-profit medical organization that develops and disseminates systematic reviews of medical research studies and clinical trials so that healthcare professionals and health consumers have access to high quality unbiased medical information. Cochrane's aim is to share the findings broadly: with medical professionals as well as the general public. Cochrane partnered with WikiProject Medicine in 2014 with the goal of helping to improve health-related content on Wikipedia using evidence produced by Cochrane. Cochrane volunteers and volunteer Wikipedia editors may consider Cochrane evidence when improving a Wikipedia article. Presently, a volunteer interested in contributing to the initiative manually searches The Cochrane Library or PubMed[1], a document cataloging website, to identify a Cochrane Review that may improve the evidence base of a particular Wikipedia article. There are presently over 7500 available Cochrane Reviews. Organizing this project and tracking which Cochrane Review may be associated with a particular Wikipedia article is a time-consuming task to do manually. We are interested in streamlining this process of matching Cochrane Reviews with potential Wikipedia articles (and vice-versa). While not all Cochrane Reviews are appropriate to add to Wikipedia, there are many reviews that have not been considered (i.e.: approximately 4000).

Using an automated system to identify potential Cochrane Reviews to consider for each Wikipedia article could lead to a more efficient editing process. As such, we have developed a system to take any review or article and return the most similar documents from the opposing corpus. To accomplish this task, a variety of existing document similarity approaches used in machine learning and information retrieval problems were explored to identify the most effective methods for finding similarity between long,

[1] https://www.ncbi.nlm.nih.gov/pubmed/

structured, medical texts. This system involved using different preprocessing approaches, document representation methods, and similarity evaluations and applying them to both corpora to explore the results.

Related Works

The problem of automated document comparison is a subject that has long been studied. There are several methods to compare different documents automatically; most of which draw on ideas involving information retrieval, semantic analysis, document clustering, and document classification [2]. One common way of finding similarity between documents is the application of different types of semantic analysis. In these analytic procedures, the similarity of the meaning of two documents is quantified through a distance measured after vectorization is performed on the text. After a model has been built on enough documents, a meaningful measure of the relatedness among all documents in a given corpus can be quantified [3][4]. As mentioned before, document classification and clustering have also been used as ways of pointing out similar documents. Through these methods, a document is assigned certain numeric values based on the features it contains. Once a method is established for labelling documents based on these features, new documents can be classified based on how they fall into given categories [5].

While these approaches have been taken to handle document grouping, the problem at hand poses some unique challenges. One notable element that separates this problem from others is that texts being compared all fall within the special domain of medical knowledge. Comparisons between medical corpora faces inherent challenges with terminology mismatch, inferred relevance, and structural specializations [6]. Exacerbating this issue further is the fact that Wikipedia articles use general language compared to the more specific reviews which they need to be matched with. Being able to understand that documents may be logically related, but use language differently, more generally for Wikipedia and more specific for the Cochrane reviews, has been explored to some extent in how it can be incorporated into information retrieval problems [7].

Data

Our data consist of two corpora of articles from Wikipedia's medical section and the Cochrane Database of Systematic Reviews. The Wikipedia data were extracted using PyWikibot and Wikimedia Dumps[2] and the Cochrane data were provided by a Cochrane representative.

I. Wikipedia Data

The Wiki data includes 33469 articles categorized as medical[3] in Wikipedia, which describe subjects including diseases, medicines, and public health issues. A total of 2218 of the Wikipedia articles have already cited relevant Cochrane Reviews. A log was kept of these Cochrane reviews and the Wikipedia articles they were on. Wikipedia articles contain various graphs and images, but for the purposes of this project only plain text was kept.

