information

Article

Semi-Automatic Systematic Literature Reviews and Information Extraction of COVID-19 Scientific Evidence: Description and Preliminary Results of the COKE Project

Davide Golinelli [1], Andrea Giovanni Nuzzolese [2], Francesco Sanmarchi [1,*], Luana Bulla [2], Misael Mongiovì [3], Aldo Gangemi [2] and Paola Rucci [1]

[1] Department of Biomedical and Neuromotor Sciences (DIBINEM), Alma Mater Studiorum—University of Bologna, 40126 Bologna, Italy; davide.golinelli@unibo.it (D.G.); paola.rucci2@unibo.it (P.R.)
[2] STLab, Institute for Cognitive Sciences and Technologies (ISTC)-CNR, 00185 Roma, Italy; andreagiovanni.nuzzolese@cnr.it (A.G.N.); lua95@hotmail.it (L.B.); aldo.gangemi@cnr.it (A.G.)
[3] STLab, Institute for Cognitive Sciences and Technologies (ISTC)-CNR, via Gaifami 18, 95126 Catania, Italy; misael.mongiovi@cnr.it
* Correspondence: francesco.sanmarchi@gmail.com

Abstract: The COVID-19 pandemic highlighted the importance of validated and updated scientific information to help policy makers, healthcare professionals, and the public. The speed in disseminating reliable information and the subsequent guidelines and policy implementation are also essential to save as many lives as possible. Trustworthy guidelines should be based on a systematic evidence review which uses reproducible analytical methods to collect secondary data and analyse them. However, the guidelines' drafting process is time consuming and requires a great deal of resources. This paper aims to highlight the importance of accelerating and streamlining the extraction and synthesis of scientific evidence, specifically within the systematic review process. To do so, this paper describes the COKE (COVID-19 Knowledge Extraction framework for next generation discovery science) Project, which involves the use of machine reading and deep learning to design and implement a semi-automated system that supports and enhances the systematic literature review and guideline drafting processes. Specifically, we propose a framework for aiding in the literature selection and navigation process that employs natural language processing and clustering techniques for selecting and organizing the literature for human consultation, according to PICO (Population/Problem, Intervention, Comparison, and Outcome) elements. We show some preliminary results of the automatic classification of sentences on a dataset of abstracts related to COVID-19.

Keywords: natural language processing; health data; systematic reviews; guidelines; evidence-based medicine; COVID-19; SARS-CoV-2; machine reading; machine learning; artificial intelligence

Citation: Golinelli, D.; Nuzzolese, A.G.; Sanmarchi, F.; Bulla, L.; Mongiovì, M.; Gangemi, A.; Rucci, P. Semi-Automatic Systematic Literature Reviews and Information Extraction of COVID-19 Scientific Evidence: Description and Preliminary Results of the COKE Project. *Information* **2022**, *13*, 117. https://doi.org/10.3390/info13030117

Academic Editor: Haridimos Kondylakis

Received: 29 December 2021
Accepted: 21 February 2022
Published: 28 February 2022

1. Introduction

A new strain of Coronavirus (SARS-CoV-2) was first documented on 31 December 2019. Although more than two years have elapsed, the human knowledge about the Coronavirus Disease 19 (COVID-19) is still pervaded by uncertainty [1,2]. Some effective public health measures, prevention strategies, and therapeutics options are now available, but international, national, and subnational policies do not always follow shared guidelines, creating confusion and mistrust among the general population [3–5]. This issue is partially determined by a gap between data collection, scientific evidence, their validation, and the creation/modification of the guidelines [6]. The forced cohabitation with SARS-CoV-2 has made it even clearer that there is great importance in making validated and updated scientific information available to policy makers, healthcare professionals, and the public as quickly as possible. The speed in disseminating reliable and verified information, as well as the subsequent policy implementation based on that information, is also essential to save as

many lives as possible [7]. Currently, the most influential public health recommendations come from guidelines developed by international and national organizations (e.g., WHO, CDC, ECDC, etc.) [8–10]. The drafting of these guidelines is time consuming and requires a great deal of resources [11]. Moreover, specific circumstances, such as the COVID-19 pandemic, preclude the development of a standard guideline. This is the case when the WHO must provide guidance in response to public health emergencies. Depending on the type of event or situation, such guidelines may need to be produced within hours, days, weeks, or months.

In general, but even more so in an emergency, it is appropriate to draft and update the guidelines in a relatively short time. This is especially true considering that the scientific community generates an exponentially growing number of scientific papers [12], and that this trend accelerated during the COVID-19 pandemic [13].

Moreover, all of this is taking place in a context in which several new solutions in the management and extraction of information in healthcare are being introduced. For instance, the introduction of information extraction software is an important way of facilitating more sophisticated healthcare research [14]. Search engines, big data search, and mining tools are continuously being introduced in healthcare processes. The past decade has seen a truly revolutionary paradigm shift to Natural Language Processing (NLP) as a result of which Deep Learning (DL) became the dominating mind-set of researchers and developers in this field [15–17] and became an extremely robust and effective tool for adequately dealing with the contents of unstructured visual, audio/speech, and textual data.

1.1. Guidelines Development and Systematic Literature Reviews

A standard guideline covers a clinical or policy area (e.g., COVID-19). Health care guidelines and their appropriate implementation are of interest to national organizations, professional associations, health care providers, policymakers, patients, and the public [11]. Plenty of the literature illustrates in detail the guideline development, implementation, and evaluation process [11,18]. Standard guidelines vary greatly in scope and focus: they might address the use of a single drug for a disease or condition, such as naloxone injection by lay persons for suspected opioid overdose, or they might encompass the full scope of a condition or public health problem, such as the diagnosis, screening, and treatment of type two diabetes mellitus. Recommendations in a standard guideline are either developed de novo or by updating previous guidelines. Standard guidelines usually take between 9 and 24 months to complete, depending on their scope. They should be supported by one or more systematic reviews of the evidence and finalized after one or two meetings of the expert panel. Specific circumstances, such as the COVID-19 pandemic, preclude the development of a standard guideline. Providing clear evidence in situations such as this is inherently challenging, both for current and possible future pandemics. Firstly, this is because of the extremely narrow time window available in emergency situations. Secondly, it is because of human limits in cognitive skills, which, in turn, are core in the process of scientific discovery [19,20].

