Impact of detecting clinical trial elements in exploration of COVID-19 literature

Simon Šuster
Fac. of Engineering and Information Tech.
The University of Melbourne
Melbourne, Australia
simon.suster@unimelb.edu.au

Karin Verspoor
School of Computing Technologies
RMIT University
Melbourne, Australia
karin.verspoor@rmit.edu.au

Timothy Baldwin
Fac. of Engineering and Information Tech.
The University of Melbourne
Melbourne, Australia
tbaldwin@unimelb.edu.au

Jey Han Lau
Fac. of Engineering and Information Tech.
The University of Melbourne
Melbourne, Australia
jeyhan.lau@unimelb.edu.au

Antonio Jimeno Yepes
Fac. of Engineering and Information Tech.
The University of Melbourne
Melbourne, Australia
antonio.jimeno@gmail.com

David Martinez Iraola
Independent researcher
Melbourne, Australia
david.martinez.iraola@gmail.com

Yulia Otmakhova
Fac. of Engineering and Information Tech.
The University of Melbourne
Melbourne, Australia
yotmakhova@student.unimelb.edu.au

Abstract—The COVID-19 pandemic has driven ever-greater demand for tools which enable efficient exploration of biomedical literature. Although semi-structured information resulting from concept recognition and detection of the defining elements of clinical trials (e.g. PICO criteria) has been commonly used to support literature search, the contributions of this abstraction remain poorly understood, especially in relation to text-based retrieval. In this study, we compare the results retrieved by a standard search engine with those filtered using clinically-relevant concepts and their relations. With analysis based on the annotations from the TREC-COVID shared task, we obtain quantitative as well as qualitative insights into characteristics of relational and concept-based literature exploration. Most importantly, we find that the relational concept selection filters the original retrieved collection in a way that decreases the proportion of unjudged documents and increases the precision, which means that the user is likely to be exposed to a larger number of relevant documents.

I. INTRODUCTION

The outbreak of Coronavirus disease 2019 (COVID-19) has led to a vigorous response from the global medical and AI communities, with efforts in the fields of information retrieval and natural language processing revolving around dataset construction and the development of tools for managing the growing literature on the virus and related diseases [1]. The release of the COVID-19 Open Research Dataset (CORD-19) [2] has stimulated the development of a large number of tools, reviewed in [3], which can be divided into more retrieval/QA-focused systems—returning a list of relevant documents for a user-specified query [4], [5], [6], [7], [8]—and those that use domain knowledge to organise and present information found in the literature [9], [10], [11], [12].

Broad clinically-relevant PICO categories derived from the structure of clinical trials [13] have been used to enable structured search as well as the visualisation of document content [10], [9], [14], [15], [11]. These categories describe the **P**atient population enrolled (e.g. *diabetics*), the **I**nterventions studied (e.g. *insulin*) and to what they were **C**ompared (e.g. *placebo*), and the **O**utcomes measured (e.g. *blood glucose levels*).

Here, we aim to examine the impact of PICO content structuring on literature exploration. Our analysis is based on the COVID-SEE system [11], [16] which offers both simple document retrieval and diverse visualisations of the retrieved document collection. We specifically focus on the information presented in the relational concept selection of COVID-SEE, which adopts Sankey diagrams to visually organise the medical concepts found in the articles according to PICO categories. The aim of presenting the retrieved documents in this way is to highlight salient P–I and I–O concept relations[1]. Figure 1 gives an example.

P–I and I–O relations often cannot be extracted from the retrieved documents, as not all articles are structured according to PICO and the identification of PICO elements is imperfect. This means that the relational concept selection shown in the visual summary is *implicitly filtering* a smaller subset of documents from the retrieval results. The question we seek to answer is how this affects the quality of the results. We approach this question by analysing the results through an