II. Cochrane Data

Cochrane Reviews use a systematic review process to determine whether or not all the available medical evidence related to a research question that meets predetermined criteria is conclusive. A research question is first formulated, and then all existing medical research on a topic that meets defined criteria is identified. Stringent guidelines are used to assess the primary research in order to determine the quality of the evidence and if it is conclusive. Cochrane Reviews are updated on a regular basis as new medical studies (e.g.: randomized controlled trials) are conducted. Unlike Wikipedia articles, which are structured in relatively random format, Cochrane Reviews follow a strict structure. Our dataset consists of 7388 Cochrane Reviews. 1983 of these publications have been cited in one or more Wikipedia articles. Each Cochrane review could have up to 5 versions of publications, where content changes slightly between versions.

Proposed Methodology

Our proposed method to address this problem consists of 3 main steps: preprocessing, document representation, and similarity measurement. The entire process is visualized in Figure 1. All models were created in Python using the libraries NLTK[4], Genism[5], and SpaCy[6].

I. Preprocessing

For both corpora of documents, the Wikipedia articles and the Cochrane Reviews, general preprocessing approaches of stopword and punctuation removal and lowercasing were performed. For the bag-of-word models (TFIDF, LSA, and LDA), additional preprocessing approaches were applied:

- **n-gram**: For many pieces of text, it can prove effective for analysis to explore words that appear frequently with other words [8]. For our models, bigrams and trigrams were used due to the predominance of 2 and 3 word sequences (i.e, vitamin C, oral cancer, etc.) found in observations of the text.
- **Part-Of-Speech Tags**: Words in a document can also be labelled based on their use in the text [9]. Different combinations of nouns, adjectives, and verbs were isolated to be used as input for our text representation methods.
- **RAKE Algorithm**: The keywords of a text describe the text's main topics. They can be used to compare the similarity between different texts. During preprocessing, we also used RAKE to extract keywords from our

[2] https://dumps.wikimedia.org/

[3] https://en.wikipedia.org/wiki/Category:WikiProject_Medicine_articles

[4] https://www.nltk.org/

[5] https://radimrehurek.com/gensim/

[6] https://spacy.io/

documents [10]. We extracted both unigram and bigram keywords for each Wikipedia article and each Cochrane review with Rake and used them in document representations.

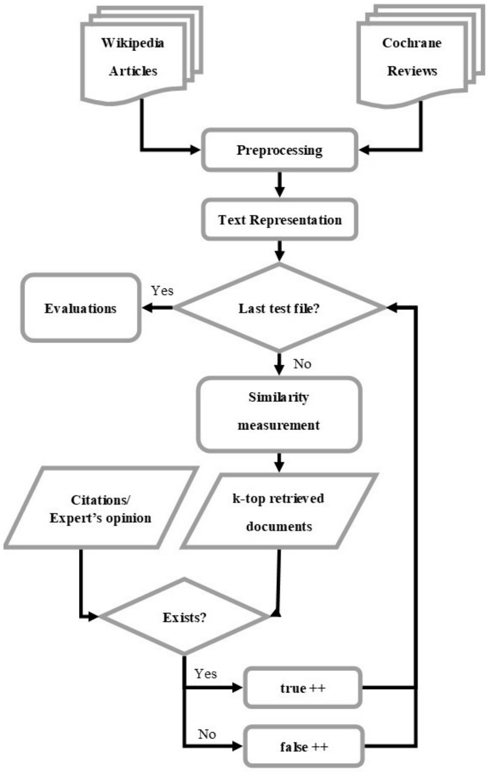

FIGURE I
METHODOLOGY

II. Document Representation

After preprocessing the documents, texts were vectorized to numerical representations using the following methods:

