Article

Trends and Features of the Applications of Natural Language Processing Techniques for Clinical Trials Text Analysis

Xieling Chen [1]**, Haoran Xie** [2]**, Gary Cheng** [1]**, Leonard K. M. Poon** [1]**, Mingming Leng** [2] **and Fu Lee Wang** [3,*]

1   Department of Mathematics and Information Technology, The Education University of Hong Kong, Tai Po, New Territories 999077, Hong Kong, China; xielingchen0708@gmail.com (X.C.); chengks@eduhk.hk (G.C.); kmpoon@eduhk.hk (L.K.M.P.)
2   Department of Computing and Decision Sciences, Lingnan University, Tuen Mun, New Territories 999077, Hong Kong, China; hrxie@ln.edu.hk (H.X.); mmleng@ln.edu.hk (M.L.)
3   School of Science and Technology, The Open University of Hong Kong, Ho Man Tin, Kowloon 999077, Hong Kong, China
*   Correspondence: pwang@ouhk.edu.hk

check for updates

Abstract: Natural language processing (NLP) is an effective tool for generating structured information from unstructured data, the one that is commonly found in clinical trial texts. Such interdisciplinary research has gradually grown into a flourishing research field with accumulated scientific outputs available. In this study, bibliographical data collected from Web of Science, PubMed, and Scopus databases from 2001 to 2018 had been investigated with the use of three prominent methods, including performance analysis, science mapping, and, particularly, an automatic text analysis approach named structural topic modeling. Topical trend visualization and test analysis were further employed to quantify the effects of the year of publication on topic proportions. Topical diverse distributions across prolific countries/regions and institutions were also visualized and compared. In addition, scientific collaborations between countries/regions, institutions, and authors were also explored using social network analysis. The findings obtained were essential for facilitating the development of the NLP-enhanced clinical trial texts processing, boosting scientific and technological NLP-enhanced clinical trial research, and facilitating inter-country/region and inter-institution collaborations.

Keywords: natural language processing; clinical trials text; bibliometrics; collaboration; structural topic modeling

1. Introduction

The adoption of natural language processing (NLP) techniques for clinical information processing has been popular for decades. Previous reviews on the NLP-enhanced clinical trial research commonly adopted manual methods like systematic literature reviews, the results of which might fail to be generalized to the whole community. Thus, it is essential to offer a thorough overview of the NLP-enhanced clinical trial studies in order to inform scholars, policymakers, and clinicians in conducting relevant research and practical activities. To review studies on a large scale, we employed bibliometrics, science mapping, and structural topic modeling (STM) [1,2] to provide a comprehensive investigation of the academic literature worldwide concerning the NLP-enhanced clinical trials research. The remainder of this section is structured as follows. Section 1.1 presents an introduction to the NLP-enhanced clinical trial research. An introduction to bibliometrics, science mapping, and topic modeling is presented in Section 1.2. Section 1.3 depicts the research aims of this study.

1.1. Introduction to the NLP-Enhanced Clinical Trials Research

Clinical trial texts involve valuable information for medical and clinical research, contributing to the solution of problems regarding the quality of healthcare and clinical decision support. However, usually, a significant portion of essential clinical trial information is documented and stored within unstructured texts, making it challenging to effectively and precisely extract useful information. Furthermore, it can be time-consuming and error-prone to convert such unstructured texts into structured ones, which may fail to capture multifarious information [3]. Thus, the extraction of appropriate features from clinical narratives consists of unlocking the hidden knowledge, as well as enabling the conduction of advanced reasoning tasks, for example, diagnosis explanation, disease progression modeling, and treatment effectiveness analytics [4]. There are at least two types of motivations worthy of being highlighted for the conversion of unstructured texts into structured ones [5], including reducing the time of manual screening and enhancing the reuse of such data to realize the automatic processing of large-scale texts.

As the production of unstructured clinical trial data is rapid and on a large scale, it is of great need to mine such textual data and to further generate structured representations through automatic approaches by adopting NLP techniques. NLP concerns natural language human–machine interactions, particularly in terms of enabling machines to process and analyze large-scale natural language data. Applications of NLP have long been widely considered among scholars (e.g., [6–8]). As a computer-driven approach to facilitate human language processing [9], the adoption of the NLP techniques for clinical information processing has been popular for decades. There are several effective applications available in our daily life, for example, MedLEE [10]. However, clinical NLP, also referring to the use of NLP techniques that are developed and adopted for supporting healthcare through unlocking clinical information hidden within clinical narratives, is still an emerging technique [11]. Although medical documents in forms of plain texts contain essential information about patients, they are seldom available for structured queries. Besides the structured data, the provision of structured information extracted from natural language documents such as electronic health records (EHRs) [12–14] is expected to facilitate clinical trial research [15]. EHRs offer chances to improve the healthcare of patients, incorporate performance measurements in clinical practice, and assist clinical research [16]. EHR systems offer significant potentials for supporting quality improvement and clinical research. In addition, to fully take advantage of the electronic medical records (EMRs), scalable methods like NLP are needed to transform unstructured text to structured data that can be adopted for evaluation or analysis [17]. To that end, NLP has increasingly become a valuable and effective technique for processing clinical trial texts, most of which are unstructured and large-scale.

Currently, several reviews on the NLP-enhanced clinical trial research and its relevant topics are available, as summarized in Table 1. For example, based on 43 studies concerning the NLP of clinical notes on chronic diseases, Sheikhalishahi et al. (2019) [18] conducted a systematic review to explore the development and uptake of NLP approaches adopted to deal with textual clinical notes data on chronic diseases. The existing relevant reviews were conducted mainly by systematic reviews, suffering from the following disadvantages. First, as most studies used manual coding methods, the numbers of reviewed articles were relatively limited. Second, manual analyses might be error-prone as they involved a tedious and laborious coding process. In addition, the results of a small sample of the NLP-enhanced clinical trial studies could not be generalized to the whole community. To that end, this study analyzed all relevant studies regarding the NLP-enhanced clinical trial research to provide a thorough and objective understanding of the entire community by using bibliometrics and topic modeling.

Table 1. Recent reviews on the natural language processing (NLP)-enhanced clinical trial research and its relevant topics.

Reviewer(s) and Year	No. of Articles	Methods	Period	Research Goals
Sheikhalishahi et al. (2019) [18]	43	Systematic review	2007–2018	To thoroughly review the development and uptake of the applications of NLP in dealing with unstructured clinical notes data regarding chronic diseases
Dreisbach et al. (2019) [19]	21	Systematic review	2003–2017	To summarize studies concerning the applications of NLP and text mining to extract and process symptom information contained in electronic patient-authored texts.
Demner-Fushman & Elhadad (2016) [11]	Less than 100	Systematic review	Till 2016	To summarize studies concerning the applications of NLP in processing clinical and consumer-produced texts.
Gonzalez-Hernandez et al. (2017) [20]	149	Systematic review	Till 2017	To summarize studies concerning the applications of NLP for processing health-associated information contained in EHRs and social media posts.
Névéol & Zweigenbaum (2015) [21]	154	Systematic review	2014	To review and select the best papers concerning clinical NLP.
Luo et al. (2017) [22]	48	Systematic review	2000–2015	To summarize the state-of-the-art in methodology advances of NLP for EHRs-based pharmacovigilance.
Koleck et al. (2019) [23]	27	Systematic review	1999–2017	To summarize studies concerning the applications of NLP for processing and analyzing symptom information contained in EHRs narratives.
Kreimeyer et al. (2017) [5]	86	Systematic review	2006–2016	To summarize current-in-use NLP solutions for clinical texts that could encode free textual data into standardized clinical terminologies and to capture common data elements to fill specified templates.
Velupillai et al. (2015) [24]	Less than 100	Systematic review	2008–2014	To summarize the latest advances in clinical NLP, particularly focusing on semantic analysis and its supporting subtasks.

1.2. Introduction to Bibliometrics, Science Mapping, and Topic Modeling

Bibliometrics involves a set of approaches or techniques for measuring research outputs using scientific publications indexed in big bibliographic databases [25]. With the sustainable increasing number of scientific papers being published in different research fields, bibliometrics has grown into a valuable method aiming to progress the assessment of scholarly quality and productivity. The contributions of bibliometrics to science progress mainly involve several aspects, for example, enabling the assessment of progress, recognizing groups of relevant domain scientists, recognizing reliable publication sources, serving as an academic foundation for evaluation of new developments, assisting the implementation of specific intelligent information systems, and proposing novel indexes to evaluate scientific output [26–30].

In the bibliometric analysis, performance analysis and science mapping are two major approaches for exploring a specific research area [31,32]. Performance analysis concerns the assessment of the impact of academic outputs of researchers, institutions, or countries. Science mapping focuses on displaying the conceptual structures of particular research fields and their evolutions across time using network analysis, for example, co-words, co-citations, or co-authorship [30]. Performance bibliometric analysis concerns the measurement of scientific outputs using quality and quantity indicators like article count, citation count, and Hirsch index (H-index). Science mapping analysis serves as a spatial representation of the correlation of different areas. It aims to monitor an academic area and to delimit research areas for the determination of its structure and evolution [33].

In addition, topic modeling generates estimations of a list of topics that are hidden within a collection of textual data [34]. STM is a newly proposed unsupervised machine learning approach for analyzing text corpora on a large scale [1,2,35,36]. Topic models use a set of machine learning algorithms and a type of dimension reduction approach to extract latent topics from data. Topic models identify the co-occurrence of terms to further sort them into three latent topics. Each topic comprises terms, each with different assigned probabilities. Additionally, topic models assume that each document comprises different latent topics, each with different assigned probabilities in each document [37].

Bibliometric analysis and topic modeling have been popularly and widely adopted to present a general overview of various research fields (e.g., [38–41]). Some representative works include bibliometric analysis on, e.g., text mining-enhanced medical research [42], research that utilize AI techniques in processing EHRs data [43], and research of event detection using social media data [44].

1.3. Research Aims

This study aimed to provide a comprehensive overview of the NLP-enhanced clinical trial research. The bibliometric approach suits the purpose. Presently, no bibliometric study regarding scientific outputs on such an interdisciplinary research area is available. Thus, the adoption of science mapping techniques to evaluate the research area is of great need and importance. To that end, in this study, the first in-depth science mapping analysis, as well as an STM analysis of the NLP-enhanced clinical trial research, was conducted, especially for disclosing its structure and evolution. Specifically, based on the full bibliographic information of 451 articles regarding the NLP-enhanced clinical trial research retrieved from Web of Science (WoS), PubMed, and Scopus, this study conducted analyses in seven aspects. First, visualize the annual trend of the NLP-enhanced clinical trial articles. Second, identify the leading publication sources, countries/regions, institutions, and authors for the publication of the NLP-enhanced clinical trial articles. Third, visualize the scientific collaborations among leading countries/regions, institutions, and authors in the NLP-enhanced clinical trial research. Fourth, uncover the major research topics covered within the NLP-enhanced clinical trial research articles. Fifth, explore the evolution of the major research topics. Besides, visualize the topical distributions across prolific countries/regions and institutions. In addition, identify the potential inter-topic directions in the NLP-enhanced clinical trial research.

2. Literature Review

In the area of literature review, there are two major types (i.e., systematic review [45] and bibliometric analysis). Compared to a systematic review, bibliometric analysis is more suitable for evaluating larger-scale literature data and has thus been widely adopted in various research areas [42,44,46,47]. For example, based on 62 articles collected from the WoS, by using content analysis, Arici et al. [48] presented the research trends and bibliometric results of the literature regarding the adoption of augmented reality in science education. Based on the obtained results, they identified several frequently used keywords, including "science education," "science learning," and "e-learning," particularly mobile learning, which has been a great concern for more recent literature. Besides, it was found that "education," "knowledge," "science education," "experiment," and "effectiveness" were the most frequently used terms in the abstracts. In addition, interests in students' knowledge and achievement were also found to be significant among authors. With the use of the publications of Computers & Industrial Engineering, one of the top academic journals in the field of the supply chain, Cancino et al. [49] identified predominant trends in the field by analyzing the most important publications, keywords, authors, institutions, as well as countries.

STM has also been applied to investigate research topics within a particular research domain. For example, based on the abstract information of research articles on community forestry published during the period of 1990–2017, Clare and Hickey [50] used bibliometric analysis and an STM method to explore trends of geographic foci, as well as research topics. Four main research issues and 20 topics were identified. The distributions of the annual topic proportions indicated that the research focus has evolved from policy analysis to regional carbon sequestration. Based on a collection of open-ended comments from a nationwide representative sample of 16,464 drivers collected in 2016 and 2017, Lee and Kolodge [51] adopted STM to identify 13 research themes, for example, "Tested for a long time" and "Hacking & glitches." With the use of STM, Bennett et al. [52] analyzed 177 responses of the mentally disabled regarding their willingness to using autonomous vehicles. In this way, they identified potential obstacles for the adoption of autonomous vehicles perceived by the mentally disabled. Their study provided suggestions to the government and manufacturers in planning public information campaigns to promote the acceptance of autonomous vehicles by the mentally disabled.

Focusing on NLP and clinical trials research, a few bibliometric studies are available. From the NLP perspective, a bibliometric analysis of NLP [53], NLP-empowered mobile computing research [54], NLP-enhanced medical research [47], and computational linguistics [55] can be found. Another aspect is research related to clinical trials or clinical. Reprehensive works include bibliometric analyses of, e.g.,

the influence measurement of clinical trials citations [56], clinical studies on Tai Chi [57], controlled clinical research concerning traditional Chinese medicine (TCM) in cancer care [58], clinical research articles on sepsis [59], and radiologic randomized controlled trials (RCTs) [60]. For example, with a large volume of clinical studies concerning the safety and advantages of Tai Chi mind-body exercise, Yang et al. [57] systematically synthesized the results or findings of clinical studies concerning Tai Chi for healthcare. Based on the controlled clinical scientific articles of TCM treatments for cancers collected from four major Chinese electronic databases, Li et al. [58] bibliometrically analyzed these articles to assess the reporting quality and summarize the evidence presented in the studied articles. Rosas [61] used a bibliometric analysis method to assess the presence, performance, and impact of scientific articles by the National Institute of Health/National Institute of Allergy and Infectious Diseases clinical trials networks. To be specific, they evaluated the articles in three aspects: (1) Assessing the impact of these articles according to journal impact measures; (2) assessing the performance of articles using bibliometric indicators; and (3) assessing the influence of articles by comparing to similar articles. Tao et al. [59] used bibliometrics to review the development of clinical studies on sepsis systematically. They found that studies contributed by scholars from the USA and published by impactful journals were the most likely to be cited. Hong et al. [60] conducted a bibliometric review of studies on radiologic RCTs to analyze their characteristics and qualities. Based on their analysis, it was found that the quantity and quality of radiologic RCTs had dramatically improved, while the quality of methodologies remained suboptimal. Kim et al. [62] assessed the characteristics and quality of acupuncture RCTs. With an increasing number of RCTs investigating the therapeutic value of yoga interventions, Cramer et al. [63] adopted bibliometrics to present a thorough overview of the characteristics of relevant randomized yoga trials. Cramer et al. [64] aimed to examine whether positive conclusions appeared more in yoga RCTs when they were carried out in India. According to their study, yoga RCTs performed in India were 25 times more likely to reach positive conclusions as compared to those performed somewhere else. In this sense, Indian trials ought to be handled cautiously while assessing yoga's effectiveness for patients from other regions. Bayram et al. [65] identified and characterized the most highly cited papers regarding ultrasonography evaluations taking place in the emergency departments. According to their analysis, the most frequently cited papers concerning the adoption of ultrasonography involved various topics, and almost half of the authors of these papers were experts in emergency medicine.

There are several research studies presenting discussions on issues regarding NLP-enhanced clinical trial research. For example, by including all the applications or methodological articles leveraging texts to support healthcare and meet customers' needs, Demner-Fushman and Elhadad [11] presented a systematic review of articles concerning the application of NLP to clinical and consumer-produced textual data. Their study indicated that although clinical NLP continued to facilitate practical applications with an increasing number of NLP approaches being adopted to process large-scale live health information, there was a need to enable the adoption of NLP in clinical applications available in our daily life. Zeng et al. [66] reviewed studies regarding the NLP-enhanced EHRs-driven computational phenotyping, involving the most advanced NLP techniques for the task. Through their review, several implications were obtained. For example, keyword searches, rule-based approaches, and supervised NLP techniques based on machine learning were the most commonly adopted approaches. Second, well-defined keyword search and rule-grounded systems usually achieved high accuracy. Third, classifications based on supervised methods could achieve higher accuracy and was easier to conduct, as compared to unsupervised methods. Besides, advances in machine learning approaches, for example, deep learning, have grown to be popular. Additionally, there was an increasing interest in extracting relations between medical concepts.

3. Materials and Methods

3.1. Data Collection and Preprocessing

3.1.1. Data Collection

Data for bibliometric analysis concerning the NLP-enhanced clinical trial research were drawn from the WoS, PubMed, and Scopus databases. WoS is a quality-assured database for evaluating research frontiers of particular research fields [67,68]. We also included relevant papers from PubMed as it is the most comprehensive medical literature database worldwide. The PubMed database is provided by the National Library of Medicine, covering the most significant number of publications on life sciences and biomedical research. In addition, we also included Scopus, which is the most reliable [69] and comprehensive [70] academic database of peer-reviewed literature, including scientific journals, books, and conference proceedings.