---

[1] The comparator ('C') category is usually merged with interventions ('I') due to their high similarity for the purposes of automatic PICO labelling [17]. We also follow this practice here.
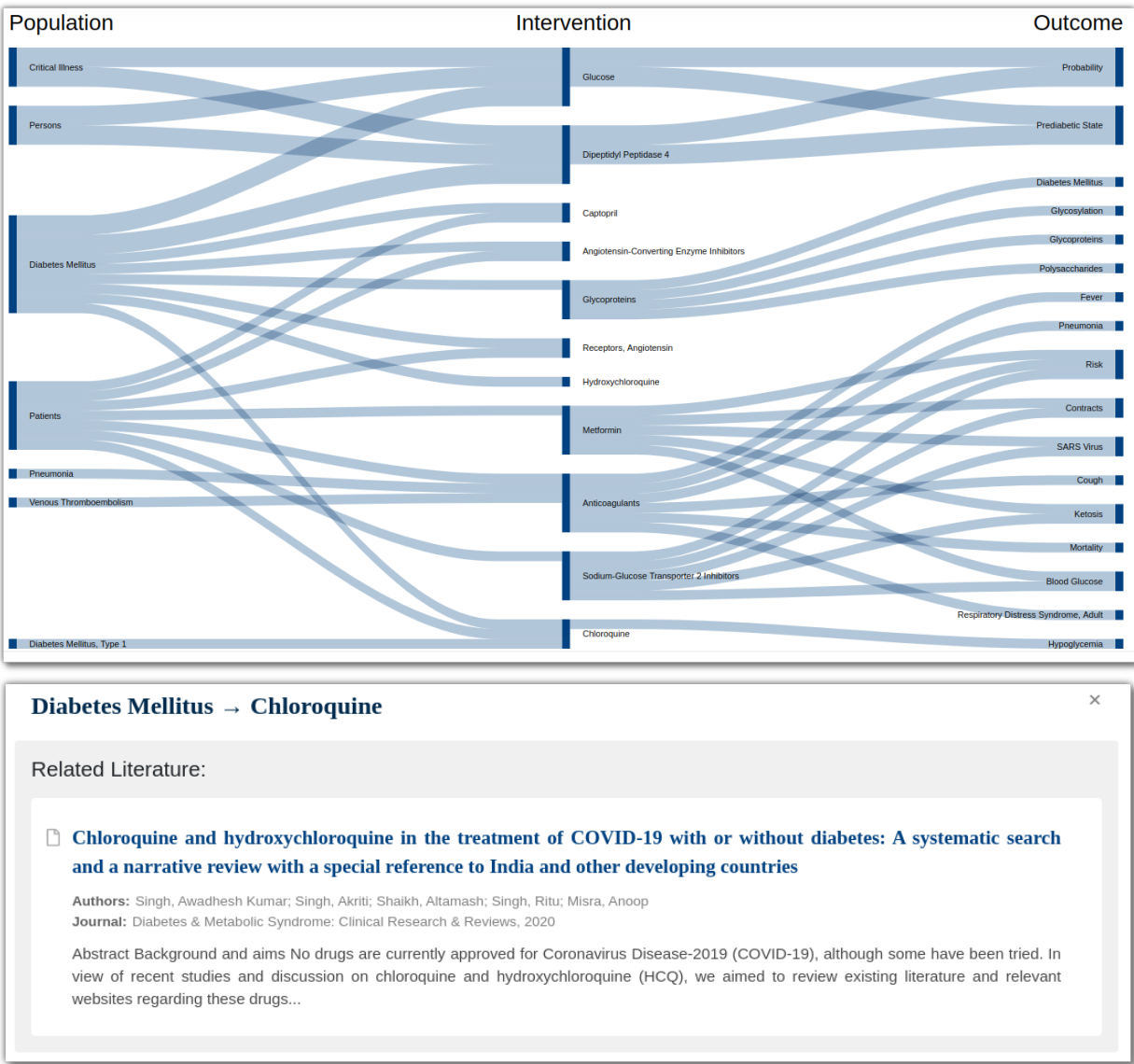
Fig. 1: Sankey diagram of PICO concepts and relations for 35 articles retrieved for *what kinds of complications related to COVID-19 are associated with diabetes* (TREC Round 5, query no. 24). Selecting a link (e.g. *Diabetes Mellitus → Chloroquine*) reveals papers that contain that relation (the displayed relation in the example only contains one document). Links are weighted by the number of documents containing the relevant relation.

existing test collection of relevance decisions, rather than carrying out a user study, which we leave for future work. To obtain a ground truth for relevance of the retrieved documents, we use the judgements for COVID-19-related literature obtained in the TREC-COVID shared task [18].

Our findings reveal the following:

- While the relational concept selection reduces the proportion of displayed documents to around 15% of the base retrieval results, the proportion of relevant documents actually *increases*. We find that this advantage comes from the fact that the relational concept selection is more likely to show the documents that received a judgment in the TREC-COVID test set (relevant or irrelevant).

- Based on manually rating a sample of unjudged documents, we find that most are irrelevant, suggesting that PICO filtering boosts the overall quality of the search results.