- **TFIDF**: Term frequency inverse document frequency is one of the long-lasting methods for text representation. Though newer methods of document representation have been heavily explored in the past decade, this method remains popular specifically for large document where words order does not matter [11]. In this approach, our corpora of medical documents are represented as rows in two separate matrices where each column represents a specific word using a factor of that word's frequency within the given document and the log of that word's total appearance in the corpus.
- **LSA**: Latent semantic analysis is an extension of TFIDF where singular-value decomposition (SVD) is applied to the TFIDF matrix with the intuition that these new columns represent document topics [6]. LSA was applied leaving a specific number of columns equivalent to the document topics in the output matrix.
- **LDA**: Unlike LSA, which learns to represent document topics through SVD representation, Latent Dirichlet allocation builds a probabilistic distribution of topics in documents. Methods such as variational inference and Gibbs' Sampling are used to generate distributions assuming a Dirichlet prior for distribution of words and topics within the text [12]. Representations could be built on our documents with varying numbers of topics ranging from 50 to 200.
- **Doc2Vec**: The goal of doc2Vec is to learn a conceptual representation of a document through training by running the corpora through a shallow neural network. Unlike the previous methods, which are varying representations of information learned from term and document frequencies in a bag-of-world model, doc2Vec learns through making prediction of words presence when running through a document [13]. Representation was similar in structure to the previous three approaches in that each representation was built for one corpus and then applied to the other separately.
- **Word Averaging:** Just as vectors can be built to represent documents, they can also be used to represent individual words [14]. For each document, we calculated this representation by averaging the word vectors in the document as wv in (1). Here, n is the number of words in the document and wj is the word vector of the jth word.

$$wv = \sum_{j}^{n} wj/n \qquad (1)$$

III. Similarity Metrics

After transforming every document into a numeric representation, the similarity between any document in one corpus and any document in the opposing corpus is calculated. This can be used to obtain the k most similar documents in one corpus for any given document in the opposing corpus.

- **Cosine Similarity**: A similarity metric calculated by measuring the angle between two vectors. The cosine of the angle is found by taking the dot product of the two vectors being measured and dividing that by the product of each vectors' norm. This metric is useful in many document similarity tasks because it is unaffected by the length of incoming documents . The range of cosine similarity is between -1 and 1 . This metric was preferred for the majority of our analysis due to the longstanding use in information retrieval and efficient computational time [15].
- **KL-Divergence**: Kullback-Leibler divergence is an asymmetric measure of the difference between two distributions. It is commonly used in information retrieval tasks when assuming the vector representations to be probability distributions [15]. In this metric, values closer to 0 are considered to be more similar. When implementing this on our data, difficulties were met in scaling to larger numbers of documents.
- **Additional**: Other ddistance methods were used at different points of our model building. Euclidean and Manhattan were initially implemented, though because they are affected by vector magnitude they were only used for baseline comparisons. Metrics such as Hellinger distance and were used to measure distributions differences like KL, but as true distance measurements.

EVALUATION

Evaluation was a challenge given the unsupervised nature of our problem. There were no labelled data to train and test our model. We used the pre-mentioned log of Cochrane Reviews already cited on Wikipedia as well as related documents found by domain experts to be considered as a test set for our evaluation. The metrics that were used for this system were accuracy, precision, and recall. Additionally, individual samples were analyzed and reported throughout the project.

I. Evaluation based on citations

Initially, we had access to a log of all Cochrane Reviews cited on Wikipedia and used this to analyze system results. For this evaluation, the top 10 Cochrane reviews were generated for each Wikipedia Article in our sample data. The measurement used for this evaluation was accuracy, (2). The accuracy, a, is the total number of correctly retrieved citations for all n Wikipedia articles divided by the total number of citations on every ith article, Ci, summed over all Wikipedia articles. Here ci represents the system retrieved results for the ith article. Accuracy was used to explore best text representation methods on the sample data.

$$a = \frac{\sum_i^n \#(Ci \cap ci)}{\sum_i^n Ci} \tag{2}$$

II. Evaluation based on expert opinion

The second evaluation approach taken used a listing of most similar Wikipedia articles to a subset of Cochrane Reviews, provided by domain experts. To have more refined evaluations of our model, the precision and recall measures commonly used in information retrieval studies were calculated on the full data set. The evaluation data given by experts were relevant Wikipedia articles for specific Cochrane Reviews. To use this data in our evaluation, the top 10 Wikipedia articles were generated for every Review.