In the WoS, we adopted "TS" as a search field, as suggested by Gao et al. [71]. Papers indexed in Science Citation Index Expanded, Social Sciences Citation Index, Arts & Humanities Citation Index, Conference Proceedings Citation Index- Science, and Conference Proceedings Citation Index - Social Science & Humanities were considered. In PubMed, [Title/Abstract] was used as a search field [72]. Literature type included article and proceedings papers written in English. Usually, research articles contain more original findings. In addition, a great deal of NLP research papers is covered by conferences.

One critical procedure in data retrieval was to identify a set of keywords related to NLP and clinical trials. The most challenging task was to maximize the identification of relevant studies. For example, publications specific to NLP research may not mention terms like "natural language processing;" thus, a query for retrieving NLP research literature should include field-specific terms like "semantic analysis," which was relevant to NLP. We followed steps proposed by Hassan et al. [73] to attain the keywords list for NLP. Step One: A seed list (e.g., "natural language processing," "semantic analysis," "bag of words," "word sense disambiguation") was provided by NLP domain experts based on their expertise and understanding of NLP. Step Two: The seed keywords were adopted to retrieve papers in the WoS. Step Three: We identified the author-defined keywords contained in those indexed as highly cited papers among the retrieved papers, and then presented them to NLP domain experts to exclude irrelevant ones. The relevant ones were added to the seed list. Similarly, we used the same steps to attain a keywords list for clinical trials with seed keywords, including "clinical trial*," "randomized controlled trial*," and "clinical research trial*." The retrieval was conducted on April 10, 2019. We chose 1999–2018 as the study period. The final keyword queries for data retrieval are shown in Tables A1 and A2 in the Appendix A.

Using the queries, 1595, 512, and 1580 papers were retrieved. The total retrieved 3687 papers were further filtered to guarantee a close relation to the research target, which was conducted by two domain experts (A and B) with comprehensive knowledge and a deep understanding of the interdisciplinary area. Expert A reviewed all the articles based on the criteria listed in Table 2 and finalized 451 papers for further data analysis. Expert B conducted the same filtering process, but for only 100 randomly selected papers. The consistency rate in the decisions made by the two experts was above 95%. Thus, the 451 articles were used as the final dataset, including 194 from the WoS, 25 from PubMed, and 232 from Scopus. As the WoS has provided the research domain of each paper, we analyzed research domains of these papers as inspired by Kowsari et al. [74], the result of which is shown in Table 3. It should be noted that as one paper might originate with more than one domain, we included all of them in our analysis. From the result, we could see that Computer Sciences was the top one domain in the publication of NLP-enhanced clinical trial research, followed by Medical Informatics, Health Care Sciences & Services, and Mathematical & Computational Biology.

Table 2. Examples of inclusion and exclusion criteria for the manual verification.

Area	Type	ID	Criteria
Clinical trials research	Inclusion criteria	I1	Randomized controlled trials
		I2	Clinical research studies
	Exclusion criteria	E2	Not focused on clinical trial
		E3	Not a scientific research
		E4	Without abstract
NLP research	Inclusion criteria	I1	Use of NLP algorithms/approaches/technologies
		I2	Improvement of NLP technology/algorithm
	Exclusion criteria	E1	Use of pure mathematical or statistical algorithms
		E4	Without abstract

Table 3. Top research domains for the 194 papers from the Web of Science (WoS).

Research Domain	Paper Count	Proportion
Computer Science	93	23.66%
Medical Informatics	79	20.10%
Health Care Sciences & Services	50	12.72%
Mathematical & Computational Biology	18	4.58%
Engineering	16	4.07%
Information Science & Library Science	16	4.07%
Psychology	13	3.31%
Public, Environmental & Occupational Health	12	3.05%
Research & Experimental Medicine	7	1.78%
Science & Technology - Other Topics	7	1.78%

3.1.2. Data Pre-Processing

Full bibliographic information of the 451 articles was downloaded in comma separated values (CSV) or plain text file (TXT) files. The CSV or TXT files contained bibliographic details such as authors' names, affiliation, publication year, publication sources, title, abstract, and keywords information. We performed feature extraction and pre-processing. Key features such as title, abstract, keywords, publication source, publication year, and author address of each paper were extracted using a Python program and stored in Microsoft Excel for further pre-processing, as described in the following. First, check all the papers to make sure all features were included in the Excel file. In particular, use Python to check the integrity of the authors' address information automatically. For those without complete information, manual supplements of information were conducted based on the original text of the paper. Second, use Python to extract the information of authors, affiliations, and countries/regions from the authors' addresses and to further transform into analyzable forms. An example of the authors' addresses transformation is shown as Table 4. Third, carry out normalization. For example, for the analysis of prolific countries/regions, articles originating from England, Scotland, Wales, and Northern Ireland were combined as the UK, while articles originating from Hong Kong, Macau, and Taiwan were individually calculated. For the analysis of prolific affiliations, expressions of the same affiliation were unified. For example, Educ Univ Hong Kong, Hong Kong Inst Educ, Educ Univ Hong Kong, The Education University of Hong Kong, Hong Kong Institute of Education, and The Hong Kong Institute of Education were unified as the Education University of Hong Kong.

For science mapping analysis, author-defined keywords, and the KeyWords Plus, PubMed MeSH, or Scopus Index keywords, were adopted. Keywords pre-processing was implemented to enhance analysis efficiency. Firstly, keywords were manually assigned to papers without keyword information, according to paper contents, especially titles. For instance, keywords "Ontology," "data management system," "post-genomic clinical trial," "Grid infrastructure," and "cancer research" were extracted from the title "Ontology based data management systems for post-genomic clinical trials within a European Grid infrastructure for cancer research." Then, all the keywords were expressed in lower case and

singular form. Keywords with similar concept meanings were grouped. For instance, "randomized controlled trial," "randomised controlled trial," "randomized controlled-trial," and "randomised controlled-trial" were grouped with the united expression "randomized controlled trial." Finally, the data were divided into three consecutive periods, 1999–2008 (91 articles), 2009–2013 (149 articles), and 2014–2018 (211 articles), for thematic evolution analysis.

Table 4. Example of the authors' addresses transformation.

Before Transformation			
Paper ID	Authors' address		
WOS:000179539400023	[Tu, Samson W.; Peleg, Mor] Stanford Univ, Stanford Ctr Biomed Informat Res, Stanford, CA 94305 USA.<;>[Peleg, Mor] Univ Haifa, Dept Management Informat Syst, IL-31999 Haifa, Israel.<;>[Carini, Simona; Bobak, Michael; Sim, Ida] Univ Calif San Francisco, Dept Med, San Francisco, CA USA.<;>[Ross, Jessica] Univ Calif San Francisco, Dept Psychiat, San Francisco, CA 94143 USA.<;>[Rubin, Daniel] Stanford Univ, Dept Radiol, Stanford, CA 94305 USA.		
After transformation			
Paper ID	Author	Affiliation	Country/region
WOS:000289030100006	Peleg, Mor	Stanford Univ	USA
WOS:000289030100006	Rubin, Daniel	Stanford Univ	USA
WOS:000289030100006	Tu, Samson W.	Stanford Univ	USA
WOS:000289030100006	Bobak, Michael	Univ Calif San Francisco	USA
WOS:000289030100006	Carini, Simona	Univ Calif San Francisco	USA
WOS:000289030100006	Ross, Jessica	Univ Calif San Francisco	USA
WOS:000289030100006	Sim, Ida	Univ Calif San Francisco	USA
WOS:000289030100006	Peleg, Mor	Univ Haifa	Israel

Before conducting the STM analysis, we conducted a pre-processing procedure to facilitate the analysis. First, abstracts of the 451 selected articles were collected, for which terms were extracted by a self-developed NLP program. Second, terms with similar meanings were unified (e.g., 'modeling' and 'modelling;' 'personalize' and 'personalise'). Third, stop terms were excluded (e.g., "of," "you," "a," and "the"). Besides, terms containing little or no meaningful information were excluded (e.g., "also," "although," "among," "another," "demonstrate," "furthermore," "however," "include," "possible," and "towards"). In addition, term frequency-inverse document frequencies (TF-IDF), as one of the common techniques of feature extractions [75], was further adopted to filter unimportant terms. We followed the suggestion by Jiang et al. [76] to include terms with TF-IDF values more than or equal to 0.05.

3.2. Performance Bibliometric Analysis

Several bibliometric indicators were implemented, including paper counts, citation counts, paper counts within citation thresholds, and the Hirsch index (H-index) [77,78]. Given that PubMed does not provide citations for a paper, for comparison convenience, we used Google scholar citation counted up to 26 April 2019, for each of the analyzed papers.

The bibliometric performance indicators can help identify prolific and influential publication sources, authors, and institutions, as well as provide data on geographic distribution. Meanwhile, the indicators can also be used to enrich the science mapping analysis workflow by measuring the comparative importance of different research themes, thus helping to establish the most leading, productive, and impactful research areas [30]. The quantitative indicators (e.g., article count) can be adopted to evaluate how productive the detected themes are, while the qualitative indicators (e.g., citation count and H-index) can evaluate their quality.

3.3. Scientific Collaboration Networks Analysis

Social network analysis was used to explore the scientific collaborations among authors, institutions, and countries/regions in the NLP-enhanced clinical trial research field. It has been increasingly important and widely adopted in various research fields [79], for example, psychology, business, and education. Taking the co-author network as an example, with this type of network, we are able to discover different features, for instance, author communities [25].

In this paper, we used Gephi (https://gephi.org/) to generate the collaboration networks, in which nodes indicated authors, institutions, or countries/regions, while lines indicated collaboration relations. The strength of collaboration relations was denoted by line width, while the node size indicated productivity, namely, the paper counts. Especially, we also embodied the country/region information for authors and institutions, as well as the continent information for countries/regions in the networks using node colors.

3.4. Science Mapping Analysis

As a spatial illustration of relations among disciplines and papers [80], science mapping has been broadly applied to various research areas to uncover hidden research topics. This study aimed to measure the evolution progress of the research topics to discover newly emerging research themes, as well as predict their developments. We implemented SciMAT (http://sci2s.ugr.es/scimat), which integrates the most advantages of the available science mapping tools [81]. We followed the analysis procedure defined by Cobo et al. [82], in which co-word analysis was adopted to explore how research areas evolved throughout consecutive time periods [83]. The approach mainly included three stages, including research themes detection, the layout of research themes, and thematic evolution and performance analysis.

3.5. Structural Topic Modeling Analysis

The STM modeling process was as follows. First, 21 models with topic numbers from five to 25 were fitted, respectively. Second, the domain experts evaluated the fitting effectiveness of the 21 models based on their domain knowledge, as well as by examining representative terms and articles for each topic. A 16 topic model was finally selected out. Third, based on the results of the 16 topic STM model, we assigned labels to each topic based on their representative terms and articles. Besides, we conducted a statistical Mann-Kendall trend test method to examine the development trend of each topic throughout the whole period. Furthermore, by incorporating metadata such as country/region, institution, and publication data information of each article, we further calculated and visualized the topical distributions across prolific countries/regions and institutions, as well as topic proportion distributions by year. In addition, we investigated the correlation between topics using a hierarchical clustering method.

4. Performance and Collaboration Analyses Results

The bibliometric analysis results of articles in the research area of NLP techniques for clinical trial text processing are displayed in this section, with article trend analysis, analysis of publication sources, countries/regions analysis, institutions, and authors included. In addition, results of scientific collaborations analyses and thematic evolutions analyses are also presented.

4.1. Analysis of Trend in Articles

The annual distribution of articles is displayed in Figure 1. The annual number of articles had increased from four in 1999 to 53 in 2017.

Until 2005, the number was around seven by year, which is quite low. From 2006 until 2011, the number increased to triple, which was around 20. Furthermore, from 2012 until 2017, it was around 45. Thus, the research area had received growing interest and attention from academia during all these

years. However, the articles collected with the publication year of 2018 had experienced a sudden drop to 24, which might have partially resulted from time-lag for academic papers to be indexed by academic databases. Thus, there was a great possibility that some articles published in 2018 had not been indexed in the retrieval databases.
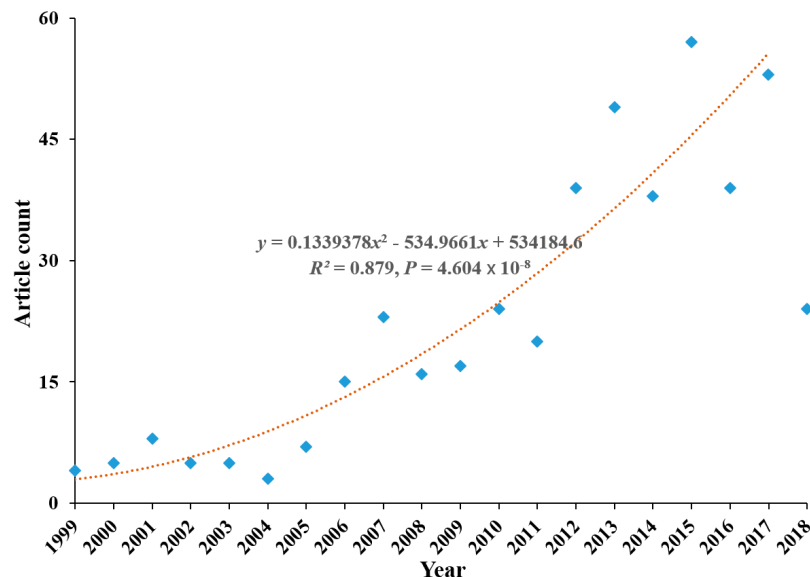

The fitted equation shown in the figure:
$$y = 0.1339378x^2 - 534.9661x + 534184.6$$
$$R^2 = 0.879, P = 4.604 \times 10^{-8}$$

Figure 1. Analysis of trends in articles.

Polynomial regression analysis was further conducted to fit the article trend with the *year* as an independent variable. For obtaining a better fitting efficiency, we utilized the period 1999–2017 for regression fitting. The fitted polynomial regression curve with a positive coefficient of x^2 ($R^2 = 0.879$) demonstrated a growing interest and enthusiasm for the research topic.

4.2. Analysis of Publication Sources

Regarding sources, results showed that two or more articles about NLP techniques for clinical trial texts processing research were published in 48 professional sources. Table 5 lists the top thirteen publication sources with an H index value ≥4. Most of the authors published their studies in sources of medical informatics such as the *Journal of Biomedical Informatics (JBI)*, *Journal of the American Medical Informatics Association (JAMIA)*, and *American Medical Informatics Association (AMIA) Annual Symposium*, which were the top three in the list. The three together contributed nearly 20% of the articles. *JBI* was the most influential source with an H index value of 20. Additionally, it was also the most productive one, with 37 articles. Thus, it not only had a wide research influence in the area but also had great productivity. *JAMIA*, with an H index value of 17, ranked in second place. Although *JAMIA* had comparatively fewer articles than *JBI*, it had an average-citations-per-article (ACP) value much higher than other sources, including JBI. This showed that it was widely and commonly influential and recognized in the research area. Regarding the article distribution for the two periods, it was interesting to find that JBI had significantly more articles in the second period, much more than other sources. This showed a fast and flourishing development of the NLP-enhanced clinical trial research in recent years.

4.3. Analyses of Country/Region, Institution, and Author

Forty-five countries/regions had contributed to the 451 articles. Twelve countries/regions had more than ten articles about the NLP-enhanced clinical trial research, while 14 countries/regions had contributed to only one article each. Table 6 depicts the top 20 ranked by H-index. The top four were the USA (42 H-index, 231 articles), Germany (17 H-index, 57 articles), the UK (16 H-index, 51 articles), and the Netherlands (14 H-index, 39 articles), which were also the top four ranked by article count.

Table 5. Distribution of articles by publication source.

Journal/Conference	H	A (R)	TC	ACP	≥100	≥50	A1	A2
Journal of Biomedical Informatics	20	37 (1)	1224	33.08	3	7	1	36
Journal of the American Medical Informatics Association	17	23 (3)	1357	59	2	5	7	16
American Medical Informatics Association (AMIA) Annual Symposium	12	22 (4)	315	14.32	0	0	13	9
Lecture Notes in Computer Science	9	16 (5)	222	13.88	0	1	4	12
Studies in Health Technology and Informatics	8	26 (2)	258	9.92	0	0	4	22
BMC Medical Informatics and Decision Making	7	11 (6)	175	15.91	0	1	1	10
Artificial Intelligence in Medicine	6	7 (12)	290	41.43	0	2	1	6
International Journal of Medical Informatics	6	8 (7)	123	15.38	0	0	1	7
PLoS ONE	6	8 (7)	101	12.63	0	0	0	8
Methods of Information in Medicine	5	8 (7)	81	10.13	0	0	2	6
BMC Bioinformatics	4	4 (14)	147	36.75	0	1	0	4
World Congress on Medical and Health Informatics (MEDINFO)	4	8 (7)	61	7.63	0	0	3	5
International Conference on Health Informatics (HEALTHINF)	4	8 (7)	52	6.5	0	0	3	5

Abbreviations: H: H-index; A: Article count; R: Ranking position; TC: Citation count; ACP: Average citations per article; ≥100, ≥50: Numbers of articles with more than 100 or 50 citations, respectively; A1, A2: Numbers of articles published during the periods 1999–2008 and 2009–2018, respectively.

Table 6. Distribution of articles by country/region.