Systematic reviews are a type of review that uses reproducible analytical methods to collect secondary data and analyse it [21]. Systematic reviews differ from traditional literature reviews because they are more replicable and transparent [22,23]. As stated by Liberati et al. [23], systematic reviews are essential tools for summarizing evidence accurately and reliably. They help clinicians keep up-to-date; provide evidence for policy makers to weigh the risks, benefits, and harms of health care behaviors and interventions; gather and summarize related research for patients and healthcare providers; provide a starting point for clinical practice guideline developers; provide summaries of previous research for funders wishing to support new research; and help editors judge the merits of publishing reports of new studies. This makes them pivotal not only for scholars, but also for clinicians, researchers, policymakers, journalists, and, ultimately, the public [24,25]. They adopt a structured, transparent, and reproducible methodology that consists of the a priori specification of a research question; clarity on the scope of the review and on the

type of studies eligible for inclusion; making every effort to find all relevant research and to ensure that possible bias is accounted for; and analysing the included studies to draw conclusions based on all the identified research in an impartial and objective way [26].

Clinical studies and questions always either explicitly or implicitly contain four aspects: Population/Problem (P), Intervention (I), Comparison (C), and Outcome (O), which are known as PICO elements. Using this structure to guide the information retrieval of medical evidence within a medical citation database is popular and advantageous. However, accurately and efficiently extracting PICO elements from non-structured information, such as a collection of medical abstracts, is challenging.

Usually, systematic reviews are conducted by expert panels chosen mainly for their scientific and clinical reputation. Those who are excluded are individuals with clear financial conflicts and those whose professional or intellectual bias can diminish the credibility of the review. Once the research question, the inclusion and exclusion criteria, and the research protocol are defined, the panel experts perform the operative part of the review process (Figure 1). The researchers perform a comprehensive query in multiple databases using search strings and download the records' information (i.e., title, abstract, and other metadata such as the authors' names, journal name, and DOI). The researchers perform title and abstract screening based on the predetermined inclusion and exclusion criteria and on the PICO elements. In this phase, usually, the majority of the articles are discarded [21]. Subsequently, the researchers extract and read the full text of each individual record included in the new set of articles and eliminate irrelevant articles. Then, from the final set of articles, the researchers extract the information and evidence relevant for the research question.

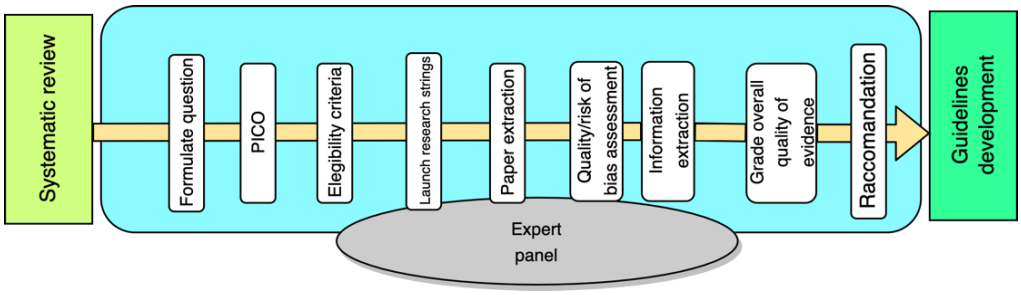

Figure 1. Guidelines development steps, including systematic literature review.

It is important to consider that, in the case of systematic reviews aimed at supporting guidelines development, it is necessary to evaluate the strength of the extracted scientific evidence. This process is usually completed through the GRADE system [27]. GRADE is a systematic approach to rating the certainty of evidence in systematic reviews and in other evidence syntheses. Specifically, the GRADE system is a transparent framework for developing and presenting summaries of evidence, and it provides a systematic approach for making clinical practice recommendations. This whole process is error-prone and inefficient (time-intensive) because of extremely imbalanced data: only a fraction of the screened studies is relevant. Moreover, the scientific literature is growing faster than the amount of time available for systematic reviews. Thus, adequate manual literature reviews become increasingly challenging at a point that, in certain research fields, they are becoming impossible [28].

Luckily, the systematic review process entails several explicit and, ideally, reproducible steps, including identifying all possibly relevant publications in a standardized way, extracting data from eligible studies, grading, and synthesizing the results. These characteristics make the process amenable to automation. This can make it much faster, more efficient, less error-prone, and actionable with the use of fewer resources.

A number of tools have been developed to streamline and semi-automate this process [29,30], including tools to generate terms, to visualize and evaluate search queries, to trace citation linkages, to deduplicate, limit, or translate searches across databases, and

to prioritize relevant abstracts for screening. Ongoing research is moving toward the development of tools that can unify searching and screening into a single step, and several prototype of such tools have already been implemented. A recent scoping review [30] identified several of these tools (LitSuggest, Rayyan, Abstractr, BIBOT, R software, RobotAnalyst, DistillerSR, ExaCT and NetMetaXL), which have the potential to be used for the automation of systematic reviews. However, these tools are not without limitations, as the majority of algorithms have not yet been developed into user-friendly tools. Some of these algorithms show high validity and reliability, but their use is conditional on user knowledge of computer science and algorithms. Furthermore, a living systematic literature review [29], which examined published approaches for data extraction from reports of clinical studies, showed that more than 90% of the publications evaluated developed classifiers that used randomized controlled trials (RCTs) as the main target texts, while only a small number of tools extracted data from observational studies. These are major limitations that should be addressed in order to implement automatic or semi-automatic support systems for experts who want to conduct systematic literature reviews on broad and mainly unknown topics, such as SARS-CoV-2/COVID-19.

1.2. Rationale and Aim

Methodological rigor and the need to rapidly produce solid scientific evidence are of utmost importance for the guideline drafting process. This is especially true during emergency situations such as the COVID-19 pandemic, in which every choice and its timing can affect the health outcomes of millions of people. Therefore, it seems necessary to conceptualize automated or semi-automated systems to support human action in the process of screening, extracting, and grading scientific evidence through the implementation of user-friendly tools that can support the expert panel during the systematic literature review process. Moreover, during an emergency situation, it is mandatory to include not only the abstract and text from RCTs, but also observational studies, which constitute a very important part of the body of the scientific literature.