The precision rate, p, measures the ratio of the model results that coincide with the test data - the number of documents that exist in both expert and system citations divided by the number of documents in the model results, seen in (3). Where Di refers to the expert suggested Wikipedia articles for the ith Cochrane Review and di refers to the suggestions given by our model.

$$p = \sum_i^n \frac{\#(Di \cap di)}{\#di} / n \tag{3}$$

Recall, denoted as r in (4), measures the ratio of the test data that coincide with the model results - the number of documents that exist in both sets divided by the number of documents in the test data. Once calculated, the precision rate and recall are separately averaged. Finally, have an average precision rate and an average recall rate to measure the performance of our results.

$$r = \sum_i^n \frac{\#(Di \cap di)}{\#Di} / n \tag{4}$$

These evaluations were performed on the full dataset as many of expert evaluation data were not present in the sample. Additionally, some of the Wikipedia articles labelled as relevant did not exist in our full database. To report performance these samples were discarded.

RESULTS AND DISCUSSION

To find the best combination of text representation and similarity metric, accuracy was calculated for each method for a sample of the data. Despite embedding techniques such as Word2Vec and Doc2Vec being heavily favored in recent years as word representations, the most success in this problem was found using more traditional TFIDF and LSA representations which can be seen in Table 1. This could stem from the fact that the documents we are using are long, dense, and domain specific texts and finding a way to capture all the information contained in a single learned vector, required for distance metrics, is challenging. Even the better performing models struggled with picking up on certain associations easily identified by human readers.

TABLE I

ACCURACY FOR REPRESENTATION AND SIMILARITY ON DEPENDENCY

	TFIDF	LSA	LDA	Do2Vec	WordAvg
Manhattan	0.54	0.43	0.36	0.11	0.47
Cosine	**0.69**	0.49	0.45	0.15	0.65
Euclidean	0.48	0.48	0.04	0.11	0.44
Jaccard	0.01	0.01	0.01	0.01	0.01
KL Divergence	0.01	0.01	0.48	0.01	0.01
Hellinger	0.01	0.01	0.52	0.01	0.01

After selecting the candidate models, we expanded to using the entire corpus and tested their performances with varying preprocessing approaches. Table 2 displays the average precision and recall for different preprocessing approaches and TFIDF representation. In every case, unigram models performed better than bigram and trigram models. The average recall rate increase as the number of features in a model increases, but the best precision model used less features. RAKE, noun isolated, and all word unigram models were able to provide consistently high results with recall with no model performing significantly better than any other. However, noun models generally performed worse in precision.

TABLE II
AVERAGE RECALL AND PRECISION ON EXPERT RESULTS

#Features	n-gram	RAKE	POS	Recall	Precision
10k	2	No	All	0.1939	0.0780
10k	2	No	Noun	0.2232	0.1024
10k	3	No	All	0.1858	0.0732
10k	3	No	Noun	0.2102	0.0878
10k	1	Yes	All	0.4569	0.2049
10k	2	Yes	All	0.3291	0.1366
10k	1	No	All	0.4630	0.2000
4k	1	No	All	0.2610	0.2098
10k	1	No	Noun	0.4528	0.1171

In many cases, we were able to capture expected results with our system. In particular, for the expert supplied test data we could often times detect the majority of their suggestions. This can be seen in the first example shown in Table III. However, the models still had trouble picking out the correct results in certain situations which is evidenced by second example in that table. The examples in this table also highlight another trend in our results: some of the system retrieved similarities were not given by the expert, but were actually found to be related upon further inspection. In fact, this issue was also found when comparing to the pre-existing citations as well. This lack of clear ground truth could be a reason for the evaluation metrics being lower than expected, despite manual evaluations indicating a better performance.