- Furthermore, when calculating precision (the proportion of relevant documents to all retrieved documents), we observe that the results vary substantially across queries. This leads us to infer that queries which are more amenable to PICO representation—based on a manual annotation of PICO elements in the queries—also tend to correlate with higher precision in the relational concept selection. In a similar vein, we find that the strength of a P–I/I–O relation (i.e. the number of articles contained in that particular

relation) is positively correlated with the proportion of relevant documents included. These results indicate that for medical queries with more easily identifiable PICO elements, the use of these elements as a central organising principle in the COVID-SEE system leads to users being exposed to more relevant results.

II. EMPIRICAL SETUP

A. Data

In our experiments, we use the CORD-19 dataset (v. 16 July 2020, totalling around 192,000 documents). It is currently the most extensive publicly available corpus of coronavirus-related publications from different sources, including PubMed's PMC open access corpus, research articles from a corpus maintained by the WHO, and bioRxiv and medRxiv pre-prints [2].

a) PICO-concept annotation: To annotate PICO elements and biomedical concepts found within PICO phrases, we follow a two-step procedure described in detail in [16], [11]. Briefly, in the first step, a BiLSTM-CRF model trained on the EBM-NLP dataset [17] is used to label textual spans with Population, Intervention, and Outcome categories. On the same dataset, the model's performance has been estimated at 0.78, 0.60 and 0.67 F1 for P, I and O categories, respectively. The next step consists of applying MetaMap [19] to identify the terms that correspond to Medical Subject Headings (MeSH) [20]. MeSH terms provide an established vocabulary for medical articles indexed by PubMed. We keep only those MeSH terms that are found within the PICO spans. The procedure is applied to all titles and abstracts in our collection.

b) TREC-COVID test set: TREC-COVID is a TREC retrieval task addressing literature retrieval related to the COVID-19 pandemic [18], [21]. The topics were developed based on searches submitted to the National Library of Medicine and suggestions from researchers on Twitter. They are representative of the high-level information needs related to the pandemic. Each topic consists of three fields of increasing granularity: a keyword-based query, a natural language question, and a longer descriptive narrative. We use the intermediate-level natural language questions as our queries. The first round of TREC-COVID introduced a set of 30 topics, and was based on the 10 April 2020 release of CORD-19. The most recent, fifth, round includes an additional set of 20 topics, and exploits the 16 July 2020 release of CORD-19.

The relevance judgments in TREC-COVID were assigned by individuals with biomedical expertise, based on a pooling approach in which only the top-ranked results from different submissions are assessed. A document is judged as 'relevant', 'partially relevant', or 'not relevant'. We consider the first two labels as *relevant* in our experiments.

B. Search and visualisation

In COVID-SEE, the CORD dataset is stored as a graph database (neo4j). Queries are executed against the full-text index over the titles and abstracts of the papers. We employ a very simple BM25 baseline which while not state of the art represents a standard (strong) baseline for IR tasks [22],

[23]. The search capability in neo4j is powered by the Apache Lucene [24], with a default VSM scoring function based on tf-idf. We limit the output of the IR search to 1000 hits.

The relational concept selection organises the identified MeSH concepts into PICO categories, and shows how they interact in a Sankey diagram (Figure 1). The diagram displays relations between pairs of concepts, where the relations 'carry' the corresponding documents, and the strength of a relation corresponds to the number of documents in which that concept pair is attested. The user interface for the relational concept selection supports varying levels of conceptual match. All our experiments involve MeSH terms of granularity 1, meaning that terms belonging to varying levels of the MeSH hierarchy get mapped to a common level, which is the one of the 16 topmost MeSH categories in combination with a tree descriptor. For example, *Canada* (MeSH tree number: Z01.107.567.176) is mapped (truncated) to *Geographic Locations* (Z01), and *Hydroxychloroquine* (D03.633.100.810.050.180.350) to *Heterocyclic Compounds* (D03).

C. Evaluation details

We evaluate the retrieval runs with the standard TREC script [25]. We report *precision*, based on the retrieved documents R that were either judged as relevant (*rel*) or irrelevant (*irrel*), or were not judged at all in the TREC test set (*unj*): $\frac{|R^{rel}|}{|R^{rel}|+|R^{irrel}|+|R^{unj}|}$. We additionally report a score that ignores the unjudged documents, referred to as *precision^{judg}*.

Note that there is no ranking in the relational concept selection, since documents are shown without ordering or scoring. Hence, we do not use evaluation metrics that are rank-dependent.