Accordingly, this paper aims to highlight the importance of accelerating and streamlining the extraction and synthesis of scientific evidence in the biomedical field, specifically within the systematic literature review process. To do so, we organized this paper into two main parts. The first (Section 2) describes the COKE (COVID-19 Knowledge Extraction framework for next generation discovery science) Project's framework, which involves the use of machine reading [31] and deep learning [32] to semi-automate the workflow of systematic literature reviews in healthcare and, more specifically, which describes our proposed strategy for selecting and navigating the relevant literature starting from a set of abstracts obtained by interrogating scientific databases (e.g., EMBASE, Pubmed). The second part (Sections 3 and 4) reports the project's methods and preliminary results on the automatic classification of sentences into PICO elements on a dataset of abstracts related to COVID-19.

2. The COKE Project

The COKE Project, funded by the Italian Ministry for University and Research (MUR) in June 2021 and which started in November 2021, is a joint venture of the Italian National Council of Research (CNR) and the University of Bologna. This project aims to design and implement a semi-automated system that supports and enhances the systematic literature review and guideline drafting processes. Specifically, it aims to expedite the "development" phase (i.e., the systematic review), the "rating/grading" of the scientific evidence, and the extraction of relevant scientific knowledge, with a particular focus on the COVID-19 literature.

The COKE Project answers the following questions: how to automatically link scientific texts or data to state-of-the-art knowledge during an emergency situation or relating to an emergent disease/pathogen, such as COVID-19? How to adapt and exploit methods of reasoning, learning, and reconciling knowledge graphs to recommend serendipitous results

to researchers and policy makers? How to achieve a good level of quality in knowledge extraction performed by an AI on articles with different formats and based on different study designs if compared to manual research and content extraction performed by a domain expert in evidence-based medicine (EBM)?

The COKE Project consists of four consecutive phases: (i) the definition of potentially automatable "nodes" in the guidelines drafting workflow, specifically in the systematic literature review process; (ii) the design of a semi-automated system to make the process faster; (iii) semi-automated system testing on research questions related to the COVID-19 pandemic; and (iv) benchmark tests (human vs machine-supported human). At the time of writing, phases (iii) and (iv) of the COKE Project are ongoing.

The guidelines' drafting process is summarized in Figure 1 and in Section 1.1. The systematic literature review workflow has been refined over the years and involves a series of standardized steps that are already reported in Section 1.1. and Figure 2, some of which can be potentially automated.

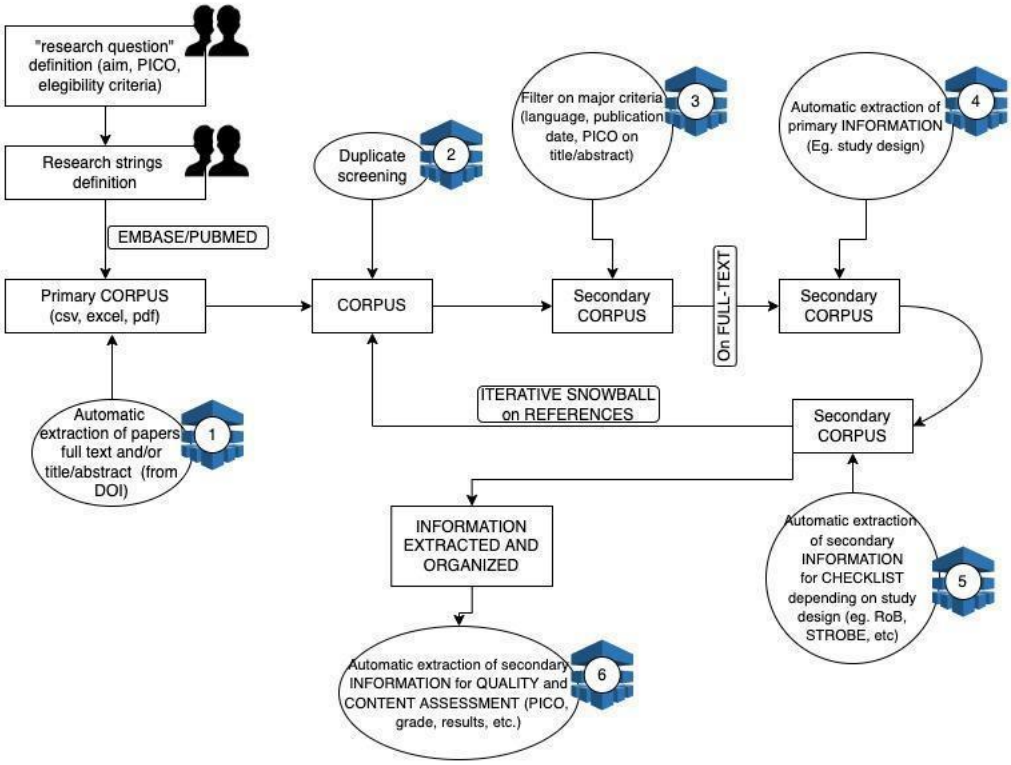

Figure 2. Systematic literature review workflow. The "blue icons" highlight the potentially automatable nodes of the evidence extraction process.

In the first phase of a systematic literature review, the researchers define the research question, the aim of the review, and the PICO and the eligibility criteria. This preparatory phase is usually human-driven. Afterwards, the researchers develop and define the research strings optimized for each search engine of the selected platforms/repository/libraries (e.g., MEDLINE, Embase). These queries generate the initial list of records that are to be screened and eventually analysed. In this phase it is possible to automate the full-text retrieval process and the duplicate record elimination process (nodes 1 and 2, Figure 2). The full-text retrieval is enabled by querying the Search API provided by Scopus. The deduplication process is performed by relying on DOIs associated with papers as unique identifiers. In case an article is not associated with a DOI, then our disambiguation approach uses the title and the list of authors for such a task.

Once the first list of records is obtained, the systematic review workflow provides a filter on some major criteria typically available in the metadata, such as language, publica-

tion date, and, above all, the PICO criteria. This screening process is usually carried out by the researchers analysing the title and abstract of each article and not on the entirety of the full text. This choice has been made to maintain high sensitivity and to save time without using too many resources. Most records are excluded in the title and abstract phase. Typically, only a small fraction of the records belongs to the relevant class, making title and abstract screening an important bottleneck in the systematic reviewing process (node 3, Figure 2) [33]. On one hand, the filtering based on metadata is fully automated. In fact, COKE performs this task by querying the Scopus API for retrieving all the required records that enable the selection. Among the various metadata, we also collect article level metrics, such as h-index and altmetrics. On the other hand, the selection based on PICO criteria is semi-automated. This is because a first step allows COKE to automatically identify candidate sentences in the title and the abstract of an article that evoke one or more of the PICO criteria, i.e., Participants/Problem (P), Intervention (I), Comparison (C) and Outcome (O). Those sentences are then annotated and presented to the expert for a manual assessment. We rely on the automatic annotation of sentences on a deep learning classifier based on three layers, namely BERT, a bi-LSTM layer, and a Conditional Random Field module, whose details are described in Section 3.