TABLE III
SYSTEM AND EXPERT RESULTS FOR TOP WIKIPEDIA ARTICLES

Bisphosphonates for Paget's disease of bone in adults		**Brief interventions for heavy alcohol users admitted to general hospital wards**	
Expert	System	Expert	System
Paget's disease of bone	Bisphosphonate	Drug rehabilitation	Long-term effects of alcohol consumption
Bisphosphonate	Osteonecrosis of the jaw	Addiction	Alcohol and health
Alendronic acid	Alendronic acid	Alcohol intoxication	Alcohol abuse
	Paget's disease of bone	Alcohol withdrawal syndrome	Alcohol and cardiovascular disease
	C-terminal telopeptide		Alcohol and cancer

Results in the other direction, inputting a Wikipedia article and outputting the 10 most similar Cochrane reviews, tended to be difficult to evaluate. Table IV provides results showing Wikipedia articles as system input. Discerning these results has proven difficult with the only reference points being the pre-linked citations and our own evaluation, which has proven challenging given our lack of domain expertise. For the examples in Table IV, our evaluation has found that the returned Cochrane Reviews are similar to the Wikipedia articles they are suggested for.

TABLE IV
TOP 5 COCHRANE REVIEWS FOR GIVEN WIKIPEDIA ARTICLES

Water fluoridation	**Hypercoagulability in pregnancy**
Water fluoridation for the prevention of dental caries	Anticoagulant therapy for deep vein thrombosis (DVT) in pregnancy
Interventions to improve water quality for preventing diarrhea	Deflation of gastric band balloon in pregnancy for improving outcomes
Interventions to improve water quality and supply	Pharmacological interventions for generalized itching (not caused by systemic disease or skin lesions) in pregnancy
Surgical removal versus retention for the management of asymptomatic disease-free impacted wisdom teeth	Methods for administering subcutaneous heparin during pregnancy
Water for wound cleansing	Reduction of the number of fetuses for women with a multiple pregnancy

Despite challenges evaluating in some cases, there are many problems in our models that we were able to identify. One example, shown in Table V, of the system lacking in discerning judgement are the results on a Cochrane Review discussing vitamin C and its' use in treating the common cold. Wikipedia articles from Vitamin A and Vitamin E all showed up in the results despite the fact that the review never mentioned them. One probable cause was the fact that only unigram keywords were implemented in our methods. As the word "Vitamin" was the keyword for all of the Wikipedia articles from Vitamin A to E, they all qualified as the results for the Cochrane reviews about Vitamin C. When changing to incorporate bigrams into the model, we were indeed able to address this challenge. However, accuracy on other samples worsened which likely stemmed from the system deciding to prefer bigrams over unigrams in all documents. This pattern continued throughout different approaches, whenever we fixed the vitamin C issue, other results would be worse than expected. We believe the same problem also happened on other bigram or trigram medical terms, but were unable to solve this without undermining the majority of the results.

CONCLUSIONS

We believe that the large size of our corpus made it especially difficult to solve all problems using one model. Furthermore, evaluation was a difficulty that limited our ability to address challenges. In fact, for further research into this field we suggest usage of a more refined evaluation system using multiple Wikipedia editors to evaluate results in place of or in conjunction with the methods used here. To solve some lasting model challenges, further exploration of different document representations may be useful. Due to time and computational constraints, only Glove word averaging and Doc2Vec embeddings were explored, it is possible that other embedding methods may have proved more effective for this problem.