III. ANALYSIS

A. Relevance of documents in the relational concept selection

The relational concept selection implicitly acts as a filter for the documents retrieved by the IR engine, as some document abstracts do not contain any identifiable P–I or I–O pairs. Overall, the median percentage of documents (across all queries) containing PICO-typed concept relations after pre-processing is 13% (with the minimum and maximum for a given query being 8% and 28%, respectively). To understand what documents are selected and the retrieval quality of that selection, we carry out several experiments.

We first compare the precision scores of documents in the raw retrieval results vs. documents in the filtered collection, across all 50 queries, as detailed in Figure 2. The median score for documents in the relational concept selection is higher (0.17) than for the raw search results (0.12). The three queries with the highest precision are as follows, where the last query (28) achieves the highest improvement over raw search results (0.28 → 0.54):

(38) *What is the mechanism of inflammatory response and pathogenesis of COVID-19 cases?*

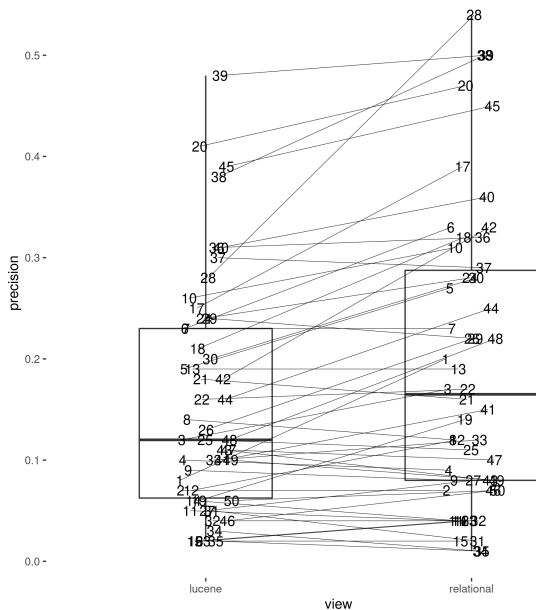(39) *What is the mechanism of cytokine storm syndrome on the COVID-19?*

Fig. 2: A box plot comparing the precision (section II-C) between the result sets retrieved for each query by the search engine (lucene) and the relational concept selection. The numbers represent the TREC-COVID query identifiers.

(28) *What evidence is there for the value of hydroxychloroquine in treating Covid-19?*

These queries can all be seen as precise statements about the required Problem/Population (*inflammatory response*, *cytokine storm syndrome*, *Covid-19*, respectively) and Intervention (*hydroxychloroquine* in query 28; the other two queries lack this criterion).

B. Role of unjudged documents

The calculation of precision (Section II-C) includes in the denominator the unjudged documents R^{unj}. Since we have little *a priori* intuition about the nature and the relevance of these documents, or whether presenting them to the user can be favourable or not, we manually analysed a sample of them, randomly selecting ten queries and picking for each the first 50 unjudged documents.[2] We then checked the abstract of each document to decide whether it is relevant to the query. The results are shown in Table I. Although some unjudged documents are relevant (cf. [26]), for all but one query the precision of the unjudged documents is far below the prior probability based on TREC-COVID judgements. We therefore infer that including fewer unjudged documents in the results should be favourable. This is in line with the view of the unjudged documents found in the literature [27].

We observe that the increased precision of results in the relational concept selection is largely due to including fewer unjudged documents: the proportion of retrieved unjudged documents is high in the raw search results (μ=0.74, min=0.5,

query	# rel.	P	P^{judg}	Δ
2	1	0.02	0.27	-0.25
11	3	0.06	0.17	-0.11
15	1	0.02	0.09	-0.07
18	10	0.20	0.89	-0.69
19	4	0.08	0.45	-0.37
22	13	0.26	0.56	-0.3
26	8	0.16	0.67	-0.51
34	8	0.16	0.06	0.10
38	20	0.40	0.85	-0.45
47	3	0.06	0.36	-0.30

TABLE I: Summary of manual annotations of relevance on top-50 unjudged documents from TREC-COVID for each query. P refers to our estimated precision among these unjudged documents in TREC-COVID; P^{judg} is the precision obtained on judged documents only (whether relevant or irrelevant, but excluding the unjudged ones); Δ quantifies the difference in the reported precision scores.

max=0.94, n=50), but somewhat lower in the relational selection (μ=0.65, min=0.4, max=0.9, n=50). In practice, this means that the user can expect to encounter—per 100 documents retrieved—9 unjudged documents less in the relational view compared to the IR list.