After the title and abstract screening process, the researchers perform the information-extraction process (i.e., study design, study quality and content analysis). During this phase, the researchers use several design-specific checklists (e.g., RoB 2 tool for randomized clinical trials, or ROBINS-I tool for non-randomized studies) to assess each study's risk of bias. This stage (node 5, Figure 2) can be automated by re-using solutions such as the RobotReviewer system [34,35] that relies on a Convolutional Neural Network (CNN) architecture trained on corpora derived from the Cochrane Database of Systematic Reviews (CDSR) for rating articles as having a "low", "high", or "unclear" risk of bias.

Finally, all the necessary information (i.e., PICO-related information, information relevant for the GRADE quality assessment of each study, and other information of interest) are extracted and organized. In this last phase (node 6, Figure 2) we construct a knowledge graph that organizes all the relevant knowledge extracted with selected articles through previous steps.

The last step, the so-called "snowball", is carried out manually by searching for relevant articles among the references of the studies included in the systematic literature review. This iterative phase is automatable too by using the entries in the reference lists of selected articles as seeds for querying Scopus again and by iteratively re-starting our process from the initial step.

Once the aforementioned steps are completed and the necessary information has been obtained, we proceed with the (manual/human) analysis of the evidence and the drafting of recommendations for the guidelines.

In summary, the COKE Project contributes to discovery science by defining a novel knowledge extraction layer in the framework of guidelines development that integrates machine reading [31] and the rigorous protocols of EBM. Machine reading is critical because it allows COKE to gather knowledge graphs directly from COVID-19-related scientific texts. Knowledge graphs are mathematical graphs representing factual, conceptual, and procedural knowledge in the form of triples (subject, predicate, object) defining binary relationships (via predicates) between entities (i.e., subject and object) [36]. Machine reading enables the generation of structured knowledge from text. An example of a state-of-the-art machine reader is FRED (http://wit.istc.cnr.it/stlab-tools/fred/ (accessed date 20 February 2022)) [34]. The advantage of FRED is that its generated knowledge graphs are centered upon events denoted in a text, its participants, and their links to existing background knowledge, thus opening the possibility of expanding the content of the text to its "interpretants" [37]. This is quite close to what human reading does, with the plus of being processable by a machine. The accuracy of extracted knowledge ranges between 0.75 and 0.90 depending on the kind of constructions found in a text, and it is represented according to logical patterns [38]. An example of a knowledge graph

produced by FRED is presented in Figure 3. The graph is formalized by using VerbNet (https://verbs.colorado.edu/verbnet/ (accessed date 20 February 2022)) [39] frames and roles. Additionally, taxonomy induction is performed, and named entities are linked to external knowledge graphs (e.g., DBpedia (https://www.dbpedia.org/ (accessed date 20 February 2022)) [40]).

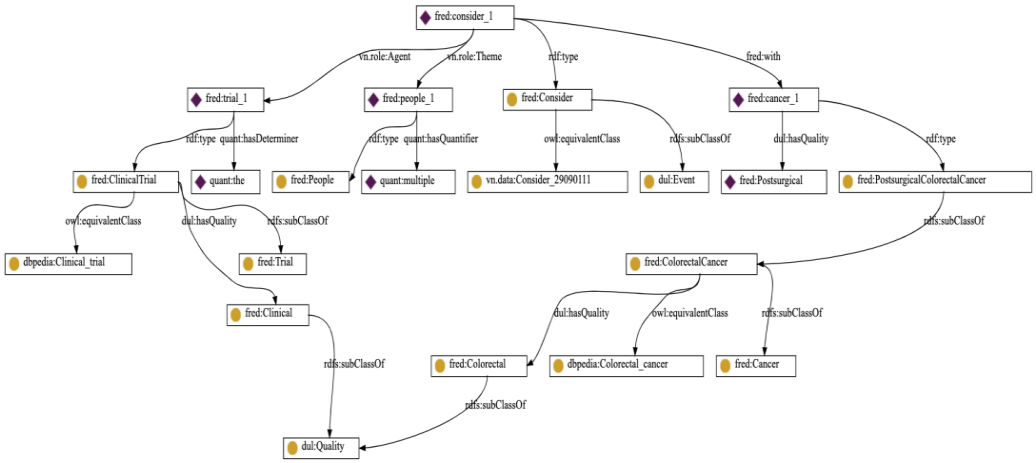

Figure 3. Example of a graph produced by FRED from the sentence, "The clinical trial considers people with postsurgical colorectal cancer".

Similarly, the analysis of procedures and practices of EBM as carried out by humans, e.g., systematic reviews, is of utmost importance for COKE. In fact, in the continuation of the COKE Project, FRED is to be used for producing knowledge graphs that are used for training a deep neural network architecture for identifying PICO- [41] or GRADE-related criteria for scientific literature selection and quality assessment.

Investigating those EBM procedures and practices is relevant for learning scientific discovery methodologies and patterns from experts by applying machine learning [42] approaches with human-in-the-loop and by converting those patterns into formal methods and schemata. Learned methodologies and patterns are to be applied on top of the knowledge graphs resulting from machine reading for automatically interpreting, querying, accessing, and inferring peculiar domain knowledge as gathered from the scientific literature. Finally, scientometrics tools, such as altmetrics [43], are useful for identifying research works that attract the attention of the community in quasi-zero-day time windows [44,45]. In fact, altmetrics are scholarly impact measures, which (i) accumulate faster than traditional metrics (i.e., h-index, citation count, impact factor) and (ii) cover a broader "impact flavor" being based on citations and activities in online tools and environments, such as social media or academic platforms such as Mendeley. Providing COKE with a layer for altmetric analysis will enable the framework to better address the information horizon problem. In fact, altmetric scores can be used as tools for filtering out works with limited scientific impact.

COKE is to be tested by performing and comparing a series of tasks in a "double blind" controlled trial between the tool and a team of researchers. More in detail, a COVID-19 research question is to be formulated. Both the COKE tool and the human researchers are to review the existing scientific literature and are to try to generate a satisfactory answer. The outputs are to be compared in terms of various metrics, such as: the number of papers identified, the quality assigned to each paper analysed, the interpretation of scientific evidence, and results and recommendations.