C	H	A (R)	TC	ACP	≥100	≥50	A1	A2	Main collaborator (CT ≥ 5)
the USA	42	231 (1)	6425	27.81	11	30	58	173	the UK; Germany; China; Canada; Australia
Germany	17	57 (2)	793	13.91	0	3	7	50	Greece; Spain; the UK; the Netherlands; USA; Belgium
the UK	16	51 (3)	1013	19.86	1	5	7	44	USA; the Netherlands; Germany
the Netherlands	14	39 (4)	541	13.87	0	2	5	34	Spain; Belgium; Germany; the UK; Greece
Canada	12	17 (9)	1164	68.47	1	6	6	11	the USA
Australia	10	19 (7)	245	12.89	0	0	2	17	the USA
Spain	9	27 (5)	423	15.67	0	3	4	23	the Netherlands; Germany; Greece; Belgium
France	8	20 (6)	366	18.30	0	3	4	16	
China	8	17 (9)	286	16.82	0	2	1	16	the USA
Greece	8	19 (7)	308	16.21	0	2	6	13	Germany; Spain; the Netherlands

Abbreviations are the same as Table 5.

The table also includes the citation count (TC), average citations per article (ACP), numbers of articles with more than 100 and 50 citations (≥100, ≥50, respectively), and numbers of articles published during 1999–2008 (A1) and 2009–2018 (A2). The USA had the top rank position for almost all index metrics (except ACP index), reflecting the high quality of its articles, as well as its dominant position in the NLP techniques for clinical trial text processing research. In addition, the USA was the main collaborator for countries of Germany, the UK, Canada, Australia, and China. Germany was the main collaborator for the USA, the UK, the Netherlands, Spain, and Greece.

There were 605 institutions that participated in the 451 studied articles. Thirty-five institutions had more than five articles on the NLP-enhanced clinical trial research, while 441 institutions had participated in only one article each. Table 7 shows the top 17 ranked by H-index, among which the top three were *Stanford University*, *the National Institutes of Health*, and *the University of California, San Francisco*. For the 17 listed institutions, eight were from the USA, three from the Netherlands, three from the UK, one from Germany, one from Greece, and one from Spain. From the article count perspective, the top three were *Stanford University* (25 articles), *Columbia University* (21 articles), and *the Technical University of Madrid* (17 articles). From the ACP perspective, the top three were *the University of California, San Francisco* (75.86 ACP), *National Institutes of Health* (59.50 ACP), and *University of Oxford* (39.67 ACP). It is worth noting that concerning the number of articles published during the period 1999–2008, *Stanford University* was ranked first with 14 articles, while most of the articles of *Columbia University* were published during the second period, demonstrating a flourishing development in the NLP-enhanced clinical trial research.

Table 7. Distribution of articles by institutions.

Institution	Country	H	A (R)	TC	ACP	≥100	≥50	A1	A2
Stanford University	the USA	16	25 (1)	770	30.80	1	5	14	11
National Institutes of Health	the USA	12	14 (5)	833	59.50	3	6	4	10
University of California, San Francisco	the USA	11	14 (5)	1062	75.86	2	4	7	7
Columbia University	the USA	10	21 (2)	392	18.67	0	1	3	18
Foundation for Research and Technology - Hellas	Greece	8	13 (7)	274	21.08	0	2	6	7
Vrije University Amsterdam	the Netherlands	8	12 (8)	175	14.58	0	1	0	12
Philips Research	the Netherlands	7	16 (4)	134	8.38	0	0	2	14
Technical University of Madrid	Spain	7	17 (3)	200	11.76	0	1	3	14
University of Washington	the USA	7	7 (16)	228	32.57	0	1	1	6
King's College London	the UK	6	8 (13)	292	36.50	1	1	1	7
Mayo Clinic	the USA	6	6 (22)	175	29.17	1	1	1	5
Northwestern University	the USA	6	6 (22)	174	29.00	0	0	1	5
Saarland University	Germany	6	8 (13)	243	30.38	0	2	3	5
University of Amsterdam	the Netherlands	6	10 (9)	208	20.80	0	1	1	9
University College London	the UK	6	8 (13)	283	35.38	1	1	1	7
University of Oxford	the UK	6	9 (11)	357	39.67	1	2	1	8
University of Texas Health Science Center at Houston	the USA	6	7 (16)	197	28.14	1	1	0	7

Abbreviations are the same as Table 5.

One thousand nine hundred ninety-nine authors had contributed to the 451 studied articles. Thirty-three authors had more than five articles on the NLP-enhanced clinical trial research, while 87.44% of authors had participated in only one article. Table 8 depicts the top 19 ranked by H-index, where the top three were *Ida Sim* (H-index as ten) from *the University of California, San Francisco, Manolis Tsiknakis* (H-index as eight) from the Foundation for *Research and Technology-Hellas*, and *Amar K. Das* (H-index as eight) from *Stanford University*. For the 19 listed authors, ten were from the USA, four from Spain, two from Germany, two from Greece, and one from the Netherlands. From the article count perspective, the top three were *Anca Bucur* (15 articles) from *Philips Research*, *Chunhua Weng* (13 articles) from *Columbia University*, and *Manolis Tsiknakis* (13 articles) from the Foundation for *Research and Technology-Hellas*. From the ACP perspective, the top three were *Ida Sim* (ACP as 84.58) from *the University of California, San Francisco, Simona Carini* (ACP as 42.67) from *the University of California, San Francisco, and Mathias Brochhausen* (ACP as 39.50) from *Saarland University*. It is worth noting that concerning the number of articles published during the period 1999–2008, *Amar K. Das* from *Stanford University* ranked first with eight articles, while *Anca Bucur* from *Philips Research* had the most articles published during the second period.

Table 8. Distribution of articles by authors.

Author	Institution	H	A (R)	TC	ACP	≥100	≥50	A1	A2
Ida Sim	*University of California, San Francisco*	10	12 (4)	1015	84.58	2	4	7	5
Manolis Tsiknakis	*Foundation for Research and Technology - Hellas*	8	13 (2)	274	42.67	0	2	6	7
Amar K. Das	*Stanford University*	8	10 (5)	176	39.50	0	0	8	2
Chunhua Weng	*Columbia University*	7	13 (2)	208	27.20	0	0	0	13
Anca Bucur	*Philips Research*	7	15 (1)	117	25.57	0	0	1	14
Martin O'Connor	*Stanford University*	7	8 (8)	136	22.00	0	0	7	1
Ravi D. Shankar	*Stanford University*	7	7 (10)	124	21.50	0	0	6	1
Simona Carini	*University of California, San Francisco*	7	9 (6)	384	21.50	1	3	4	5
Stelios Sfakianakis	*Foundation for Research and Technology - Hellas*	6	7 (10)	179	21.08	0	2	3	4
Mathias Brochhausen	*Saarland University*	6	6 (18)	237	20.57	0	2	3	3
Alberto Anguita	*Technical University of Madrid*	6	7 (10)	144	17.71	0	1	3	4
Hua Xu	*University of Texas Health Science Center at Houston*	5	6 (18)	95	17.60	0	0	0	6
Todd Lingren	*Cincinnati Children's Hospital Medical Center*	5	6 (18)	129	17.00	0	0	0	6
Imre Solti	*Cincinnati Children's Hospital Medical Center*	5	6 (18)	129	16.00	0	0	0	6
David B. Parrish	*Immune Tolerance Network*	5	5 (23)	110	15.83	0	0	4	1
Martin Dugas	*University of Munster*	5	9 (6)	97	10.78	0	0	0	9
Victor Maojo	*Technical University of Madrid*	5	7 (10)	65	9.29	0	0	2	5
Luis Martin	*Technical University of Madrid*	5	5 (23)	136	7.80	0	1	3	2
David Perez-Rey	*Technical University of Madrid*	5	8 (8)	39	4.88	0	0	0	8

Abbreviations are the same as Table 5.

4.4. Analysis of Scientific Collaboration

The scientific collaboration relations among the 44 countries/regions (http://home.eduhk.hk/~{}hxie/applied_sciences/countries.html) participating in research of NLP techniques for clinical trials text processing are illustrated as Figure 2 with 44 nodes and 124 links. Among the 44 countries/regions,

23 were from Europe (orange nodes), 12 from Asia (red nodes), five from South America (purple nodes), two from North America (blue nodes), and two from Oceania (green nodes). The collaborations among European countries/regions were close. The USA and the UK had the most collaborations (in 13 articles), followed by Germany-Greece (11 articles), the Netherlands-Belgium (nine articles), Germany-Spain (nine articles), and the Netherlands-Spain (nine articles). Of these 58 countries, the USA was the most collaborative country with collaborations with 29 countries/regions. Thus, it was located in the center of the network. Other collaborative countries/regions included the UK, Germany, the Netherlands, and Greece, with collaborations with 29, 21, 18, 15, and 15 countries/regions, respectively.
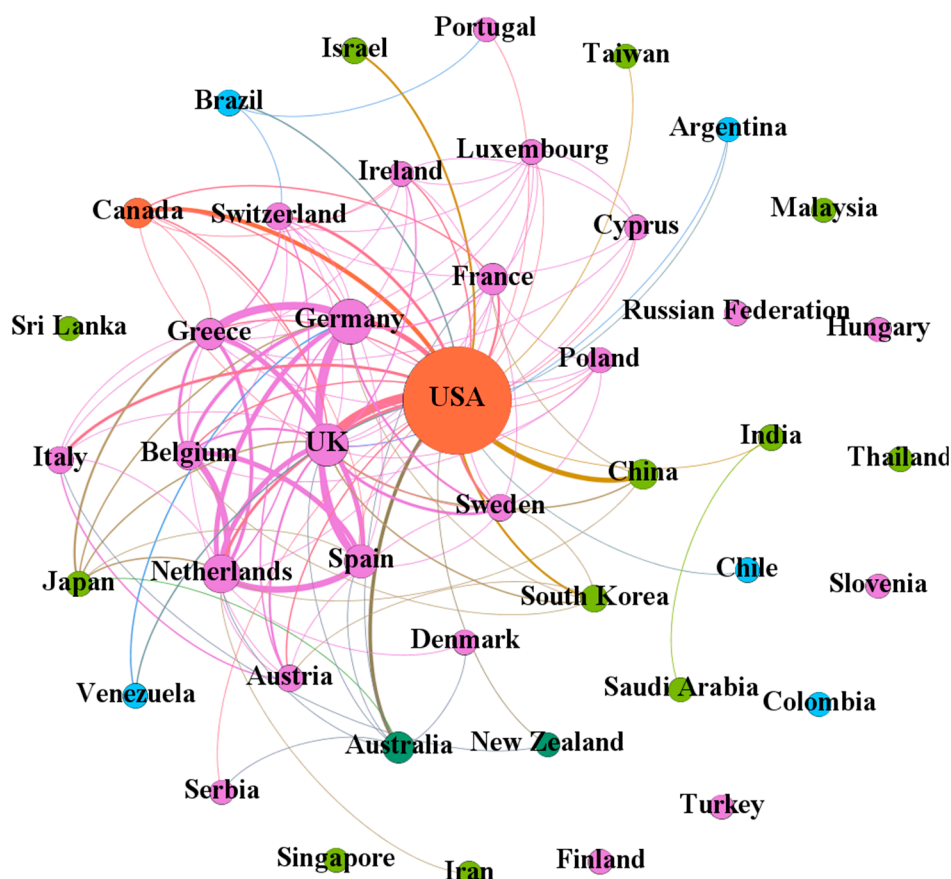
Figure 2. Collaboration network among countries/regions.

Figure 3 shows the scientific collaboration network among 35 institutions (http://home.eduhk.hk/~{}hxie/applied_sciences/institutions.html) with an article count ≥5 and with 35 nodes and 65 links. Of them, about 50% are from the USA (blue nodes). The collaborations among them were very close. *The Technical University of Madrid* from Spain and *Philips Research* from the Netherlands were the most collaborative partners (nine articles), followed by *Philips Research* from the Netherlands and *Custodix NV* from Belgium (six articles), as well as *the Technical University of Madrid* from Spain and *the Foundation for Research and Technology-Hellas* from Greece (six articles). Of these 35 institutions, *the University of Oxford* from the UK (11 collaborators) was the most collaborative institution, followed by *the National Institutes of Health* from the USA (10 collaborators) and *Stanford University* from the USA (nine collaborators).

The scientific collaboration relations among 33 authors (http://home.eduhk.hk/~{}hxie/applied_sciences/authors.html) with an article count ≥5 are displayed in Figure 4 with 33 nodes and 70 links. Among the 33 authors, 13 were from the USA (orange nodes). Interestingly, the collaborations among authors from the same countries/regions were close, with several clusters formed. For example, a cluster was formed by *Philipp Neuhaus*, *Mathias Brochhausen*, and *Julian Varghese*, all of which were from Germany. A cluster was formed by *Efthymios Chondrogiannis*, *Vassiliki Andronikou*, and *Theodora*

Varvarigou, all of which were from Greece. In addition, three clusters formed by authors from the USA were significantly and clearly found. *Simona Carini* from the USA and *Ida Sim* from the USA had the most collaborations (9 articles). Of these 33 authors, *Anca Bucur* from the Netherlands was the most collaborative author (13 collaborators), followed by *Manolis Tsiknakis* from Greece and *Victor Maojo* from Spain, each with nine collaborators.
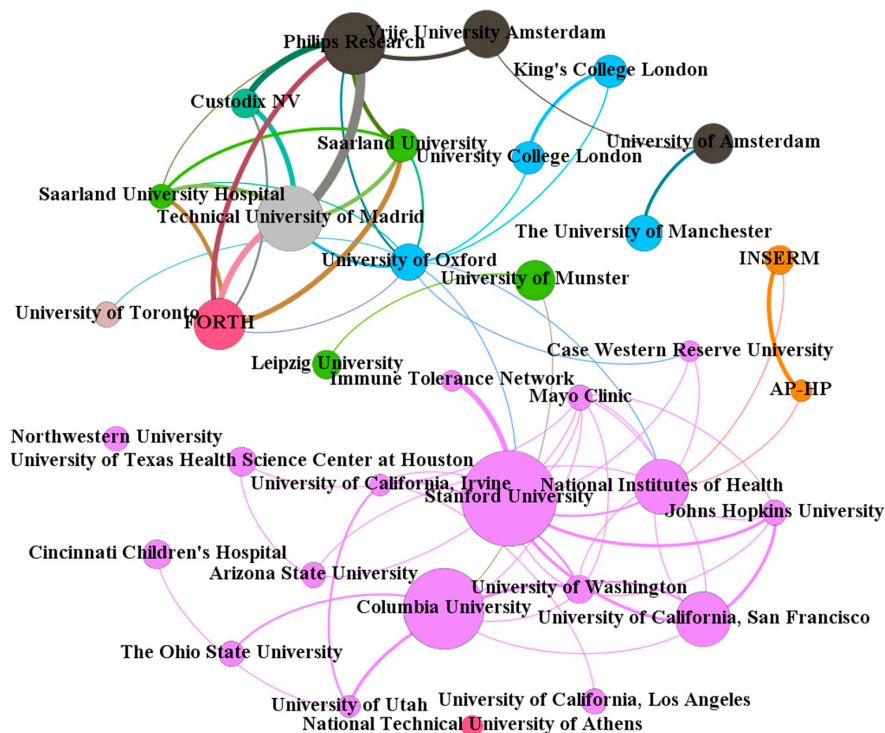

Figure 3. Collaboration network among institutions with ≥5 articles.
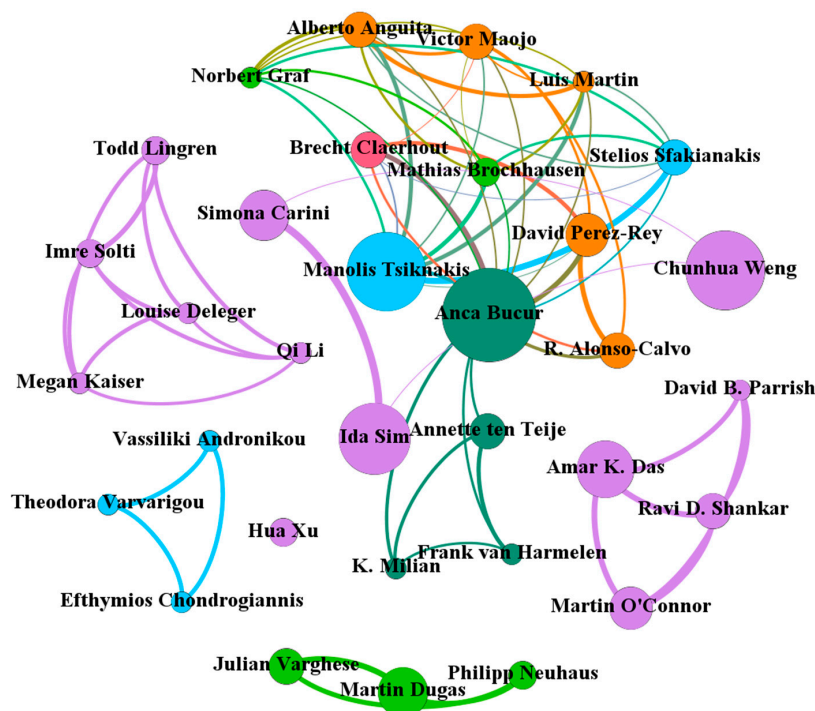
Figure 4. Collaboration network among authors with ≥5 articles.

5. Science Mapping Analysis Results

5.1. Analysis of the Content of the Papers Published

Strategic diagrams were drawn to explore the research themes of high importance in the NLP-enhanced clinical trials research field, where the sphere size was proportional to articles of each research theme, and citations of each theme were also integrated, as in brackets.