While the number of hits is on average close to 1000 for the search engine, the relational concept view typically captures only around 15% of the retrieved articles (μ=148 documents, min=78, max=278, n=50). The PICO view filters from the IR search list without any particular order, and in a way that is not guided by the original search ranking.

C. Exploring the relational concept selection

a) ***The nature of queries and their fit to the relational selection:*** Given the design of COVID-SEE in which *all* P- and I-typed concepts (as well as I- and O-typed concepts) are considered to be related if they co-occur within the same abstract without any further constraints, there is no guarantee that the P–I and I–O pairs truly correspond to the initial query. The query may not clearly state any PICO criteria, or the abstract may introduce PICO elements that are not mentioned in the query. Therefore, we would like to better understand what proportion of PICO-typed MeSH terms (or *nodes* of the graph) in the relational concept selection match the PICO criteria expressed in a query. Since the TREC-COVID queries are not annotated with PICO criteria or MeSH terms, we manually annotated all 30 queries from round 1 of the shared task with this information.

The results in Table II show that the percentage of nodes that match the query directly is generally low, on average around 7% for Population and Intervention, and 12% for Outcome. We find a moderate correlation between the proportions of Intervention concepts fitting the query and the precision scores (ρ=0.53), but the number of queries with a clearly stated Intervention is small (n=12).

b) ***Extent of grouping:*** We next explore to what extent the documents represented in the relational concept selection are grouped (pooled) together under the same relation, and

type	μ	min	max	n
P	7	1	14	20
I	7	3	17	12
O	12	3	18	4

TABLE II: Percentage of PICO-typed MeSH concepts included in the relation concept selection that are directly relevant to the query. μ is the mean over n queries (not all queries express PICO criteria).

whether more evidence of grouping suggests a higher precision or a better fit to the query. Analysis (not shown) reveals that: (a) most relations are sparse, only corresponding to a single document; (b) the ratio of relations representing more than one document varies between 0.17 and 0.47, depending on the query; and (c) for some queries, relations cover a large number of documents, e.g. for query 7, some relations contain more than 25 documents, whereas for query 8, fewer relations cover a large number of documents. The correlation between the number of documents in a relation and the precision scores for each relation (across all queries) reveals a slight positive correlation, with ρ=0.22 on 85,287 relations.

IV. FURTHER DISCUSSION & CONCLUSION

From this analysis, we can conclude that a query processing approach that directly relies on matching a PICO-based structured query, such as the semantic search available as an alternative search strategy in COVID-SEE as well as in SciSight [9] and DOC Search [14], is unlikely to be effective for the TREC-COVID queries. This reflects the fact that these queries are generally quite open-ended [21], and in many cases do not clearly follow the PICO structure. The finding is also consistent with previous research examining the impact of explicitly using defined PICO criteria in search strategies for systematic reviews, which found that the inclusion of some PICO elements, specifically Outcome, significantly reduced recall [28].

However, our analysis of information presented in the selection of documents based on PICO categories and medical concepts suggests some benefits of integrating PICO. Specifically, the user is more likely to be exposed to more relevant documents, and fewer unjudged documents, compared to the raw search results. Additionally, the user is more likely to find the PICO concepts related to the initial query in those relations that cover a larger number of articles. These stronger relations then tend to represent to a greater extent the articles relevant to the query. A limitation is that different queries express the PICO concepts in various degrees, which affects the selection of articles and their overall relevance.

REFERENCES

[1] M. Hutson, "Artificial-intelligence tools aim to tame the coronavirus literature," *Nature*, June 2020.

[2] L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Burdick, D. Eide, K. Funk, Y. Katsis, R. M. Kinney, Y. Li, Z. Liu, W. Merrill, P. Mooney, D. A. Murdick, D. Rishi, J. Sheehan, Z. Shen, B. Stilson, A. D. Wade, K. Wang, N. X. R. Wang, C. Wilhelm, B. Xie, D. M. Raymond, D. S. Weld, O. Etzioni, and S. Kohlmeier, "CORD-19: The COVID-19 open research dataset," in *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, (Online), Association for Computational Linguistics, July 2020.

[3] L. L. Wang and K. Lo, "Text mining approaches for dealing with the rapidly expanding literature on COVID-19," *Briefings in Bioinformatics*, 12 2020. bbaa296.