To ensure the quality of the COKE procedures, the output is to be compared to human performance in terms of the classification of study quality and the interpretation of study contents. In fact, benchmark testing is crucial to understand the real-world performance of any machine-learning-aided system, but such benchmark options are currently mostly lacking [46].

3. Methods

3.1. Organizing and Filtering the Relevant Literature Based on PICO Elements

A fundamental component of our framework concerns the organization of the literature to enable fast and user-friendly filtering of relevant articles (node 3, Figure 2). As previously reported, the medical literature has identified four key aspects in clinical studies, known as PICO elements. Since such information is often unavailable in a structured form and is described in the abstract text, organizing the retrieved literature based on PICO elements requires a machine understanding of such text.

Next we describe a framework for the organization and filtering task based on PICO elements. An overall description is reported in Figure 4.

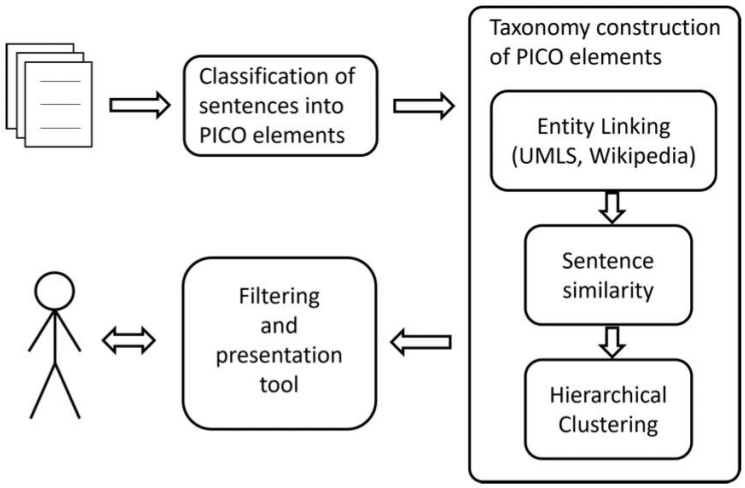

Figure 4. The proposed framework for organizing and filtering the relevant literature based on PICO elements.

Abstracts are first analysed to classify each sentence into PICO elements (e.g., the sentence "*Oral co-amoxiclav (@ mg/kg/day in three doses for @ days) or parenteral ceftriaxone (@ mg/kg/day in a single parenteral dose) for three days, followed by oral co-amoxiclav (@ mg/kg/day in three divided doses for seven days)*" for abstract #1319 was classified as I, i.e., Intervention). We then construct a taxonomy for each PICO element with the purpose of organizing the abstracts and enabling filtering. This step is performed separately for each PICO element after aggregating related sentences. It consists of three sub-steps. Clinical terms mentioned in the text are first linked to reference ontologies and vocabularies of clinical terms (e.g., the unified medical language system (UMLS) [47], SNOMED CT [48], etc.) and other resources (e.g., Wikidata [49]), to reduce ambiguity. The similarity of corresponding sentences across abstracts is then computed, and hierarchical clustering is performed. Eventually, the results are passed to a user presentation tool for filtering and navigation. We discuss each of the main steps in the remainder of this section.

3.2. Sentence Classification into PICO Elements

Automatic classification of the abstracts' sentences is the first step to solve the challenge mentioned above. To identify PICO elements in text, we adopt a linguistic model [50] composed of three main modules: sentence encoder, sentence contextualization, and label sequence optimization layer.

Given a sentence, the first module generates a vector from the sequence of tokens that compound it. To fulfill this task, the authors used a pre-trained language model called BERT [51]. Pre-trained on English Wikipedia and BooksCorpus, the model was fine-tuned on a large corpus that combined PubMed abstracts and PubMed Central (PMC) full-text articles in an unsupervised way [52] in order to learn the lexicon relevant to the medical

domain. The pre-training phase was followed by a supervised learning phase conducted on targeted downstream datasets [50].

The encoded sentences are subsequently processed by a bi-LSTM layer that contextualizes each vector with information taken from nearby sentences. The output obtained is processed by a feed-forward neural network with only one hidden layer, which returns the probability that a sentence belongs to each label.

Finally, the results obtained from the prediction are processed through a Conditional Random Field (CRF) module that optimizes the sequence of labels by modeling the dependencies between them.

3.3. Extracting PICO Elements Taxonomies

Once sentences are classified, the text related to each PICO element can be extracted and employed to organize the abstracts accordingly. For instance, a population-based organization would differentiate clinical studies on men from those on women or on children. The user might be interested in a finer partitioning, e.g., men with a specific disease or people in an age range. Therefore, we propose to hierarchically organize the abstracts according to each of the PICO elements extracted from the abstract text. To obtain meaningful results, it is crucial that the similarity function stresses the appropriate text features that refer to the population (or the specific PICO element) and disregards unrelated parts.

We propose to perform entity linking [51,52] to medical and general resources (e.g., UMLS, SNOMED CT, MeSH, Wikipedia) in order to (i) reduce the ambiguity of mentions, (ii) to identify the parts of the text that are more relevant for each PICO element, and (iii) to gather hierarchical information from external relevant ontologies and vocabularies. We plan to employ the available general [52] and specific [51] entity linking tools trained on annotated biomedical corpora [53]. We plan to develop a similarity function between entity linked text portions by combining semantic text similarity (e.g., Sentence-BERT) and similarity among linked entities. Finally, abstracts are hierarchically organized by clustering [54]. The result of these activities is a knowledge graph (KG) that provides structured knowledge about sentences gathered from abstracts with respect to PICO elements as well as clinical terms.

3.4. Filtering Tool

The goal of this component is reducing the cognitive load associated with the analysis of the scientific literature for carrying out systematic reviews by panels of experts. This goal is obtained in our framework by providing panels with a graphical solution that enables exploratory capabilities for interacting with the KG. A systematic review process can be classified as an exploratory search task. Exploratory search is extremely time consuming and cognitively complex, as it is typically associated with undefined and uncertain goals [55–57]. For example, a significant amount of literature is associated with "transmission of COVID-19", and it is inherently hard to identify evidence that can be used for defining guidelines for preventing COVID-19 spread. Hence, in our approach, KG exploration supports domain experts in browsing the literature and selecting relevant research works only.