During the first sub-period 1999–2008 (see Figure 5a), *HUMAN*, *COMPUTER-PROGRAM*, and *SYSTEM* were motor themes, meaning that these research themes were essential and well developed for structuring the NLP-enhanced clinical trial research field. The performance measures of each theme are depicted in Table 9, including article count, citation count, and H-index. Based on the three performance measures, *HUMAN* and *COMPUTER-PROGRAM* could be highlighted. Compared to the remaining themes, they received a significant impact rate, with more citations and higher H-indexes. As for the second sub-period 2009–2013, based on the strategic diagram depicted in Figure 5b, five major themes were highlighted, including *HUMAN*, *ARTIFICIAL-INTELLIGENCE*, *FACTUAL-DATABASE*, *MEDICAL-APPLICATION*, and *MEDICAL-RECORD*. In Table 9, two research themes were highlighted with more citations and higher H-indexes, namely, *HUMAN* and *ARTIFICIAL-INTELLIGENCE*. Based on the strategic diagram depicted in Figure 5c, during the last sub-period 2014–2018, the NLP-enhanced clinical trials research field had become more diverse with 12 research themes with seven major ones (including motor and basic themes), including *PROCEDURE*, *MACHINE-LEARNING*, *ELECTRONIC-HEALTH*-RECORD, *ACCURACY*, *FEMALE*, *COMPUTER-ASSISTED*, and *GENETICS*. In Table 9, *PROCEDURE* with more citations should be highlighted. In addition, the following two themes were notable with a higher H-index: *MACHINE-LEARNING* and *RANDOMIZED-CONTROLLED-TRIAL*.
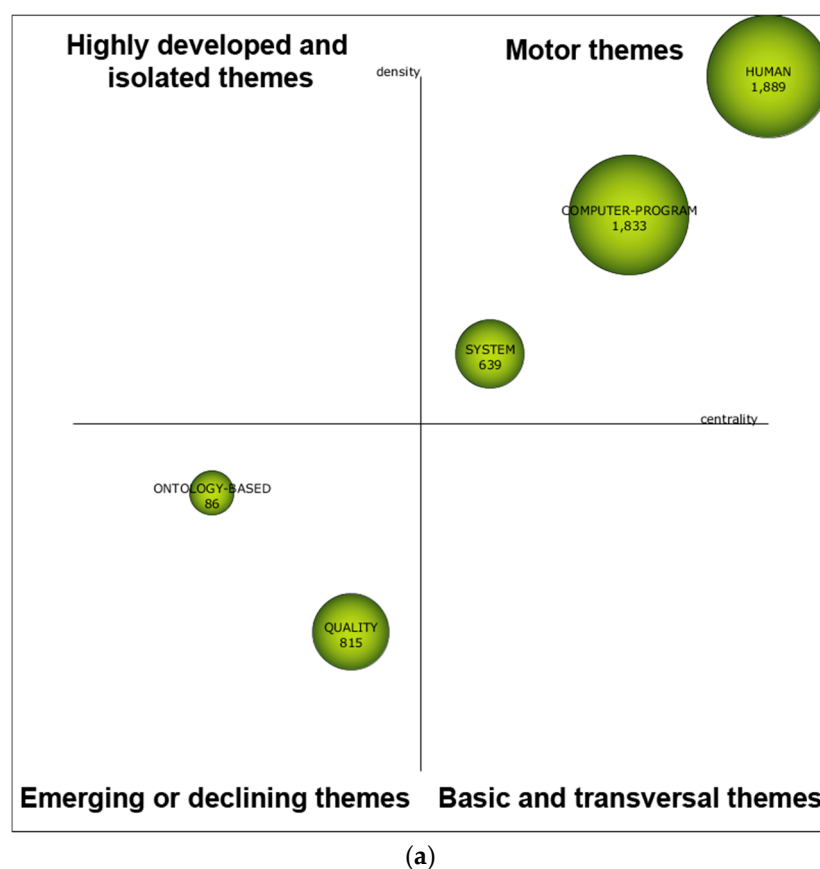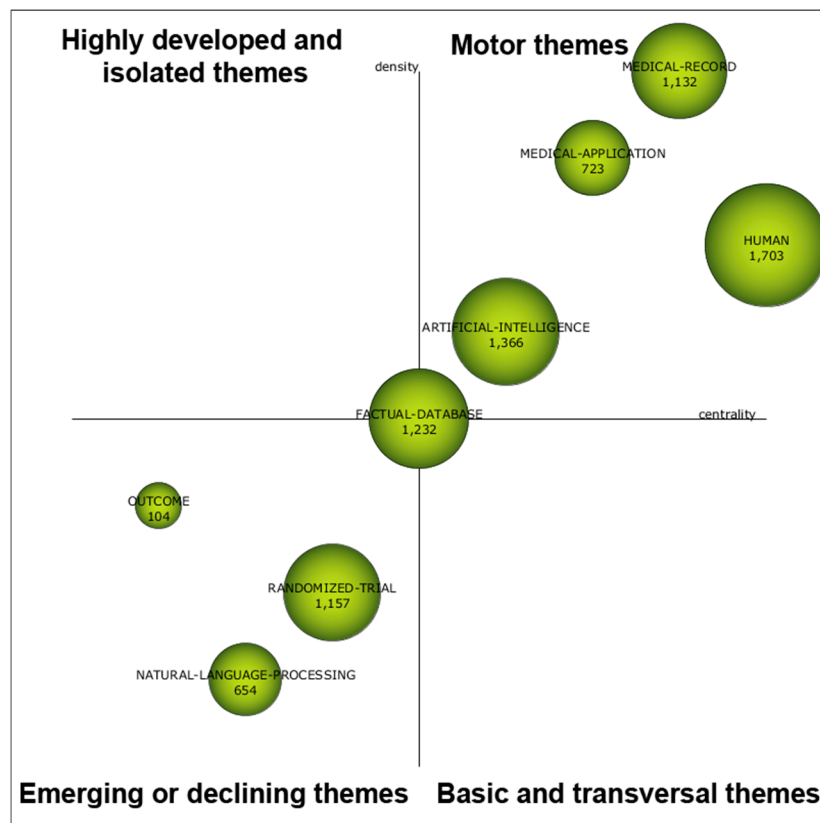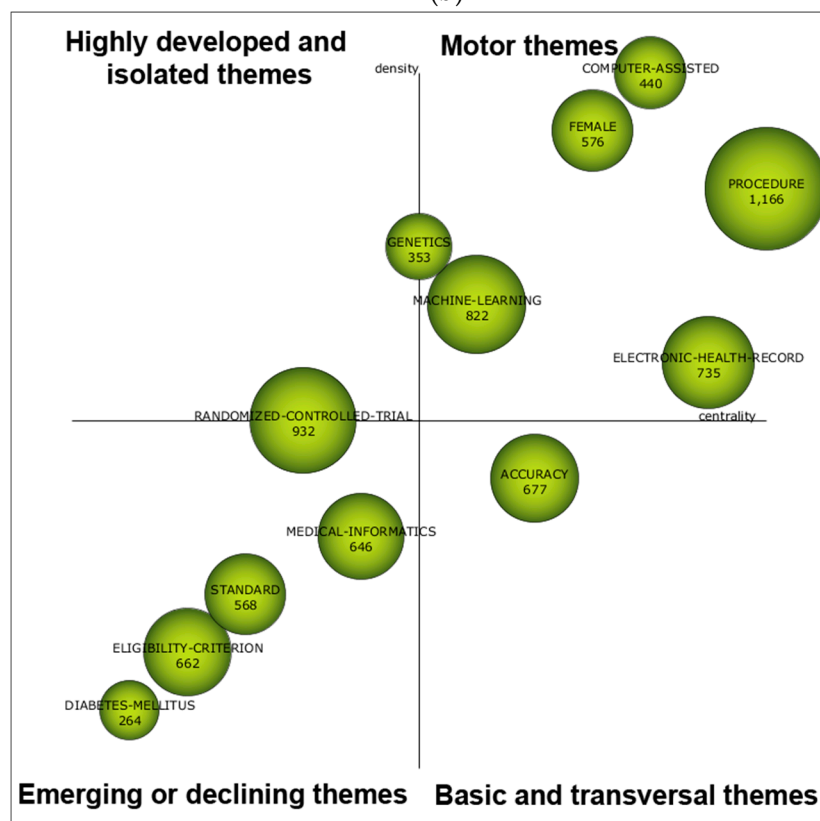

(a)

Figure 5. *Cont.*

(**b**)

(**c**)

Figure 5. Strategic diagram for the periods 1999–2008 (**a**), 2009–2013 (**b**), and 2014–2018 (**c**).

Table 9. Examples of inclusion and exclusion criteria for the manual verification.

Period	Name	Articles	Citations	H-index
1999–2008	*HUMAN*	43	1889	20
	COMPUTER-PROGRAM	35	1833	16
	SYSTEM	26	639	17
	QUALITY	11	815	7
	ONTOLOGY-BASED	3	86	3
2009–2013	*HUMAN*	79	1703	24
	MEDICAL-RECORD	51	1132	19
	ARTIFICIAL-INTELLIGENCE	48	1366	23
	FACTUAL-DATABASE	46	1232	18
	RANDOMIZED-TRIAL	43	1157	19
	MEDICAL-APPLICATION	41	723	14
	NATURAL-LANGUAGE-PROCESSING	24	654	13
	OUTCOME	8	104	6
2014–2018	*PROCEDURE*	106	1166	20
	MACHINE-LEARNING	73	822	17
	RANDOMIZED-CONTROLLED-TRIAL	72	932	17
	ELECTRONIC-HEALTH-RECORD	69	735	14
	ACCURACY	60	677	14
	MEDICAL-INFORMATICS	53	646	14
	STANDARD	51	568	14
	FEMALE	49	576	13
	ELIGIBILITY-CRITERION	42	662	15
	COMPUTER-ASSISTED	36	440	11
	GENETICS	33	353	10
	DIABETES-MELLITUS	23	264	8

5.2. Conceptual Evolution Map

Referring to the strategies by Ramírez and Sánchez-Cañi [84], we conducted a conceptual evolution analysis. The period 1999–2008 contained 798 keywords, 328 of which appeared again in the period 2009–2013, while the remaining 470 keywords did not. Thus, the similarity index between the two periods equaled 41%. The period 2009–2013 contained 1429 keywords, 1101 of which were newly emerged. In the period 2014–2018, 557 of these keywords appeared again, while 872 did not. The similarity index between the latter two periods equaled 39%. Given that there were more articles in the period 2014–2018, there were 1951 keywords included, of which 1394 have newly emerged. The importance of keywords was constantly changing across time (Figure 6). With terminology changes, diverse keywords were adopted to depict contents of the NLP-enhanced clinical trial studies, with some new keywords emerging while others vanished. For instance, keywords *NURSING-EDUCATION*, *SHORT-TERM-MEMORY*, *ASSOCIATE-LEARNING*, *ATTITUDE-TO-COMPUTER*, *BIOMEDICAL-GRID*, and *CANCER-SCREENING* appeared only in the first period. However, keywords such as *SEMANTICS*, *SYSTEM*, *ONTOLOGY*, *RANDOMIZED-CONTROLLED-TRIAL*, *ARTIFICIAL-INTELLIGENCE*, *EVIDENCE-BASED-MEDICINE*, *NATURAL-LANGUAGE-PROCESSING*, and *DECISION-SUPPORT-SYSTEM* were presented in all three periods. The evolution map depicted in Figure 6 was very dense. Some keywords belonged simultaneously to more than one area, while others were isolated (e.g., *ONTOLOGY-BASED* and *QUALITY*) or failed to be continuously prominent.
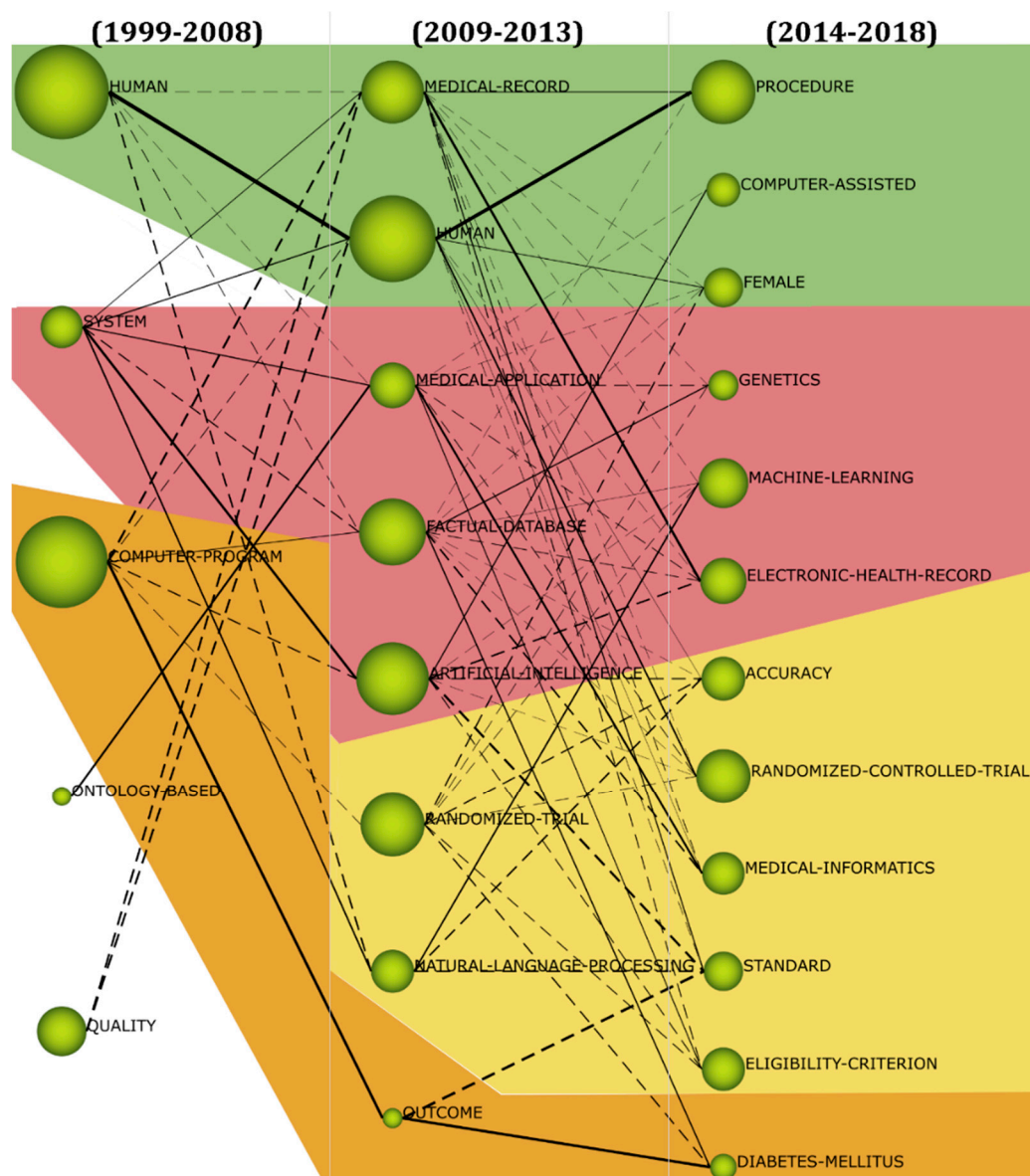
Figure 6. Thematic evolution (1999–2018).

6. Structural Topic Modeling Analysis Results

6.1. Topic Identification and Trends

Of the extracted terms, the most frequent ones included "clinical (in 344 articles)," "trial (335)," "patient (199)," "information (197)," "system (188)," "knowledge (179)," "semantic (161)," and "support (137)." Table 10 displays the 16 topic STM results, with an integration of the most discriminating terms, topical proportions, suggested labels, as well as topical trend test results. Of the 16 topics, some topics were about technologies or methods such as Information extraction and Text mining. Two topics were drug-related, namely, Drug and Drug side effect. The top five most frequently discussed topics included Interoperability (13.21%), Clinical decision support (10.38%), Information extraction (9.36%), Eligibility criteria (8.68%), and Protocol (8.08%).

Table 10. The 16 structural topic modeling (STM) results with the discriminating terms, topical proportions, suggested topic labels, and trend test results.