TABLE V

TOP 10 WIKIPEDIA ARTICLES FOR COCHRANE REVIEW: "VITAMIN C FOR PREVENTING AND TREATING THE COMMON COLD"

Unigrams	Bigrams
Vitamin C and the common cold	Vitamin C and the common cold
Vitamin C and the Common Cold (book)	Vitamin C and the Common Cold (book)
Vitamin C	Vitamin C
Vitamin D	Scurvy
Vitamer	Common cold
Retinol	Cold medicine
Vitamin A	Human coronavirus 229E
Vitamin A deficiency	Upper respiratory tract infection
Vitamin D deficiency	Autoschizis
Tocopherol	Dehydroascorbic acid

The ultimate goal for this research was to develop a system that could reduce the time needed for Wikipedia editors find high quality evidence and improve medical articles. The system that we have created, while facing some inherent problems, is capable of providing reasonable starting points for similar documents while being able to provide results efficiently. We believe this system provides a useable alternative to the keyword searching currently employed by the volunteers doing this work manually.

ACKNOWLEDGEMENTS

We would like to acknowledge the Cochrane team of volunteers including Camila Margarita Montesinos Guevara, Andrés Viteri García, Paula Zambrano. We would also like to acknowledge the Data Science Institute Wikimedian-in-Residence, Lane Rasberry.

REFERENCES

[1] J. M. Heilman and A. G. West, "Wikipedia and medicine: Quantifying readership, editors, and the significance of natural language," *J. Med. Internet Res.*, 2015.

[2] L. Huang, D. Milne, E. Frank, and I. H. Witten, "Learning a concept-based document similarity measure," *J. Am. Soc. Inf. Sci. Technol.*, vol. 63, no. 8, pp. 1593–1608, 2012.

[3] Y. Feng, E. Bagheri, F. Ensan, and J. Jovanovic, "The state of the art in semantic relatedness: a framework for comparison," *Knowledge Engineering Review*. pp. 1–30, 2017.

[4] B. Pincombe, "Comparison of Human and Latent Semantic Analysis (LSA) Judgements of Pairwise Document Similarities for a News Corpus," *Inf. Sci. (Ny).*, 2004.

[5] C. SaranyaJothi and D. T. D.Thenmozhi, "Machine Learning approach to Document Classification using Concept based Features," *Int. J. Comput. Appl.*, vol. 118, no. 20, pp. 33–36, 2015.

[6] P. Bruza, G. Zuccon, B. Koopman, L. Sitbon, and M. Lawley, "An evaluation of corpus-driven measures of medical concept similarity for information retrieval," 2012, p. 2439.

[7] D. R. Swanson, "Two medical literatures that are logically but not bibliographically connected," *J. Am. Soc. Inf. Sci.*, vol. 38, no. 4, pp. 228–233, 1987.

[8] J. Daniel and J. H. Martin, "N-gram Language Models," in *Speech and Language Processing*, 2018.

[9] D. Cutting, J. Kupiec, J. Pedersen, and P. Sibun, "A practical part-of-speech tagger," 2007, p. 133.

[10] S. Rose, D. Engel, N. Cramer, and W. Cowley, "Automatic Keyword Extraction from Individual Documents," in *Text Mining: Applications and Theory*, 2010.

[11] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manag.*, vol. 24, no. 5, pp. 513–523, 1988.

[12] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation David," *J. Mach. Learn. Res.*, vol. 3, no. Jan, pp. 993–1022, 2003.

[13] Q. Le and T. Mikolov, "Distributed Representations of Sentences and Documents," vol. 32, 2014.

[14] J. Pennington, R. Socher, and C. Manning, "Glove: Global Vectors for Word Representation," 2015.

[15] A. Huang, "Similarity measures for text document clustering," in *New Zealand Computer Science Research Student Conference (NZCSRSC)*, 2008.

AUTHOR INFORMATION

Jingnan Yang, Master of Science in Data Science (MSDS) Candidate, Data Science Institute, University of Virginia.

Justin Ward, MSDS Candidate, Data Science Institute, University of Virginia.

Erfaneh Gharavi, Research Assistant, Department of Engineering Systems & Environment, University of Virginia.

Jennifer Dawson, PhD, Research Associate. Wikipedia Consultant for Cochrane, Ottawa, Ontario, Canada.

Rafael Alvarado, PhD, Professor, Data Science Institute, University of Virgini