We remind that selected works are then used by experts for finalizing a systematic review process. Filtering is performed by allowing a panel of domain experts to interact with the knowledge graph visually. Accordingly, the filtering tool allows a panel to explore the literature organized in the graph with respect to PICO elements and clinical terms. This allows the panel to select one or more concepts in the graph.

In the next Section we report the preliminary test's results of the COKE Project on the automatic classification of the COVID-19 abstracts.

4. Preliminary Results: A Case Study on the COVID-19 Literature

We provide the findings of a case study on the COVID-19 literature in evaluating the sentence classifier for PICO elements described in Section 3.2. The purpose is to

evaluate how the tool performs on a specific topic (COVID-19) with characteristics that are different from the dataset from the authors. We employ the model trained on the authors' dataset [41,50], a corpus composed by 24,668 abstracts of randomized controlled trials derived from Embase. The annotation of the dataset is carried out automatically through a keyword detection operation. This methodology streamlines the processing of large amounts of data, allowing for faster validation than manual execution. Hence, we assess how the model trained on such a dataset adapts to a specific topic that has not been covered during training.

We considered abstracts concerning the pre-exposure prophylaxis related to COVID-19 that was published within nine months in Embase, a popular free access database of medical articles. To enable manual validation, we randomly selected 50 abstracts and divided them into sentences, obtaining in total 752 sentences. The sentences were annotated by two domain experts. Following [45], we considered seven labels: Aim (A), Participants (P), Intervention (I), Outcome (O), Method (M), Results (R) and Conclusion (C), where Comparison was incorporated with the Intervention category. Since the labels Results (R) and Outcome (O) are similar in meaning and difficult to distinguish even for an expert, we collapsed them into a single label, Outcome, obtaining eventually six labels: A, P, I, O, M, and C. The annotators discarded short sentences corresponding to subsections (i.e., "Results:", "Conclusions:") and ambiguous elements, obtaining 584 annotated sentences (e.g., the sentence "PrEP-users identified convenience as a key benefit along with access to PrEP with reduced potential for COVID-19 exposure" of abstract #30 was annotated both manually and automatically as R, Results).

Table 1 summarizes the results for each label in terms of precision, recall and F1. Precision measures the percent of predictions with a given label that are correct, whereas recall represents the percent of true sentences with a given label that are correctly predicted. F1 is the harmonic mean between precision and recall, and therefore it can be considered a measure of the overall performance of the classifier. We also report for each label its support, i.e., the number of sentences annotated to that label. In the bottom we report the overall performances in terms of accuracy, i.e., the percent of correct predictions; macro average, i.e., the arithmetic average of the performance metrics; and weighted average, i.e., the average of the performance metrics weighted by support.

Table 1. Performance of the PICO classification (Section 3.2) on a dataset of 50 Embase abstracts related to the pre-exposure prophylaxis for COVID-19.

Label	Precision	Recall	F1	Support
A	0.884	0.850	0.867	153
C	0.796	0.854	0.824	96
I	0.312	0.625	0.417	8
M	0.750	0.543	0.630	94
O	0.748	0.821	0.782	184
P	0.479	0.469	0.474	49
Accuracy			0.757	584
Macro avg.	0.662	0.694	0.666	584
Weighted avg.	0.763	0.757	0.756	584

The results show a 76% overall accuracy of the classifier. Precision, recall, and F1 are above or attaining 75% for all labels except I, M, and P, which also have a lower support. As expected, the performances are slightly worse than those on the test set of [50]. This is due to two main factors. First, the annotation criteria of our dataset are different from those of the training set since our dataset has been manually annotated by experts, and the training dataset has been annotated automatically by considering the section headings of abstracts and mapping them to labels. Second, our dataset has different features since it refers to a specific topic (COVID-19) not covered by the training dataset and is not restricted to randomized clinical trials.

The results show that the employed classifier, trained on automatically annotated data, maintains adequate performances in predicting real (expert validated) labels even on abstracts of different topics and characteristics.

5. Conclusions

A rapidly changing healthcare requires fast decisions supported by sound scientific evidence. This is not compatible with the human limits in cognitive skills that reduce the ability to extract and process information. This paper describes such an urgent gap in the scientific field, specifically in the biomedical field, and a project that aims at filling it. Specifically, it aims to speed up the creation of healthcare guidelines and, in particular, systematic literature reviews, semi-automating parts of the workflow, and supporting the human-performed process of extracting and analysing contents.

In this paper we describe a framework for aiding in the systematic literature review process by analysing, organizing, and filtering medical abstracts. The tool we envision is based on NLP techniques able to detect and classify PICO elements and medical terms and organize abstracts accordingly. Preliminary results on the PICO element classification of abstract sentences show that a BERT + bi-LSTM language model trained on an automatically generated dataset performs adequately on a real case. Our proposed tool is expected to significantly reduce the effort for producing medical guidelines and therefore have a strong, positive impact, particularly in emergency scenarios. This paper also represents a call-to-action for similar initiatives, aimed at strengthening and enhancing the information and knowledge extraction process in medicine, which is particularly relevant in the fight against the current COVID-19 pandemic and other possible health crises.

Author Contributions: Conceptualization, D.G., A.G.N., F.S., P.R. and A.G.; methodology, A.G.N., L.B., M.M., A.G. and P.R.; software, A.G.N., L.B., M.M. and A.G.; formal analysis, A.G.N., L.B., M.M., A.G. and P.R.; investigation, D.G., A.G.N. and F.S.; resources, P.R., A.G.N. and A.G.; data curation, A.G.N., L.B., M.M. and A.G.; writing—original draft preparation, D.G., A.G.N., F.S. and L.B.; writing—review and editing, P.R., A.G. and M.M.; supervision, P.R. and A.G.; project administration, P.R. and A.G.; funding acquisition, P.R. and A.G. All authors have read and agreed to the published version of the manuscript.

Funding: The research leading to these results Project COKE (COVID-19 Knowledge Extraction framework for next—Piattaforma di estrazione della conoscenza basata su COVID19 per una nuova scienza della scoperta", CUP J55F21001990001—COKE_FISR2020IP_00621) has received funding by Ministero dell'Università e della Ricerca (Italian Ministry of University and Research, MUR) within the Fondo Integrativo Speciale per la Ricerca (FIRS) call 2020 (Decreto Direttoriale n. 562 del 05/05/2020)—PI Prof. Rucci.