Discriminating Terms	%	Suggested Topic	Trend
ACGT, RDF, ODM, project, genomic, MDM integration, metadata, grid, interoperability, clinico, infrastructure, service, CDISC, portal	13.21	*Interoperability*	↑↑
CDSS, PCROM, archetype, ontology, terminology, interoperability, EC, EHR, EMR, healthcare, practice, AskHERMES, nutritional, clinical, supporting,	10.38	*Clinical decision support*	↑
extraction, ADRs, natural, radiology, extract, precision, summarization, information, NLP, biomedical, free, text, processing, SR, mining	9.36	*Information extraction*	↑↑
annotation, eligibility, criterion, UMLS, SNOMED, automatic, semantic, relatedness, AnnSim, FSTs, CT clinicaltrials.gov, pattern, matching, exclusion	8.68	*Eligibility criteria*	↑↑
protocol, temporal, constraint, TBNS, management, representation, accrual, TA, enrollment, eligibility, eligible, criterion, trial, querying, conflict	8.08	*Protocol*	↓↓
landmark, protein, gene, graph, Taxane, interaction, phenotype, diacetylmorphine, DTOME, candidate, QA, anti, caffeine, genotype, PGX	6.76	*Protein & Gene*	↑
bank, RCTS, systematic, cochrane, registration, pro, randomized, HTA, controlled, search, delivery, alteration, CDA, register, update	5.57	*Randomized controlled trials*	↓
QI, resident, professional, factor, sexual, team, alert, health, feedback, importance, HRA, cluster, CR, DDI, mental	5.55	*Life quality*	↑↑↑
IMS, glycemic, nursing, indicator, diabetes, obesity, disclosure, APN, EBM, intervention, control, family, evidence, management, telemedicine	5.32	*Intervention*	↑↑
OAR, ADHD, rapidplan, AI, EOL, exercise, RD, HIV, treatment, disorder, EFV, TDM, BDETs, therapy, therapist	5.22	*Treatments*	↑
classifier, sentence, corpus, binary, naive, consent, performance, PECODR, ethic, bayes, machine, feature, SVM, vector, classification	5.18	*Text mining*	↑↑
COIS, ECIG, abutment, dissection, delirium, PWPD, HCPs, similarity, OAE, connection, category, term, financial, minority, AES	4.22	*Terms & management*	↓
meta, rectum, SDCT, IMR, IR, RDCT, activation, dose, PC, sealant, tagging, consistency, plan, rule, DRPs	3.76	*Drug*	↑↑↑
HCI, cohesion, display, CTPT, time, trend, visitor, iconic, pneumonia, T2DM, exploration, representativeness, comprehension, temporal, imaging	3.18	*Human–computer interaction*	↑↑
NRT, comment, sentiment, ADE, video, cessation, exposure, negative, positive, engagement, self, twitter, depression, emotion, compassion	3.05	*Social & emotional*	↑↑↑
side, industry, event, adverse, drug, pipeline, authored, safety, adjective, approved, effect, product, market, serious, repositioning	2.5	*Drug side effect*	↑↑

Note: Topics are ranked by the proportion in descending order. Frequent and exclusive terms have a high frequency-exclusivity (FREX) value. %: Topic proportions. The corresponding full-names of FREX in abbreviation form are shown in Table A3 in the Appendix A. ↑(↓): Increasing (decreasing) trend but not statistically significant ($p > 0.05$); ↑↑(↓↓), ↑↑↑(↓↓↓), ↑↑↑↑(↓↓↓↓): Statistically significant increasing (decreasing) trend ($p < 0.05$, $p < 0.01$, and $p < 0.001$, respectively).

According to the Mann–Kendall trend test, most of the 16 topics presented an increasing trend. In particular, ten topics, namely, Interoperability, Information extraction, Eligibility criteria, Life quality, Intervention, Text mining, Drug, Human–computer interaction, Social & emotional, and Drug side effect, exhibited statistically significant increasing trends. Topic Protocol was found to experience a statistically significant decreasing trend. The annual proportion distributions of the 16 topics are also visualized in Figure 7, from which it was obvious to find that some topics appeared to be increasing, for example, Text mining, Information extraction, and Life quality.
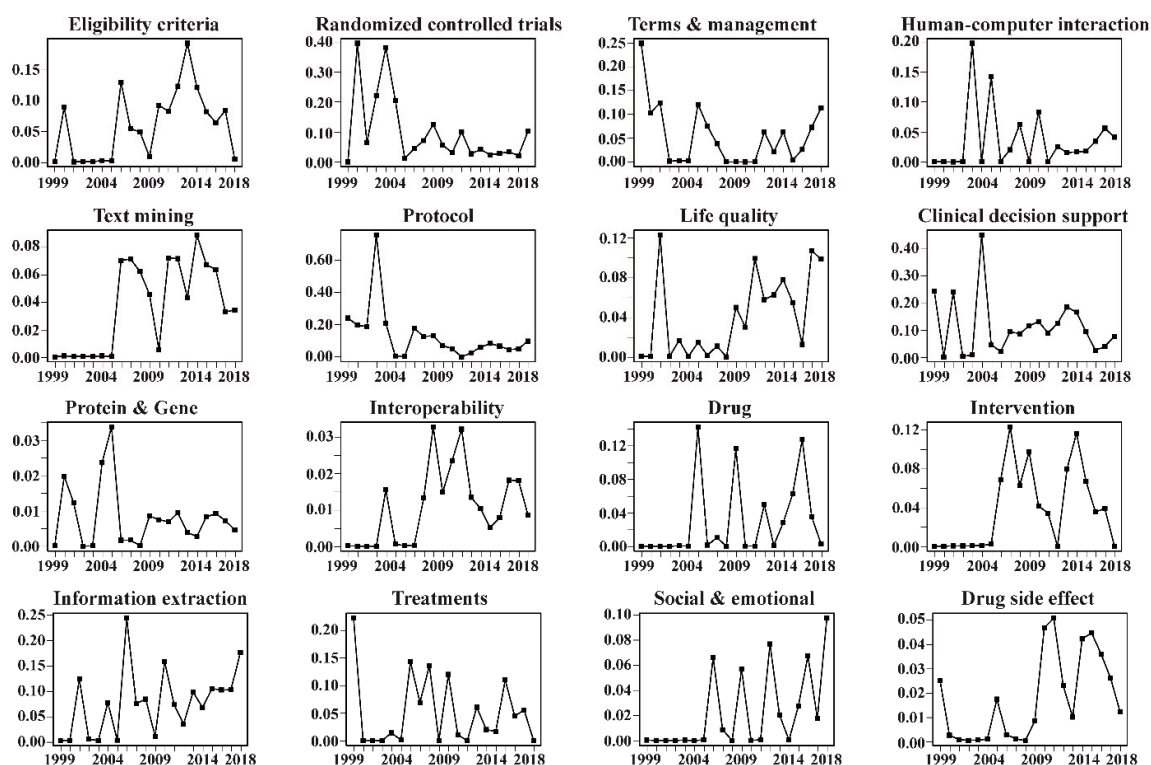
Figure 7. Temporal trends of the 16 topics in the whole corpus (2009–2018).

6.2. Topical Distributions Across Prolific Countries/Regions and Institutions

The topical distributions for prolific countries/regions and institutions with respect to NLP-enhanced clinical trial research were also explored, shown in Figure 8, in which the upper panel is for countries/regions, while the lower panel is for institutions.

From the perspective of countries/regions, several countries such as the USA, the UK, Australia, Canada, and China seemed to show balanced interest in diverse topics in the research. However, comparatively, the others showed particular interest in one or more topics, among which Interoperability was being favored most. In addition to Interoperability, there were several topics receiving much interest from particular countries. For example, France was more productive in Protocol. China had focused much on Clinical decision support and Eligibility criteria. Switzerland was very active in Information extraction.

From the perspective of institutions, there were several institutions mainly focusing on the topic Interoperability, including the *Technical University of Madrid*, *Philips Research*, *the National Institutes of Health*, *Foundation for Research and Technology-Hellas*, *University of Munster*, and *the University of Oxford*. As for the topic *Eligibility criteria*, there were three institutions showing a specific interest for it, including *Columbia University*, *Philips Research*, and *Vrije Universiteit Amsterdam*. In addition, *Stanford University* was particularly interested in *Protocol*, while *the University of Amsterdam* was particularly concerned about *Life quality*.
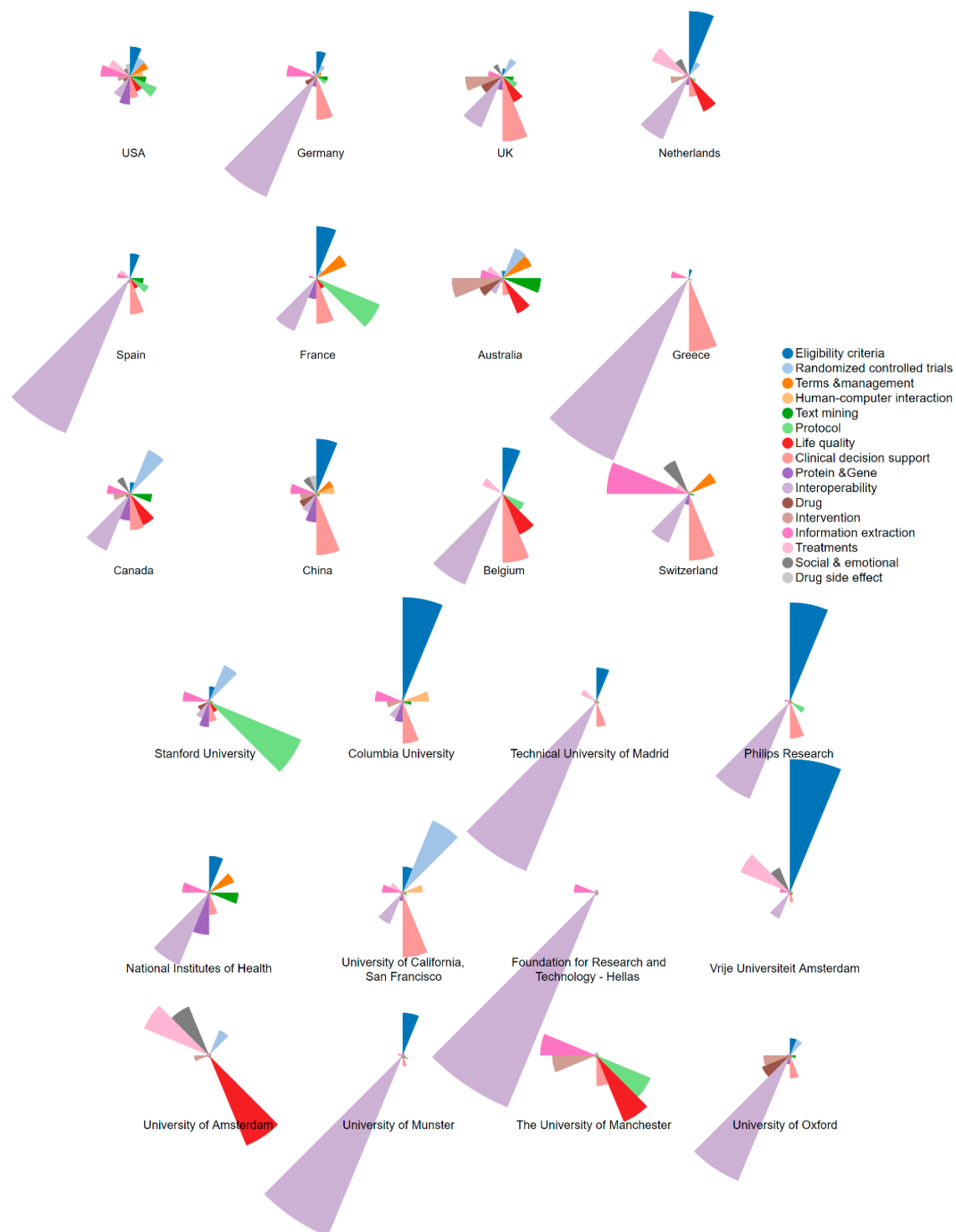
Figure 8. Topic proportion distributions across the prolific countries/regions and institutions.

6.3. Topic Proportion Distributions by Year and Topic Clustering

Figure 9 shows the topic proportion distributions by year in the NLP-enhanced clinical trial research literature. Overall, the results had shown a diversity in research topics with time, with more topics being discussed more among authors. In previous years, only several topics received particularly close attention, for example, *Randomized controlled trials* in 2000, *Protocol* in 2002, and *Clinical decision support* in 2004. However, with time, research interest among authors showed more balance in different aspects of the research.
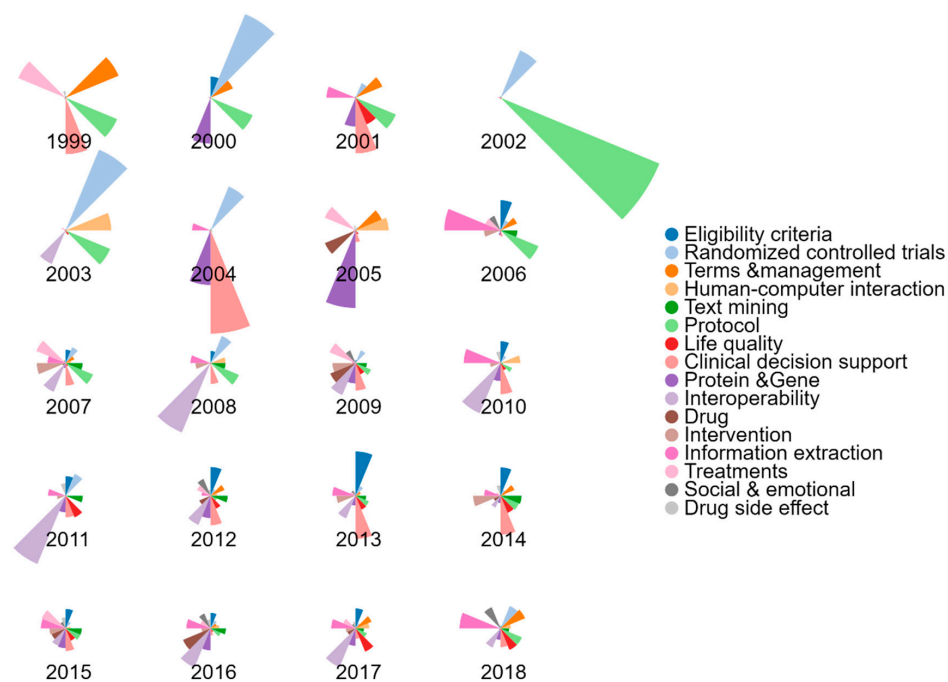
Figure 9. Topic proportion distributions by year in the NLP-enhanced clinical trial research.

Figure 10 shows the results of the hierarchical clustering based on document-level similarity. Such analysis provided insights into potential inter-topic research. For example, *Eligibility criteria* and *Clinical decision suppor*t had a high document-level similarity, indicating that articles with a high relevance to *Eligibility criteria* also had higher relevance to *Clinical decision support*. Thus, document-level similarity clustering had the potential to graph emerging trends of the NLP-enhanced clinical trial research during the study period.
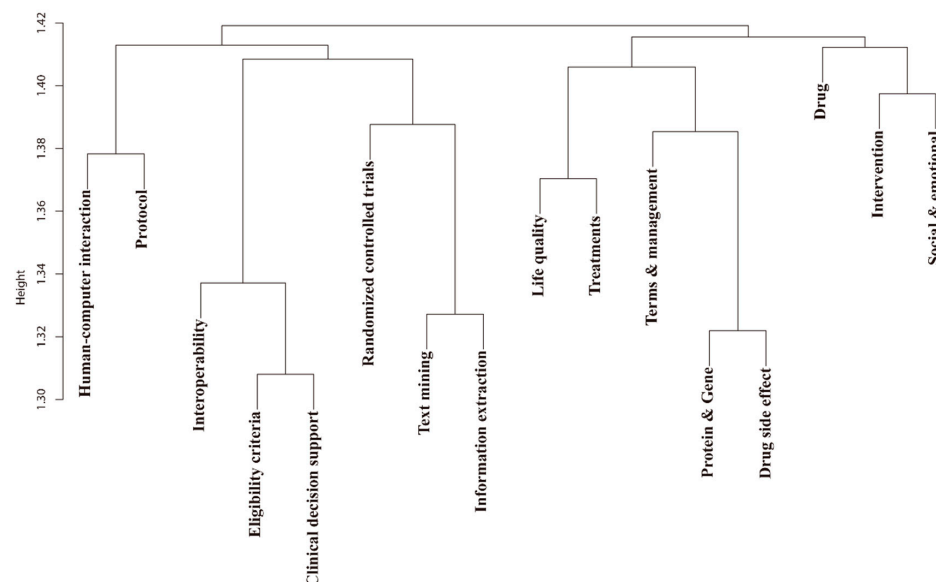
Figure 10. Hierarchical clustering results based on the document-topic distribution matrix.

7. Discussions

In this study, based on 451 articles retrieved from the WoS, PubMed, and Scopus databases, we displayed a thorough explanation of a bibliometric analysis of the NLP-enhanced clinical trials research

domain. Results obtained provided answers to the six research questions, based on which some critical issues were highlighted, as follows.

7.1. Insights From Performance Analysis and Scientific Collaboration Analysis

The analysis of the trend of the annual publication demonstrated that, overall, the annual publication of the NLP-enhanced clinical trial research had experienced a dramatic increase, although with a slight decrease in the final year, which was mainly caused by the time delay for the publication to be indexed by the database. The results significantly highlighted increasing attention and interest shown toward the NLP-enhanced clinical trial research from academia, and it was of high potential to continue to be more popular. The results of the analysis of the publication sources identified the leading position of the *JBI*, *JAMIA*, and *AMIA Symposium* in terms of H-index. *JBI* was also the most prolific journal in the publication of the NLP-enhanced clinical trial research. *JAMIA* was particularly famous for the impact and influence of its publication, with an ACP value up to 59. The results of the analyses of country/region, institution, and author identified the USA, Stanford University, and Ida Sim as the most influential ones in terms of H-index in the NLP-enhanced clinical trial research.

The scientific collaboration analysis between countries/regions, institutions, and authors highlighted that the USA and the UK, the *Technical University of Madrid* and *Philips Research*, as well as *Simona Carini* and *Ida Sim,* were the closest collaborators in the NLP-enhanced clinical trial research. The analyses also demonstrated the close collaboration between institutions and authors, particularly those from the USA. However, it was highly suggested that cross-region and cross-country collaboration could be strengthened to take advantage of regional research advantages.