[4] *CORD-19 Search*, accessed February 11, 2021.

[5] *Discovid*, accessed February 11, 2021.

[6] *Covidex*, accessed February 11, 2021.

[7] *CORD-19 Search*, accessed February 11, 2021.

[8] *COVID Scholar*, accessed February 11, 2021.

[9] T. Hope, J. Portenoy, K. Vasan, J. Borchardt, E. Horvitz, D. S. Weld, M. A. Hearst, and J. D. West, "SciSight: Combining faceted navigation and research group detection for COVID-19 exploratory scientific search," *bioRxiv*, 2020.

[10] B. Nye, A. Nenkova, I. Marshall, and B. C. Wallace, "Trialstreamer: Mapping and browsing medical evidence in real-time," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, (Online), pp. 63–69, Association for Computational Linguistics, July 2020.

[11] K. Verspoor, S. Šuster, Y. Otmakhova, S. Mendis, Z. Zhai, B. Fang, J. H. Lau, T. Baldwin, A. Jimeno Yepes, and D. Martinez, "Brief description of covid-see: The scientific evidence explorer for covid-19 related research," in *Advances in Information Retrieval* (D. Hiemstra, M.-F. Moens, J. Mothe, R. Perego, M. Potthast, and F. Sebastiani, eds.), (Cham), pp. 559–564, Springer International Publishing, 2021.

[12] *COVID-19 Navigator*, accessed February 11, 2021.

[13] W. S. Richardson, M. C. Wilson, J. Nishikawa, R. S. Hayward, *et al.*, "The well-built clinical question: a key to evidence-based decisions," *Acp j club*, vol. 123, no. 3, pp. A12–3, 1995.

[14] *DOC Search*, accessed February 11, 2021.

[15] *COVID-19 Living OVerview of Evidence*, accessed February 11, 2021.

[16] K. Verspoor, S. Šuster, Y. Otmakhova, S. Mendis, Z. Zhai, B. Fang, J. H. Lau, T. Baldwin, A. J. Yepes, and D. Martinez, "Covid-see: Scientific evidence explorer for covid-19 related research," *arXiv preprint arXiv:2008.07880*, 2020.

[17] B. Nye, J. J. Li, R. Patel, Y. Yang, I. Marshall, A. Nenkova, and B. Wallace, "A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Melbourne, Australia), pp. 197–207, Association for Computational Linguistics, July 2018.

[18] K. Roberts, T. Alam, S. Bedrick, D. Demner-Fushman, K. Lo, I. Soboroff, E. Voorhees, L. L. Wang, and W. R. Hersh, "TREC-COVID: Rationale and Structure of an Information Retrieval Shared Task for COVID-19," *Journal of the American Medical Informatics Association*, 05 2020. ocaa091.

[19] *MetaMap - A Tool For Recognizing UMLS Concepts in Text*, accessed February 11, 2021.

[20] *MeSH (Medical Subject Headings)*, accessed February 11, 2021.

[21] E. Voorhees, T. Alam, S. Bedrick, D. Demner-Fushman, W. R. Hersh, K. Lo, K. Roberts, I. Soboroff, and L. L. Wang, "Trec-covid: Constructing a pandemic information retrieval test collection," *arXiv preprint arXiv:2005.04474*, 2020.

[22] C. Kamphuis, A. P. de Vries, L. Boytsov, and J. Lin, "Which bm25 do you mean? a large-scale reproducibility study of scoring variants," in *European Conference on Information Retrieval*, pp. 28–34, Springer, 2020.

[23] A. Trotman, A. Puurula, and B. Burgess, "Improvements to bm25 and language models examined," in *Proceedings of the 2014 Australasian Document Computing Symposium*, pp. 58–65, 2014.

[24] *Apache Lucene*, accessed February 11, 2021.

[25] *TREC evaluation script*, accessed February 11, 2021.

[26] J. Zobel, "How reliable are the results of large-scale information retrieval experiments?," in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, (New York, NY, USA), p. 307–314, Association for Computing Machinery, 1998.

[27] M. Sanderson, "Test collection based evaluation of information retrieval systems," in *Foundations and Trends in Information Retrieval*, pp. 247–375, 2010.

[28] T. F. Frandsen, M. F. Bruun Nielsen, C. L. Lindhardt, and M. B. Eriksen, "Using the full pico model as a search tool for systematic reviews resulted in lower recall for some pico elements," *Journal of Clinical Epidemiology*, vol. 127, pp. 69 – 75, 2020.