Institutional Review Board Statement: This study did not require ethical approval.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Koffman, J.; Gross, J.; Etkind, S.N.; Selman, L. Uncertainty and COVID-19: How are we to respond? *J. R. Soc. Med.* **2020**, *113*, 211–216. [CrossRef] [PubMed]
2. Sanmarchi, F.; Golinelli, D.; Lenzi, J.; Esposito, F.; Capodici, A.; Reno, C.; Gibertoni, D. Exploring the gap between excess mortality and COVID-19 deaths in 67 countries. *JAMA Netw. Open* **2021**, *4*, e2117359. [CrossRef] [PubMed]
3. Hua, J.; Shaw, R. Corona virus (COVID-19) "infodemic" and emerging issues through a data lens: The case of China. *Int. J. Environ. Res. Public Health* **2020**, *17*, 2309. [CrossRef] [PubMed]
4. Bin Naeem, S.; Bhatti, R. The Covid-19 'infodemic': A new front for information professionals. *Health Inf. Libr. J.* **2020**, *37*, 233–239. [CrossRef] [PubMed]
5. The Lancet Infectious Diseases. The COVID-19 infodemic. *Lancet Infect. Dis.* **2020**, *20*, 875. [CrossRef]

6. Saperstein, Y.; Ong, S.Y.; Al-Bermani, T.; Park, J.; Saperstein, Y.; Olayinka, J.; Jaiman, A.; Winer, A.; Salifu, M.O.; McFarlane, S.I. COVID-19 guidelines changing faster than the virus: Implications of a clinical decision support app. *Int. J. Clin. Res. Trials* **2020**, *5*, 148. [CrossRef]
7. Gao, F.; Tao, L.; Huang, Y.; Shu, Z. Management and data sharing of COVID-19 pandemic information. *Biopreserv. Biobank.* **2020**, *18*, 570–580. [CrossRef]
8. WHO. Therapeutics and COVID-19: Living Guideline. Available online: https://www.who.int/publications/i/item/WHO-2019-nCoV-therapeutics-2021.1 (accessed on 31 March 2021).
9. Centers for Disease Control and Prevention. Interim Public Health Recommendations for Fully Vaccinated People. Available online: https://stacks.cdc.gov/view/cdc/105629 (accessed on 6 July 2021).
10. European Centre for Disease Prevention and Control; World Health Organization Regional Office for Europe. *Methods for the Detection and Characterisation of SARS-CoV-2 Variants*; ECDC; WHO Regional Office for Europe: Stockholm, Sweden; Copenhagen, Denmark, 2021.
11. Schünemann, H.J.; Wiercioch, W.; Etxeandia-Ikobaltzeta, I.; Falavigna, M.; Santesso, N.; Mustafa, R.; Ventresca, M.; Brignardello-Petersen, R.; Laisaar, K.-T.; Kowalski, S.; et al. Guidelines 2.0: Systematic development of a comprehensive checklist for a successful guideline enterprise. *Can. Med. Assoc. J.* **2014**, *186*, E123–E142. [CrossRef]
12. Boetto, E.; Golinelli, D.; Carullo, G.; Fantini, M.P. Frauds in scientific research and how to possibly overcome them. *J. Med. Ethics* **2021**, *47*, e19. [CrossRef]
13. Aviv-Reuven, S.; Rosenfeld, A. Publication patterns' changes due to the COVID-19 pandemic: A longitudinal and short-term scientometric analysis. *Scientometrics* **2021**, *126*, 6761–6784, published online ahead of print. [CrossRef]
14. Malmasi, S.; Hosomura, N.; Chang, L.-S.; Brown, C.J.; Skentzos, S.; Turchin, A. Extracting healthcare quality information from unstructured data. *AMIA Annu. Symp. Proc.* **2018**, *2017*, 1243–1252. [PubMed]
15. Hahn, U.; Oleynik, M. Medical information extraction in the age of deep learning. *Yearb. Med. Inform.* **2020**, *29*, 208–220. [CrossRef] [PubMed]
16. Wang, L.; Wang, J.; Wang, M.; Li, Y.; Liang, Y.; Xu, D. Using internet search engines to obtain medical information: A comparative study. *J. Med. Internet Res.* **2012**, *14*, e74. [CrossRef] [PubMed]
17. Dash, S.; Shakyawar, S.K.; Sharma, M.; Kaushik, S. Big data in healthcare: Management, analysis and future prospects. *J. Big Data* **2019**, *6*, 54. [CrossRef]
18. World Health Organization. *WHO Handbook for Guideline Development*, 2nd ed.; World Health Organization: Geneva, Switzerland, 2014.
19. Kitano, H. Artificial intelligence to win the Nobel prize and beyond: Creating the engine for scientific discovery. *AI Mag.* **2016**, *37*, 39–49. [CrossRef]
20. Nature Index. Available online: https://www.natureindex.com/news-blog/the-top-coronavirus-research-articles-by-metrics (accessed on 23 June 2020).
21. Khan, K.S.; Kunz, R.; Kleijnen, J.; Antes, G. Five steps to conducting a systematic review. *J. R. Soc. Med.* **2003**, *96*, 118–121. [CrossRef]
22. Cooper, H. *Research Synthesis and Meta-Analysis: A Step-by-Step Approach*; SAGE Publications: Thousand Oaks, CA, USA, 2015.
23. Liberati, A.; Altman, D.G.; Tetzlaff, J.; Mulrow, C.; Gøtzsche, P.C.; Ioannidis, J.P.A.; Clarke, M.; Devereaux, P.J.; Kleijnen, J.; Moher, D. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *J. Clin. Epidemiol.* **2009**, *62*, e1–e34. [CrossRef]
24. Boaz, A.; Ashby, D.; Young, K. *Systematic Reviews: What Have They Got to Offer Evidence Based Policy and Practice?* ESRC UK Centre for Evidence Based Policy and Practice: London, UK, 2002.
25. Oliver, S.; Dickson, K.; Bangpan, M. *Systematic Reviews: Making Them Policy Relevant. A Briefing for Policy Makers and Systematic Reviewers*; UCL Institute of Education: London, UK, 2015.
26. Booth, A. Searching for qualitative research for inclusion in systematic reviews: A structured methodological review. *Syst. Rev.* **2016**, *5*, 74. [CrossRef]
27. Guyatt, G.; Oxman, A.D.; Akl, E.A.; Kunz, R.; Vist, G.; Brozek, J.; Norris, S.; Falck-Ytter, Y.; Glasziou, P.; DeBeer, H.; et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. *J. Clin. Epidemiol.* **2011**, *64*, 383–394. [CrossRef]
28. Marshall, I.J.; Wallace, B.C. Toward systematic review automation: A practical guide to using machine learning tools in research synthesis. *Syst. Rev.* **2019**, *8*, 163. [CrossRef]
29. Schmidt, L.; Olorisade, B.K.; McGuinness, L.A.; Thomas, J.; Higgins, J.P.T. Data extraction methods for systematic review (semi)automation: A living systematic review. *F1000Research* **2021**, *10*, 401. [CrossRef] [PubMed]
30. Khalil, H.; Ameen, D.; Zarnegar, A. Tools to support the automation of systematic reviews: A scoping review. *J. Clin. Epidemiol.* **2021**, *144*, 22–42. [CrossRef]
31. Etzioni, O.; Banko, M.; Cafarella, M.J. *Machine Reading*; Association for the Advancement of Artificial Intelligence: Palo Alto, CA, USA, 2006; Volume 6, pp. 1517–1519.
32. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
33. Borah, R.; Brown, A.; Capers, P.L.; Kaiser, K.A. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ Open* **2017**, *7*, e012545. [CrossRef] [PubMed]