7.2. Insights From Science Mapping Analysis

With the use of SciMAT, research themes were analyzed by exploring keywords and the evolution in each period of time. Four major thematic areas in NLP-enhanced clinical trial research were identified, including *Human-focused issues*, *Systems and applications*, *Clinical trials processing*, as well as *Diabetes-related issues*. First, within the theme *Human-focused issue*, there were three keywords, including *HUMAN*, *MEDICAL RECORD*, *PROCEDURE*, *FEMALE*, and *COMPUTER-ASSISTED*, where *HUMAN* was the largest one. EMRs are about the collection and description of events and patients' health history, concerning their interaction with clinical trials [85]. With the wide adoption of EMR databases in healthcare, an extraordinary amount of clinical data in electronic forms have been produced. There was a growing trend toward the collaborations between institutions from different regions on medical research with the use of EMRs [86]. However, such enormous unstructured medical texts had also become a major challenge for the enrollment of patients to test targeted therapies, which is very time- and energy-consuming [87]. Thus, many scholars had been attempting to advance clinical research by reducing the time spent on screening, identifying, and recruiting patients to realize automatic information extraction and facilitate the decision-making process.

Second, within the theme Systems and applications, there were seven keywords, including *SYSTEM*, *MEDICAL APPLICATION*, *GENETICS*, *MACHINE LEARNING*, *FACTURAL DATABASE*, *ELECTRONIC HEALTH RECORD*, and ARTIFICIAL INTELLIGENCE, where *ARTIFICIAL INTELLIGENCE* was the largest one. Artificial intelligence techniques had been widely adopted for the processing of clinical trial data. For example, with the use of semantic, structural, and NLP techniques, Dhuliawala et al. [88] proposed an automatic system for transforming activity schedule tables within clinical trial protocols from the portable document format into machine-interpretable forms, Their work was the first effort to adopt the artificial intelligence method for extracting procedural logic within clinical trial texts and creating a protocol knowledge base. Alonso et al. [89] aimed to apply AI techniques to provide support to scientists engaging in bio-nutritional trials. In addition, there were scholars focusing on the application of machine learning techniques for clinical trial data processing. For example, with the use of discriminative sequence classifiers for named entity recognition, Hassanpour and Langlotz [90] proposed a machine learning system for radiology reports annotation and report

contents extraction. With the use of logistic regression and decision trees, Bakal et al. [91] focused on accurately predicting new treatment and causative relationships between biomedical entities based on only the semantic graph pattern features obtained from biomedical knowledge graphs.

Third, within the theme Clinical trials processing, there were seven keywords, including *ACCURACY*, *RANDOMIZED TRIAL*, *RANDOMIZED CONTROLLED TRIAL*, *MEDICAL INFORMATICS*, *STANDARD*, *NATURAL LANGUAGE PROCESSING*, and *ELIGIBILITY CRITERION*, where *RANDOMIZED TRIAL* was the largest one. One critical application of NLP in clinical trial research is to resolve eligibility criteria for recruiting patients onto clinical trials. To evaluate patients' eligibility for a particular clinical trial, it usually requires manual evaluation of the coincidence between each eligibility criterion and the condition of the patient, which is very time-consuming [92]. Thus, many scholars had been devoted to facilitate the assessment process. For example, using information extraction techniques, Jonnalagadda et al. [87] proposed an automatic method for converting unstructured texts into structured ones. Such an approach was cross-referenced against eligibility criteria using a rule-based system for determining the qualification of patients for a particular clinical trial. Furthermore, standardization is a critical issue in clinical trial research. It aims to transform useful information hidden within clinical trial texts into machine-interpretable data to facilitate its applications in clinical studies (e.g., patient screening, trial recruitment, trial duplication detection, and clinical decision support [93]). Boscá et al. [94] worked on the task of obtaining semantically interoperable genetic testing reports to not only enhance communication but also facilitate clinical decision support. van Leeuwen et al. [93] proposed a standards-driven trial metadata repository for capturing structured trial information, application-related formalisms, as well as execution logic, to support various relevant applications.

Finally, within the theme Diabetes-related issues, there were three keywords, including *COMPUTER PROGRAM*, *OUTCOME*, and *DIABETES MELLITUS*, where *COMPUTER PROGRAM* was the largest one. Nowadays, diabetes mellitus has gradually become a worldwide disease, serving as the main cause of morbidity and mortality [95]. Caballero-Ruiz et al. [96] proposed an innovative system named Sinedie to enhance the accessibility of specialized healthcare aid, protect patients from unnecessary displacements, facilitate the process of patient evaluation, and avoid the gestational diabetes negative effects. Cole-Lewis et al. [97] aimed to propose an extendible knowledge base concerning the self-management of diabetes to serve as a foundation for supporting patient-centered decisions. The knowledge base was designed with the participation of academic diabetes educators using knowledge acquisition approaches and was further verified by a scenario-driven method.

7.3. Important Research Topics Identified by STM

Within the 16 identified research topics, there were several critical topics worthy of being highlighted, all of which were frequently discussed among authors regarding the NLP-enhanced clinical trial research.

First of all, *Interoperability* accounted for the largest proportion (13.21%), and it had also received increasing interest and attention from the research community, indicating its important role in the research area. Currently, there were generally several major problems of medical documentation, including a lack of transparency, an overwhelming amount of medical contents to be documented, and missing interoperability [98]. As indicated by Priyatna et al. [99], semantic interoperability was important while conducting post-genomic clinical trials in which different affiliations cooperate. There were a great number of scholars attempting to solve such a challenge. For example, Alonso-Calvo et al. [100] proposed a standards-driven method to guarantee semantic integration to facilitate clinico-genomic clinical trials analysis, by developing semantic interoperability based on various well-defined biomedical standards. Gazzarata et al. [101] adopted a service-centered architecture paradigm to facilitate technical interoperability.

The second important topic was *Information extraction*, accounting for a proportion of 9.36%. The reuse of EHRs can facilitate scientific research and patient recruitment for clinical trials.

Unfortunately, much information is hidden in unstructured texts such as discharge letters. Thus, it is required to be processed and transformed into structured data through the use of information extraction, which can be a time-consuming and labor-intensive task [102]. There were scholars working on such issues to facilitate clinical trial research. For example, Patterson et al. [103] proposed an NLP system with the use of dictionaries, rules, and patterns to extract measures of heart function recorded in echocardiogram reports. Further, by using a Curated semantic bootstrapping, a novel disambiguation approach based on semantic constraints, and a scalable framework, they enabled the proposed system to process large-scale clinical data. Milosevic et al. [104] investigated the feasibility of information extraction from tables that contained essential information such as characteristics of clinical trials within academic literature. They proposed a rule-driven approach to decompose tables into cell-level structures and to further extract information from them.

In addition, *Eligibility criteria* were also an essential issue in the NLP-enhanced clinical trial research. An essential step for discovering novel treatments for diseases is to match subjects with potential clinical trials. Eligibility criteria for clinical trials are inclusion and exclusion criteria for clinical trials and are written in the format of free text. Standardization of the eligibility criteria is able to improve both the efficiency and accuracy of the recruitment process [105]. As indicated by Milian et al. [106], the automated interpretation and the evaluation of patients' eligibility were challenging as the eligibility criteria of clinical trials is typically expressed as free text. There were many scholars trying to seek effective ways to facilitate the process.

7.4. Developing Trends of Research Foci

Results from the Mann–Kendall trend test (Table 10) and topic temporal trends visualization (Figure 7) provided several insights into the evolution trends of the NLP-enhanced clinical trial research. On the whole, most of the 16 identified topics had shown a growing tendency, with some topics becoming important. In addition to the three topics discussed in Section 5.1, there were several other topics that had experienced an increase in studies by authors, for example, *Life quality*, *Social & emotional*, *Text mining*, and *Drug side effect*.

First, since about 2009, the topic Life quality started to receive increasing attention among authors, although with a decrease in 2016. Such a tendency indicated growing concern on the quality of life within the NLP-enhanced clinical trial research community. For example, aiming to improve the quality of cardiac rehabilitation care, van Engen-Verheul et al. [107,108] first adopted concept mapping to determine factors for the successful implementation of web-driven quality improvement systems based on audit and feedback intervention with outreach visits, and they then evaluated the effectiveness of the proposed system.

Second, the topic *Social & emotional*, overall, had enjoyed a continually increasing attention since about 2005. As indicated by Chen [109], the internet is widely adopted to search health information and to seek social support. In addition, the large volume of information on the internet requires an effective tool or method for entity resolution in biomedical records to identify name abbreviations for diseases or various symptoms for the same diseases [110]. Thus, social media data have been widely adopted to support clinical trial research, particularly in terms of emotional disclosure analysis. Emotional disclosure refers to "the process of writing or talking about personally stressful or traumatic events and has been associated with improvements in mood and immune parameters in clinical and non-clinical groups [111] (Page 1)." For example, by adopting automatic sentiment analysis, Livas et al. [112] investigated the popularity and content of Invisalign patient testimonials on YouTube, as well as comments' sentiments.

Third, *Text mining* had also experienced a growth in research interest among scholars devoted to the NLP-enhanced clinical trial research, particularly since about 2005, although with an obvious decrease in about 2010. For example, Bekhuis and Demner-Fushman [113] aimed to investigate the following three issues: (1) Whether machine learning classifiers were able to facilitate the screening process of nonrandomized studies eligible for a systematic review on a particular topic; (2) whether

the performance of classifiers varied with optimization; (3) whether it was possible to reduce the number of citations for screening. With the use of a binary classification approach based on a support vector machine and a multi-layered sequence labeling approach based on conditional random fields, Li et al. [114] conducted evaluations of the use of machine learning approaches, binary classification, and sequence labeling to detect the medication-attribute linkage in two clinical corpora.

In addition, the topic *Drug side effect* started to gain increasing attention since about 2008. As an essential role in medicine, drugs are produced for treating, controlling, and preventing diseases. However, in addition to therapeutic effects, drugs also bring adverse effects. As indicated by Rajapaksha and Weerasinghe [115], in spite of post-marketing drug surveillance, adverse drug reactions (ADRs) have been one of the most usual causes of deaths worldwide. Even though with expensive clinical trials, it is still difficult to discover all types of ADRs, which can be troublesome for both customers and health care providers. There are many scholars devoted to the issues of drug side effects. For example, with the use of a probabilistic topic model, Bisgin et al. [116] proposed a drug repositioning method to make use of complete safety information as compared to the conventional method.

7.5. Insights Into Potential Inter-Topical Research

Several potential inter-topical research directions were identified from the results of document-topic hierarchical clustering (Figure 10). For example, there was a trend toward research on both *Human–computer interaction and Protocol*, particularly protocol eligibility criteria. For example, Petkov et al. [117] found the optimum approach based on a clinical trial eligibility database integrated system for rapidly identifying patients with metastatic diseases by automatically screening unstructured radiology reports. In addition, the proposed approach could be employed at different affiliations to improve the screening efficiency of clinical trial eligibility.

Second, a potential research direction was brought by *Eligibility criteria* and *Clinical decision support*. By proposing a novel system for clinical trial processing, Huang et al. [118] aimed to not only realize the interoperability by semantically integrating heterogeneous data into clinical trials but also enable services of automated reasoning and data processing in clinical trials decisions supporting systems.

Third, there was also a tendency for the application combinations of *Text mining* and *Information extraction* to the clinical trial research. For example, Oellrich et al. [119] compared the performances of four existing concept recognition systems, including the clinical text analysis and knowledge extraction system, the National Center for Biomedical Ontology annotator, the biomedical concept annotation system, and MetaMap.

Fourth, there was also a tendency for combined research on *Protein & Gene* and *Drug side effect*. For example, based on the assumption that machine learning was able to make predictions of unknown adverse reactions according to existing knowledge, Bean et al. [120] created a knowledge graph involving four kinds of nodes (i.e., drugs, protein targets, indications, and adverse reactions). They further proposed an innovative machine learning approach based on a simple enrichment test. Experiments indicated that the proposed approach performed tremendously well for the classification of known causes of adverse reactions.

In addition, a potential research direction was brought *by Intervention and Social & emotional*. Tarazona-Santabalbina et al. [121] conducted an interventional, controlled, and simple randomized study to examine the following two issues. The first was to determine whether a supervised-facility multicomponent exercise program was able to reverse frailty and improve cognition, emotional, and social networking. The second was to explore biological biomarkers of frailty by comparing it to an untrained and controlled population.

7.6. The Latest Trends in NLP-Enhanced Clinical Trials Research

We also provided several latest research achievements and outcomes in the field of the NLP-enhanced clinical trials. First, there was a trend toward drug side-effect signals. Cohort selection is an especially challenging task to which NLP and deep learning techniques can make a

significant difference [122]. For example, based on a collection of online posts from known statin users, Timimi et al. [123] adopted NLP techniques and hands-on linguistic analysis for the identification of drug side-effect signals. Experimental results indicated that statin users correlated statistically and significantly with memory impairment discussion.

Second, there was a growing trend in the development of innovative methods to select a cohort for clinical trials. For example, to address track one of the 2018 national NLP clinical challenges, Stubbs et al. [124] first selected criteria from current clinical trials to represent various NLP tasks, for example, concept extraction, temporal reasoning, and inference. They then labeled clinical narratives of 288 patients by examining their qualifications for the criteria. Vydiswaran et al. [125] proposed an innovative hybrid model by combining pattern-driven, knowledge-enriched, and feature weighting approaches. Segura-Bedmar and Raez [122] investigated several deep learning architectures for cohort selection for clinical trials, for example, simple convolutional neural networks (CNNs), deep CNNs, recurrent neural networks (RNNs), and CNNs–RNNs. Xiong et al. [126] proposed a hierarchical neural network consisting of five components. The first was a sentence representation using a CNN or the long short-term memory (LSTM) network. The second was a highway network for adjusting information flow. The third was a self-attention neural network for reweighting sentences. The fourth was a document representation using LSTM. The final was a fully connected neural network for determining whether each criterion was met or not. By employing an innovative generic rule-driven framework plugged in with lexical-, syntactic-, and meta-level, task-specific knowledge inputs, Chen et al. [127] proposed a novel and integrated rule-driven clinical NLP system. In addition, Chen et al. [128] proposed a clinical text representation model to learn patients' clinical trial criteria eligibility status for participating in cohort research. Experimental results demonstrated that the proposed model outperformed other medical knowledge-related learning architectures by around 3%.

Furthermore, there was also a growing trend for reducing information overload for trial seekers. For example, Liu et al. [129] proposed a novel system using dynamic questionnaires, by adopting a real-time and dynamic question generation approach, and dynamically updating the trial set through excluding unqualified trials according to users' responses to the generated questions.

Moreover, there was also a growing trend in the use of deep NLP of radiologic reports for measuring real-world oncologic outcomes (e.g., disease progressions and responses to therapies). For instance, Kehl et al. [130] aimed to explore whether deep NLP could extract related cancer outcomes from radiologic reports. Their study demonstrated that deep NLP had potential at speeding up the curation of related cancer outcomes, as well as facilitating fast learning from EHRs.

In addition, there was also a growing need for structuring and learning from the EHR using novel NLP techniques. For example, Thompson et al. [131] proposed an innovative method for assisting structuring, based on the exploration of how the most relevant words related to the class predicts a text from a position perspective.

8. Conclusions

This study reported the first endeavor toward the conduction of a detailed bibliometric analysis of the existing body of knowledge in the NLP-enhanced clinical trial research domain. Results obtained from the bibliometric analysis, science mapping, and STM analysis demonstrated several significant implications and insights into relevant scholars before pursuing further explorations in the NLP-enhanced clinical trial research field. From the methodology perspective, this study was the first to apply the STM approach to the NLP-enhanced clinical trial research, which was able to get rid of the limitations by using manual coding or word frequency analysis. Major findings and suggestions for future research on the NLP-enhanced clinical trials are summarized in the following subsections.

8.1. Major Findings

We summarized the major findings in the following aspects.

First, analysis of the annual trend of publications demonstrated that there was a dramatic growth in the research interests in the NLP-enhanced clinical trial research. In addition, research interests had become diverse among scholars.

Second, the USA, as well as several institutions such as *Stanford University*, *Columbia University*, and *the National Institutes of Health*, had contributed the most to the NLP-enhanced clinical trial research.

Third, the collaborations among European countries/regions in conducting NLP-enhanced clinical trial research were very close. The collaborations among institutions and authors from the same countries/regions were also close, particularly those from the USA.

Fourth, science mapping revealed several major research themes, including *Human-focused issues*, *Systems and applications*, *Clinical trials processing*, as well as *Diabetes-related issues*.

Fifth, STM identified 16 major topics, within which the most discussed included *Interoperability*, *Clinical decision support*, *Information extraction*, and *Eligibility criteria*.

Besides, several topics with an increase in research interest among authors had been identified, for example, *Life quality*, *Drug, Social & emotional*, *Information extraction*, *Eligibility criteria*, and *Drug side effect*.

In addition, several potential inter-topic research directions had also been uncovered, for example, *Human-computer interaction and Protocol*, *Eligibility criteria and Clinical decision support*, *Text mining and Information extraction*, *Protein & Gene* and *Drug side effect*, as well as *Intervention and Social & emotional*.

8.2. Suggestions for Future Research on the NLP-Enhanced Clinical Trials

Based on the major findings and discussions in Section 6, suggestions for NLP-enhanced clinical trials research in the future are proposed as follows.

First, it is important for scholars to continuously focus on the applications of text mining and information extraction techniques for the processing of clinical trial data.

Second, scholars from different regions and institutions are encouraged to enhance scientific communications and collaborations with advantageous regions in specific research areas to bolster scientific activities regarding the NLP-enhanced clinical trial research.

Third, scholars are encouraged to seek the potential to combine advanced deep learning techniques, for example, deep reinforcement learning, deep neural networks, recurrent neural network, and convolutional neural networks, with NLP techniques to deal with issues regarding clinical trials.

Besides, there is a growing need for improving and ensuring the quality of life in clinical trials research with the assistance of NLP techniques.

In addition, professionals, academics, and clinicians from different areas are encouraged to conduct collaborative research on potential promising issues, for example, *Eligibility criteria* and *Clinical decision support*, *Protein & Gene* and *Drug side effect*, as well as *Intervention* and *Social & emotional*.

Appendix A

Table A1. Final keywords list for NLP research in data retrieval.

"automatic summarization" or "bag of words" or "bayesian n-gram model" or "concept mapping" or "concept normalization" or "conceptual graphs" or "conceptual processing" or "contextual diversity" or "contextual semantics" or "corpus linguistics" or "dirichlet allocation" or "discourse interpretation" or "document ranking" or "information extraction" or "keyword ranked search" or "knowledge bases" or "knowledge graph" or "knowledge maps" or "knowledge reduction" or "knowledge representation" or "language model*" or "latent dirichlet allocation" or "latent semantic analysis" or "lemmatization" or "lexica creation" or "lexical decision" or "lexical knowledge" or "lexical semantics" or "linguistic information" or "linguistic model*" or "linguistic similarity" or "linguistic style match" or "multi-keyword search" or "named entity recognition" or "n-grams" or "part-of-speech" or "parts-of-speech" or "relation extraction" or "semantic analysis" or "semantic interpretation" or "semantic networks" or "semantic orientation" or "semantic processing" or "semantic role labeling" or "semantic roles" or "semantic tag" or "semantic topic analysis" or "semantic vector representation" or "semantical representation" or "semantics" or "semi-supervised learning" or "sentence boundary detection" or "sentence boundary disambiguation " or "sentence breaking" or "sentiment analysis" or "sentiment classification" or "similarity matching" or "smart semantic search" or "spatial similarity" or "string similarity" or "symbolic linguistic computing models" or "syntactic ambiguity resolution" or "syntactic analysis" or "terminology extraction" or "text analysis" or "text analytics" or "text classification" or "text clustering" or "text indexing" or "text simplification" or "text summarization" or "textual analysis" or "texture classification" or "thematic role-assignment" or "thermal-analysis" or "thermogravimetric analysis" or "tokenization" or "topic analysis" or "topic model*" or "word associations" or "word embedding" or "word sense disambiguation" or "word2vec" or "word-emotion associations" or "wordnet"

Table A2. Final keywords list for clinical trial research in data retrieval.

"clinical trial*" or "randomized controlled trial*" or "clinical research trial*" or "clinical research study" or "clinical research studies" or "clinical-trial*" or "randomized controlled-trial*" or "randomized trial*" or "controlled trial*" or "randomized-clinical-trial*" or "randomised trial*" or "randomized-trial*" or "randomization trial*" or "controlled-trial*" or "randomised-trial*"

Table A3. Full names of the abbreviations (in capitals) in Table 10.

Abbr.	Full names	Abbr.	Full names
ACGT	Advancing Clinico-genomic Trials on Cancer	HRA	Health-risk Assessments
ADE	Adverse Drug Events	HTA	Health Technology Assessment
ADHD	Attention-deficit/Hyperactivity Disorder	IMR	Iterative Model Reconstruction
ADRs	Adverse Drug Reactions	IMS	Intervention-management-system
AES	Adverse Events	IR	Iterative Reconstruction
AI	Artificial Intelligence	MDM	Medical Data Model
AnnSim	Annotation Similarity Measure	NLP	Natural Language Processing
APN	Advanced Practice Nurse	NRT	Nicotine Replacement Therapy
AskHERMES	A Clinical Question Answering System	OAE	Ontology of Adverse Events
BDETs	Bipolar Disorder Efficacy Trials	OAR	Organ-at-risk
CDA	Computerized Decision Aid	ODM	Operational Data Model
CDISC	Clinical Data Interchange Standards Consortium	PC	Public Consultations
CDSS	Clinical Decision-support System	PCROM	Primary Care Research Object Model
COIs	Conflicts of Interest	PECODR	Patient-population-problem, Exposure-intervention, Comparison, Outcome, Duration and Results
CR	Cardiac Rehabilitation	PGX	Pharmacogenomics
CT	Clinical Trials	PWPD	People with Parkinson's Disease
CTPT	Clinical Trial Participant Tracking	QA	Quality Assurance
DDI	Drug–drug Interactions	QI	Quality Improvement
DRPs	Drug-related Problems	RCTs	Randomized Controlled Trials
DTOME	Drug-target Interactome Tool	RD	Reading Disorder
EBM	Evidence-Based Medicine	RDCT	Reduced-dose Computed Tomography
EC	Eligibility Criteria	RDF	Resource Description Framework
ECIG	Electronic Cigarettes	SDCT	Standard-dose Computed Tomography
EFV	Efavirenz	SNOMED	The Systematized Nomenclature of Human and Veterinary Medicine
EMR	Electronic Medical Records	SR	Structured Reports
EOL	End-of-life	SVM	Support Vector Machine
FSTs	Frequent Semantic Tags	T2DM	Type 2 Diabetes Mellitus
HCI	Human–computer Interaction	TA	Temporal Abstraction
HCPs	Health Care Professionals	TBNs	Temporal Bayesian Networks
HER	Electronic Health Record	TDM	Therapeutic Drug Monitoring
HIV	Human Immunodeficiency Virus	UMLS	Unified Medical Language System

References

1. Roberts, M.E.; Stewart, B.M.; Airoldi, E.M. A model of text for experimentation in the social sciences. *J. Am. Stat. Assoc.* **2016**, *111*, 988–1003. [CrossRef]

2. Roberts, M.E.; Stewart, B.M.; Tingley, D.; Lucas, C.; Leder-Luis, J.; Gadarian, S.K.; Albertson, B.; Rand, D.G. Structural topic models for open-ended survey responses. *Am. J. Political Sci.* **2014**, *58*, 1064–1082. [CrossRef]

3. Jung, H.; Hong, S.; Park, J.; Park, M.; Sun, J.; Lee, S.; Ahn, J.; Ahn, M.; Park, K. MA19. 06 Successful Development of Realtime Automatically Updated Data Warehouse in Health Care (ROOT-S). *J. Thorac. Oncol.* **2019**, *14*, S328. [CrossRef]

4. Jonnalagadda, S.; Cohen, T.; Wu, S.; Gonzalez, G. Enhancing clinical concept extraction with distributional semantics. *J. Biomed. Inform.* **2012**, *45*, 129–140. [CrossRef] [PubMed]

5. Kreimeyer, K.; Foster, M.; Pandey, A.; Arya, N.; Halford, G.; Jones, S.F.; Forshee, R.; Walderhaug, M.; Botsis, T. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. *J. Biomed. Inform.* **2017**, *73*, 14–29. [CrossRef] [PubMed]

6. Heidarysafa, M.; Kowsari, K.; Odukoya, T.; Potter, P.; Barnes, L.E.; Brown, D.E. Women in ISIS Propaganda: A Natural Language Processing Analysis of Topics and Emotions in a Comparison with Mainstream Religious Group. Available online: https://www.researchgate.net/publication/337855242_Women_in_ISIS_Propaganda_A_Natural_Language_Processing_Analysis_of_Topics_and_Emotions_in_a_Comparison_with_Mainstream_Religious_Group (accessed on 1 March 2020).

7. Chen, T.; Dredze, M.; Weiner, J.P.; Hernandez, L.; Kimura, J.; Kharrazi, H. Extraction of geriatric syndromes from electronic health record clinical notes: Assessment of statistical natural language processing methods. *JMIR. Med. Inform.* **2019**, *7*, e13039. [CrossRef] [PubMed]

8. Bustos, A.; Pertusa, A. Learning eligibility in cancer clinical trials using deep neural networks. *Appl. Sci.* **2018**, *8*, 1206. [CrossRef]

9. Spyns, P. Natural language processing in medicine: An overview. *Methods Inf. Med.* **1996**, *35*, 285–301. [CrossRef] [PubMed]

10. Friedman, C.; Alderson, P.O.; Austin, J.H.; Cimino, J.J.; Johnson, S.B. A general natural-language text processor for clinical radiology. *J. Am. Med. Inform. Assoc.* **1994**, *1*, 161–174. [CrossRef] [PubMed]

11. Demner-Fushman, D.; Elhadad, N. Aspiring to unintended consequences of natural language processing: A review of recent developments in clinical and consumer-generated text processing. *Yearb. Med. Inform.* **2016**, *25*, 224–233.

12. Zhang, J.; Kowsari, K.; Harrison, J.H.; Lobo, J.M.; Barnes, L.E. Patient2vec: A personalized interpretable deep representation of the longitudinal electronic health record. *IEEE Access* **2018**, *6*, 65333–65346. [CrossRef]

13. Zhang, J.; Gong, J.; Barnes, L. HCNN: Heterogeneous convolutional neural networks for comorbid risk prediction with electronic health records. In Proceedings of the 2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE), 2017, Philadelphia, PA, USA, 17–19 July 2017.

14. Bian, J.; Barnes, L.E.; Chen, G.; Xiong, H. Early detection of diseases using electronic health records data and covariance-regularized linear discriminant analysis. In Proceedings of the 2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), 2017, Orlando, FL, USA, 16–19 February 2017.

15. Löbe, M.; Stäubert, S.; Goldberg, C.; Haffner, I.; Winter, A. Towards Phenotyping of Clinical Trial Eligibility Criteria. *Stud. Health Technol. Inform.* **2018**, *248*, 293–299. [PubMed]

16. Cowie, M.R.; Blomster, J.I.; Curtis, L.H.; Duclaux, S.; Ford, I.; Fritz, F.; Goldman, S.; Janmohamed, S.; Kreuzer, J.; Leenay, M. Electronic health records to facilitate clinical research. *Clin. Res. Cardiol.* **2017**, *106*, 1–9. [CrossRef] [PubMed]

17. Huhdanpaa, H.T.; Tan, W.K.; Rundell, S.D.; Suri, P.; Chokshi, F.H.; Comstock, B.A.; Heagerty, P.J.; James, K.T.; Avins, A.L.; Nedeljkovic, S.S. Using Natural Language Processing of Free-Text Radiology Reports to Identify Type 1 Modic Endplate Changes. *J. Digit. Imaging* **2018**, *31*, 84–90. [CrossRef] [PubMed]

18. Sheikhalishahi, S.; Miotto, R.; Dudley, J.T.; Lavelli, A.; Rinaldi, F.; Osmani, V. Natural language processing of clinical notes on chronic diseases: Systematic review. *JMIR Med. Inform.* **2019**, *7*, e12239. [CrossRef] [PubMed]

19. Dreisbach, C.; Koleck, T.A.; Bourne, P.E.; Bakken, S. A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. *Int. J. Med. Inform.* **2019**, *125*, 37–46. [CrossRef]

20. Gonzalez-Hernandez, G.; Sarker, A.; O'Connor, K.; Savova, G. Capturing the patient's perspective: A review of advances in natural language processing of health-related text. *Yearb. Med. Inform.* **2017**, *26*, 214–227. [CrossRef]

21. Névéol, A.; Zweigenbaum, P. Clinical natural language processing in 2014: Foundational methods supporting efficient healthcare. *Yearb. Med. Inform.* **2015**, *24*, 194–198. [CrossRef]

22. Luo, Y.; Thompson, W.K.; Herr, T.M.; Zeng, Z.; Berendsen, M.A.; Jonnalagadda, S.R.; Carson, M.B.; Starren, J. Natural language processing for EHR-based pharmacovigilance: A structured review. *Drug Saf.* **2017**, *40*, 1075–1089. [CrossRef]

23. Koleck, T.A.; Dreisbach, C.; Bourne, P.E.; Bakken, S. Natural language processing of symptoms documented in free-text narratives of electronic health records: A systematic review. *J. Am. Med. Inform. Assoc.* **2019**, *26*, 364–379. [CrossRef]

24. Velupillai, S.; Mowery, D.; South, B.R.; Kvist, M.; Dalianis, H. Recent advances in clinical natural language processing in support of semantic analysis. *Yearb. Med. Inform.* **2015**, *24*, 183–193.

25. Gutiérrez-Salcedo, M.; Martínez, M.Á.; Moral-Munoz, J.; Herrera-Viedma, E.; Cobo, M.J. Some bibliometric procedures for analyzing and evaluating research fields. *Appl. Intell.* **2018**, *48*, 1275–1287. [CrossRef]

26. De la Flor-Martínez, M.; Galindo-Moreno, P.; Sánchez-Fernández, E.; Piattelli, A.; Cobo, M.J.; Herrera-Viedma, E. H-classic: A new method to identify classic articles in Implant Dentistry, Periodontics, and Oral Surgery. *Clin. Oral Implant. Res.* **2016**, *27*, 1317–1330. [CrossRef] [PubMed]

27. De Maio, C.; Fenza, G.; Loia, V.; Parente, M. Biomedical data integration and ontology-driven multi-facets visualization. In Proceedings of the 2015 International Joint Conference on Neural Networks (IJCNN); Killarney, Ireland, 12–16 July 2015; pp. 1–8.

28. Houari, N.S.; Taghezout, N. Integrating Agents into a Collaborative Knowledge-based System for Business Rules Consistency Management. *Int. J. Interact. Multimed. Artif. Intell.* **2016**, *4*, 61–72. [CrossRef]

29. Settouti, N.; Bechar, M.E.A.; Chikh, M.A. Statistical comparisons of the top 10 algorithms in data mining for classification task. *Int. J. Interact. Multimed. Artif. Intell.* **2016**, *4*, 46–51.

30. Cobo, M.J.; Martínez, M.A.; Gutiérrez-Salcedo, M.; Fujita, H.; Herrera-Viedma, E. 25years at Knowledge-Based Systems: A bibliometric analysis. *Knowl. Based Syst.* **2015**, *80*, 3–13. [CrossRef]

31. Noyons, E.C.; Moed, H.F.; Luwel, M. Combining mapping and citation analysis for evaluative bibliometric purposes: A bibliometric study. *J. Am. Soc. Inf. Sci.* **1999**, *50*, 115–131. [CrossRef]

32. Moed, H.F.; Glänzel, W.; Schmoch, U. Handbook of quantitative science and technology research. In *the Use of Publication and Patent Statistics in Studies of S&T Systems*; Springer: Berlin/Heidelberg, Germany, 2004.

33. Cobo, M.J.; Chiclana, F.; Collop, A.; de Ona, J.; Herrera-Viedma, E. A bibliometric analysis of the intelligent transportation systems research based on science mapping. *IEEE Trans. Intell. Transp. Syst.* **2013**, *15*, 901–908. [CrossRef]

34. Grajzl, P.; Murrell, P. Toward understanding 17th century English culture: A structural topic model of Francis Bacon's ideas. *J. Comp. Econ.* **2019**, *47*, 111–135. [CrossRef]

35. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.

36. Blei, D.M. Probabilistic topic models. *Commun. ACM* **2012**, *55*, 77–84. [CrossRef]

37. Lester, C.A.; Wang, M.; Vydiswaran, V.V. Describing the patient experience from Yelp reviews of community pharmacies. *J. Am. Pharm. Assoc.* **2019**, *59*, 349–355. [CrossRef] [PubMed]

38. Chen, X.; Zou, D.; Cheng, G.; Xie, H. Detecting latent topics and trends in educational technologies over four decades using structural topic modeling: A retrospective of all volumes of computer & education. *Comput. Educ.* **2020**. [CrossRef]

39. Chen, X.; Zou, D.; Xie, H. Fifty years of British Journal of Educational Technology: A topic modeling based bibliometric perspective. *Br. J. Educ. Technol.* **2020**. [CrossRef]

40. Chen, X.; Yu, G.; Cheng, G.; Hao, T. Research topics, author profiles, and collaboration networks in the top-ranked journal on educational technology over the past 40 years: A bibliometric analysis. *J. Comput. Educ.* **2019**, *6*, 563–585. [CrossRef]

41. Chen, X.; Lun, Y.; Yan, J.; Hao, T.; Weng, H. Discovering thematic change and evolution of utilizing social media for healthcare research. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 50. [CrossRef] [PubMed]

42. Hao, T.; Chen, X.; Li, G.; Yan, J. A bibliometric analysis of text mining in medical research. *Soft Comput.* **2018**, *22*, 7875–7892. [CrossRef]

43. Chen, X.; Liu, Z.; Wei, L.; Yan, J.; Hao, T.; Ding, R. A comparative quantitative study of utilizing artificial intelligence on electronic health records in the USA and China during 2008–2017. *BMC Med. Inform. Decis. Mak.* **2018**, *18*, 117. [CrossRef]

44. Chen, X.; Wang, S.; Tang, Y.; Hao, T. A bibliometric analysis of event detection in social media. *Online Inf. Rev.* **2019**, *43*, 29–52. [CrossRef]

45. Xie, H.; Chu, H.-C.; Hwang, G.-J.; Wang, C.-C. Trends and development in technology-enhanced adaptive/personalized learning: A systematic review of journal publications from 2007 to 2017. *Comput. Educ.* **2019**, *140*, 103599. [CrossRef]

46. Song, Y.; Chen, X.; Hao, T.; Liu, Z.; Lan, Z. Exploring two decades of research on classroom dialogue by using bibliometric analysis. *Comput. Educ.* **2019**, *137*, 12–31. [CrossRef]

47. Chen, X.; Xie, H.; Wang, F.L.; Liu, Z.; Xu, J.; Hao, T. A bibliometric analysis of natural language processing in medical research. *BMC Med. Inform. Decis. Mak.* **2018**, *18*, 14. [CrossRef] [PubMed]

48. Arici, F.; Yildirim, P.; Caliklar, Ş.; Yilmaz, R.M. Research trends in the use of augmented reality in science education: Content and bibliometric mapping analysis. *Comput. Educ.* **2019**, *142*, 103647. [CrossRef]

49. Cancino, C.A.; Amirbagheri, K.; Merigó, J.M.; Dessouky, Y. A bibliometric analysis of supply chain analytical techniques published in Computers & Industrial Engineering. *Comput. Ind. Eng.* **2019**, *137*, 106015.