34. Gangemi, A.; Presutti, V.; Recupero, D.R.; Nuzzolese, A.G.; Draicchio, F.; Mongiovì, M. Semantic web machine reading with FRED. *Semant. Web* **2017**, *8*, 873–893. [CrossRef]
35. Marshall, I.J.; Kuiper, J.; Wallace, B.C. RobotReviewer: Evaluation of a system for automatically assessing bias in clinical trials. *J. Am. Med. Inform. Assoc.* **2015**, *23*, 193–201. [CrossRef]
36. Hogan, A.; Blomqvist, E.; Cochez, M.; d'Amato, C.; de Melo, G.; Gutierrez, C.; Kirrane, S.; Labra Gayo, J.E.; Navigli, R.; Neumaier, R.; et al. *Knowledge Graphs. Synthesis Lectures on Data, Semantics, and Knowledge*; Morgan and Claypool Publishers: San Rafael, CA, USA, 2021; Volume 12, pp. 1–257.
37. Eco, U. Peirce's Notion of Interpretant. *Comp. Lit.* **1976**, *91*, 1457–1472. [CrossRef]
38. Gangemi, A.; Recupero, D.R.; Mongiovì, M.; Nuzzolese, A.G.; Presutti, V. Identifying motifs for evaluating open knowledge extraction on the web. *Knowl. Based Syst.* **2016**, *108*, 33–41. [CrossRef]
39. Schuler, K.K. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*; University of Pennsylvania: Philadelphia, PA, USA, 2005.
40. Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, G.; Lehmann, J.; Cyganiak, R.; Ives, Z. *Dbpedia: A Nucleus for a Web of Open Data*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 722–735.
41. Jin, D.; Szolovits, P. Pico element detection in medical text via long short-term memory neural networks. In Proceedings of the BioNLP Workshop, Melbourne, Australia, 19 July 2018; pp. 67–75.
42. Džeroski, S.; Kocev, D.; Panov, P. Special issue on discovery science. *Mach. Learn.* **2016**, *105*, 1–12. [CrossRef]
43. Priem, J.; Groth, P.; Taraborelli, D. The altmetrics collection. *PLoS ONE* **2012**, *7*, 48753. [CrossRef]
44. Nuzzolese, A.G.; Ciancarini, P.; Gangemi, A.; Peroni, S.; Poggi, F.; Presutti, V. Do altmetrics work for assessing research quality? *Scientometrics* **2019**, *118*, 539–562. [CrossRef]
45. Boetto, E.; Fantini, M.P.; Gangemi, A.; Golinelli, D.; Greco, M.; Nuzzolese, A.G.; Presutti, V.; Rallo, F. Using altmetrics for detecting impactful research in quasi-zero-day time-windows: The case of COVID-19. *Scientometrics* **2021**, *126*, 1189–1215. [CrossRef] [PubMed]
46. Van de Schoot, R.; de Bruin, J.; Schram, R.; Zahedi, P.; de Boer, J.; Weijdema, F.; Kramer, B.; Huijts, M.; Hoogerwerf, M.; Ferdinands, G.; et al. An open source machine learning framework for efficient and transparent systematic reviews. *Nat. Mach. Intell.* **2021**, *3*, 125–133. [CrossRef]
47. Bodenreider, O. The unified medical language system (UMLS): Integrating biomedical terminology. *Nucleic Acids Res.* **2004**, *32*, D267–D270. [CrossRef]
48. Donnelly, K. SNOMED-CT: The advanced terminology and coding system for eHealth. *Stud. Health Technol. Inform.* **2006**, *121*, 279.
49. Vrandečić, D.; Krötzsch, M. Wikidata: A free collaborative knowledgebase. *Commun. ACM* **2014**, *57*, 78–85. [CrossRef]
50. Jin, D.; Szolovits, P. Advancing PICO element detection in biomedical text via deep neural networks. *Bioinformatics* **2020**, *36*, 3856–3862. [CrossRef]
51. Devlin, J.; Chang, M.-W.; Lee, K.-t.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; Volume 1, pp. 4171–4186.
52. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; Ho So, C.; Kang, J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2019**, *36*, 1234–1240. [CrossRef]
53. Fraser, K.C.; Nejadgholi, I.; De Bruijn, B.; Li, M.; LaPlante, A.; El Abidine, K.Z. Extracting UMLS concepts from medical text using general and domain-specific deep learning models. *arXiv* **2019**, arXiv:1910.01274.
54. Raiman, J.; Raiman, O. Deeptype: Multilingual entity linking by neural type system evolution. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
55. Mohan, S.; Li, D. Medmentions: A large biomedical corpus annotated with UMLS concepts. In Proceedings of the Automated Knowledge Base Construction (AKBC), Amherst, MA, USA, 20–21 May 2019.
56. Murtagh, F.; Contreras, P. Algorithms for hierarchical clustering: An overview. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2012**, *2*, 86–97. [CrossRef]
57. White, R.W.; Kules, B.; Bederson, B. Exploratory search interfaces: Categorization, clustering and beyond: Report on the xsi 2005 workshop at the human-computer interaction laboratory, University of Maryland. *ACM SIGIR Forum.* **2005**, *39*, 52–56. [CrossRef]