50. Clare, S.M.; Hickey, G.M. Modelling Research Topic Trends in Community Forestry. *Small Scale For.* **2019**, *18*, 149–163. [CrossRef]

51. Lee, J.D.; Kolodge, K. Exploring trust in self-driving vehicles through text analysis. *Hum. Factors* **2019**, *62*, 260–277. [CrossRef] [PubMed]

52. Bennett, R.; Vijaygopal, R.; Kottasz, R. Willingness of people with mental health disabilities to travel in driverless vehicles. *J. Transp. Health* **2019**, *12*, 1–12. [CrossRef]

53. Chen, X.; Chen, B.; Zhang, C.; Hao, T. Discovering the recent research in natural language processing field based on a statistical approach. In Proceedings of the International Symposium on Emerging Technologies for Education, Cape Town, South Africa, 20–22 September 2017.

54. Chen, X.; Ding, R.; Xu, K.; Wang, S.; Hao, T.; Zhou, Y. A bibliometric review of natural language processing empowered mobile computing. *Wirel. Commun. Mob. Comput.* **2018**, *2018*, 1827074. [CrossRef]

55. Radev, D.R.; Joseph, M.T.; Gibson, B.; Muthukrishnan, P. A bibliometric and network analysis of the field of computational linguistics. *J. Am. Soc. Inf. Sci. Technol.* **2009**, *1001*. [CrossRef]

56. Romero, A.; Cortés, J.; Escudero, C.; López, J.; Moreno, J. Measuring the influence of clinical trials citations on several bibliometric indicators. *Scientometrics* **2009**, *80*, 747–760. [CrossRef]

57. Yang, G.-Y.; Wang, L.-Q.; Ren, J.; Zhang, Y.; Li, M.-L.; Zhu, Y.-T.; Luo, J.; Cheng, Y.-J.; Li, W.-Y.; Wayne, P.M. Evidence base of clinical studies on Tai Chi: A bibliometric analysis. *PLoS ONE* **2015**, *10*, e0120655. [CrossRef] [PubMed]

58. Li, X.; Yang, G.; Li, X.; Zhang, Y.; Yang, J.; Chang, J.; Sun, X.; Zhou, X.; Guo, Y.; Xu, Y. Traditional Chinese medicine in cancer care: A review of controlled clinical studies published in Chinese. *PLoS ONE* **2013**, *8*, e60338.

59. Tao, T.; Zhao, X.; Lou, J.; Bo, L.; Wang, F.; Li, J.; Deng, X. The top cited clinical research articles on sepsis: A bibliometric analysis. *Crit. Care* **2012**, *16*, R110. [CrossRef] [PubMed]

60. Hong, S.J.; Yoon, D.Y.; Cho, Y.K.; Yoon, S.J.; Moon, J.Y.; Baek, S.; Lim, K.J. Characteristics and quality of radiologic randomized controlled trials: A bibliometric analysis between 1995 and 2014. *Am. J. Roentgenol.* **2016**, *206*, 917–923. [CrossRef] [PubMed]

61. Rosas, S.R.; Kagan, J.M.; Schouten, J.T.; Slack, P.A.; Trochim, W.M. Evaluating research and impact: A bibliometric analysis of research by the NIH/NIAID HIV/AIDS clinical trials networks. *PLoS ONE* **2011**, *6*, e17428. [CrossRef]

62. Kim, S.; Sagong, H.S.; Kong, J.C.; Choi, J.-Y.; Lee, M.S.; Wieland, L.S.; Manheimer, E.; Shin, B.-C. Randomised clinical trials on acupuncture in the Korean literature: Bibliometric analysis and methodological quality. *Acupunct. Med.* **2014**, *32*, 160–166. [CrossRef]

63. Cramer, H.; Lauche, R.; Dobos, G. Characteristics of randomized controlled trials of yoga: A bibliometric analysis. *BMC Complement. Altern. Med.* **2014**, *14*, 328. [CrossRef]

64. Cramer, H.; Lauche, R.; Langhorst, J.; Dobos, G. Are Indian yoga trials more likely to be positive than those from other countries? A systematic review of randomized controlled trials. *Contemp. Clin. Trials* **2015**, *41*, 269–272. [CrossRef]

65. Bayram, B.; Limon, Ö.; Limon, G.; Hancı, V. Bibliometric analysis of top 100 most-cited clinical studies on ultrasound in the Emergency Department. *Am. J. Emerg. Med.* **2016**, *34*, 1210–1216. [CrossRef]

66. Zeng, Z.; Deng, Y.; Li, X.; Naumann, T.; Luo, Y. Natural language processing for EHR-based computational phenotyping. *IEEE Trans. Comput. Biol. Bioinform.* **2018**, *16*, 139–153. [CrossRef]

67. Zhang, Y.; Huang, Y.; Porter, A.L.; Zhang, G.; Lu, J. Discovering and forecasting interactions in big data research: A learning-enhanced bibliometric study. *Technol. Forecast. Soc. Chang.* **2019**, *146*, 795–807. [CrossRef]

68. Zhang, Y.; Chen, H.; Zhu, D. Semi-automatic technology roadmapping composing method for multiple science, technology, and innovation data incorporation. In *Anticipating Future Innovation Pathways Through Large Data Analysis*; Springer: Berlin/Heidelberg, Germany, 2016.

69. Adriaanse, L.S.; Rensleigh, C. Web of Science, Scopus and Google Scholar: A content comprehensiveness comparison. *Electron. Libr.* **2013**, *31*, 727–744. [CrossRef]

70. Mongeon, P.; Paul-Hus, A. The journal coverage of Web of Science and Scopus: A comparative analysis. *Scientometrics* **2016**, *106*, 213–228. [CrossRef]

71. Gao, W.; Chen, Y.; Liu, Y.; Guo, H.-C. Scientometric analysis of phosphorus research in eutrophic lakes. *Scientometrics* **2015**, *102*, 1951–1964. [CrossRef]

72. Wang, Z.; Chen, Y.; Cai, G.; Jiang, Z.; Liu, K.; Chen, B.; Jiang, J.; Gu, H. A bibliometric analysis of PubMed literature on middle east respiratory syndrome. *Int. J. Environ. Res. Public Health* **2016**, *13*, 583. [CrossRef]

73. Hassan, S.-U.; Haddawy, P.; Zhu, J. A bibliometric study of the world's research activity in sustainable development and its sub-areas using scientific literature. *Scientometrics* **2014**, *99*, 549–579. [CrossRef]

74. Kowsari, K.; Brown, D.E.; Heidarysafa, M.; Meimandi, K.J.; Gerber, M.S.; Barnes, L.E. Hdltex: Hierarchical deep learning for text classification. In Proceedings of the 2017 16th IEEE international conference on machine learning and applications (ICMLA), Cancun, Mexico, 18–21 December 2017.

75. Kowsari, K.; Jafari Meimandi, K.; Heidarysafa, M.; Mendu, S.; Barnes, L.; Brown, D. Text classification algorithms: A survey. *Information* **2019**, *10*, 150. [CrossRef]

76. Jiang, H.; Qiang, M.; Lin, P. A topic modeling based bibliometric exploration of hydropower research. *Renew. Sustain. Energy Rev.* **2016**, *57*, 226–237. [CrossRef]

77. Peng, B.; Guo, D.; Qiao, H.; Yang, Q.; Zhang, B.; Hayat, T.; Alsaedi, A.; Ahmad, B. Bibliometric and visualized analysis of China's coal research 2000–2015. *J. Clean. Prod.* **2018**, *197*, 1177–1189. [CrossRef]

78. Hirsch, J.E.; Buela-Casal, G. The meaning of the h-index. *Int. J. Clin. Health Psychol.* **2014**, *14*, 161–164. [CrossRef]

79. Serrat, O. Social network analysis. In *Knowledge Solutions*; Springer: Berlin/Heidelberg, Germany, 2017.

80. Small, H. Visualizing science by citation mapping. *J. Am. Soc. Inf. Sci.* **1999**, *50*, 799–813. [CrossRef]

81. Cobo, M.J.; López-Herrera, A.G.; Herrera-Viedma, E.; Herrera, F. Science mapping software tools: Review, analysis, and cooperative study among tools. *J. Am. Soc. Inf. Sci. Technol.* **2011**, *62*, 1382–1402. [CrossRef]

82. Cobo, M.J.; Herrera, F. An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the Fuzzy Sets Theory field. *J. Informetr.* **2011**, *5*, 146–166. [CrossRef]

83. Garfield, E. Scientography: Mapping the tracks of science. *Curr. Contents Soc. Behav. Sci.* **1994**, *7*, 5–10.

84. Cabeza-Ramírez, L.J.; Sánchez Cañizares, S.M.; Fuentes García, F.J. Past Themes and Tracking Research Trends in Entrepreneurship: A Co-Word, Cites and Usage Count Analysis. *Sustainability* **2019**, *11*, 1–32.

85. Canino, G.; Suo, Q.; Guzzi, P.H.; Tradigo, G.; Zhang, A.; Veltri, P. Feature selection model for diagnosis, electronic medical records and geographical data correlation. In Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, Seattle, WA, USA, 2–5 October 2016.

86. Wu, Y.; Lei, J.; Wei, W.-Q.; Tang, B.; Denny, J.C.; Rosenbloom, S.T.; Miller, R.A.; Giuse, D.A.; Zheng, K.; Xu, H. Analyzing differences between Chinese and English clinical text: A cross-institution comparison of discharge summaries in two languages. *Stud. Health Technol. Inform.* **2013**, *192*, 662.

87. Jonnalagadda, S.R.; Adupa, A.K.; Garg, R.P.; Corona-Cox, J.; Shah, S.J. Text mining of the electronic health record: An information extraction approach for automated identification and subphenotyping of HFPEF patients for clinical trials. *J. Cardiovasc. Transl. Res.* **2017**, *10*, 313–321. [CrossRef]

88. Dhuliawala, M.; Fay, N.; Gruen, D.; Das, A. What Happens When? Interpreting Schedule of Activity Tables in Clinical Trial Documents. In Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, Washington, DC, USA, 29 August–1 September 2018.

89. Alonso, O.; De La Peña, P.; Moreno, A.; Anzaldi, G.; Domingo, X. Data Mining and Query Answer techniques applied to a bio-nutritional trials focused Expert System. In Proceedings of the CCIA, Alicante, Spain, 24–26 October 2012.

90. Hassanpour, S.; Langlotz, C.P. Information extraction from multi-institutional radiology reports. *Artif. Intell. Med.* **2016**, *66*, 29–39. [CrossRef]

91. Bakal, G.; Talari, P.; Kakani, E.V.; Kavuluru, R. Exploiting semantic patterns over biomedical knowledge graphs for predicting treatment and causative relations. *J. Biomed. Inform.* **2018**, *82*, 189–199. [CrossRef]

92. Bucur, A.; Van Leeuwen, J.; Chen, N.-Z.; Claerhout, B.; De Schepper, K.; Pérez del Rey, D.; Alonso Calvo, R.; Pugliano, L.; Saini, K. Supporting patient screening to identify suitable clinical trials. *Stud. Health Technol. Inform.* **2014**, *205*, 823–827.

93. van Leeuwen, J.; Bucur, A.I.; Claerhout, B.; de Schepper, K.; Pérez-Rey, D.; Alonso-Calvo, R. BRIDG-based Trial Metadata Repository-Need for Standardized Machine Interpretable Trial Descriptions. In Proceedings of the HEALTHINF, Loire Valley, France, 3–6 March 2014.

94. Boscá, D.; Marco, L.; Burriel, V.; Jaijo, T.; Millán, J.M.; Levin, A.M.; Pastor, O.; Robles, M.; Maldonado, J.A. Genetic testing information standardization in HL7 CDA and ISO13606. In Proceedings of the MedInfo, Copenhagen, Denmark, 20–13 August 2013.

95. Fernandes, F.; Vicente, H.; Abelha, A.; Machado, J.; Novais, P.; Neves, J. Artificial neural networks in diabetes control. In Proceedings of the 2015 Science and Information Conference (SAI), London, UK, 28–30 July 2015.

96. Caballero-Ruiz, E.; García-Sáez, G.; Rigla, M.; Villaplana, M.; Pons, B.; Hernando, M.E. A web-based clinical decision support system for gestational diabetes: Automatic diet prescription and detection of insulin needs. *Int. J. Med. Inform.* **2017**, *102*, 35–49. [CrossRef] [PubMed]

97. Cole-Lewis, H.J.; Smaldone, A.M.; Davidson, P.R.; Kukafka, R.; Tobin, J.N.; Cassells, A.; Mynatt, E.D.; Hripcsak, G.; Mamykina, L. Participatory approach to the development of a knowledge base for problem-solving in diabetes self-management. *Int. J. Med. Inform.* **2016**, *85*, 96–103. [CrossRef] [PubMed]

98. Geßner, S.; Neuhaus, P.; Varghese, J.; Bruland, P.; Meidt, A.; Soto-Rey, I.; Storck, M.; Doods, J.; Dugas, M. The Portal of Medical Data Models: Where Have We Been and Where Are We Going? In Proceedings of the MedInfo, Hangzhou, China, 21–25 August 2017.

99. Priyatna, F.; Alonso-Calvo, R.; Paraiso-Medina, S.; Corcho, O. Querying clinical data in HL7 RIM based relational model with morph-RDB. *J. Biomed. Semant.* **2017**, *8*, 49. [CrossRef] [PubMed]

100. Alonso-Calvo, R.; Paraiso-Medina, S.; Perez-Rey, D.; Alonso-Oset, E.; van Stiphout, R.; Yu, S.; Taylor, M.; Buffa, F.; Fernandez-Lozano, C.; Pazos, A. A semantic interoperability approach to support integration of gene expression and clinical data in breast cancer. *Comput. Biol. Med.* **2017**, *87*, 179–186. [CrossRef]

101. Gazzarata, R.; Giannini, B.; Giacomini, M. A SOA-based platform to support clinical data sharing. *J. Healthc. Eng.* **2017**, *2017*, 1–25. [CrossRef]

102. Dietrich, G.; Krebs, J.; Fette, G.; Ertl, M.; Kaspar, M.; Störk, S.; Puppe, F. Ad hoc information extraction for clinical data warehouses. *Methods Inf. Med.* **2018**, *57*, e22–e29. [CrossRef]

103. Patterson, O.V.; Freiberg, M.S.; Skanderson, M.; Fodeh, S.J.; Brandt, C.A.; DuVall, S.L. Unlocking echocardiogram measurements for heart disease research through natural language processing. *BMC Cardiovasc. Disord.* **2017**, *17*, 151. [CrossRef]

104. Milosevic, N.; Gregson, C.; Hernandez, R.; Nenadic, G. Extracting Patient Data from Tables in Clinical Literature-Case Study on Extraction of BMI, Weight and Number of Patients. In Proceedings of the HEALTHINF, Rome, Italy, 21–23 February 2016.

105. Lee, Y.; Krishnamoorthy, S.; Dinakarpandian, D. A semantic framework for intelligent matchmaking for clinical trial eligibility criteria. *ACM Trans. Intell. Syst. Technol.* **2013**, *4*, 71. [CrossRef]

106. Milian, K.; Bucur, A.; van Harmelen, F.; ten Teije, A. Identifying Most Relevant Concepts to Describe Clinical Trial Eligibility Criteria. In Proceedings of the HEALTHINF, Barcelona, Spain, 11–14 February 2013.

107. van Engen-Verheul, M.M.; Peek, N.; Haafkens, J.A.; Joukes, E.; Vromen, T.; Jaspers, M.W.; de Keizer, N.F. What is needed to implement a web-based audit and feedback intervention with outreach visits to improve care quality: A concept mapping study among cardiac rehabilitation teams. *Int. J. Med. Inform.* **2017**, *97*, 76–85. [CrossRef]

108. van Engen-Verheul, M.M.; de Keizer, N.F.; van der Veer, S.N.; Kemps, H.M.; op Reimer, W.J.S.; Jaspers, M.W.; Peek, N. Evaluating the effect of a web-based quality improvement system with feedback and outreach visits on guideline concordance in the field of cardiac rehabilitation: Rationale and study protocol. *Implement. Sci.* **2014**, *9*, 780. [CrossRef]

109. Chen, Z.; Koh, P.W.; Ritter, P.L.; Lorig, K.; Bantum, E.O.C.; Saria, S. Dissecting an online intervention for cancer survivors: Four exploratory analyses of internet engagement and its effects on health status and health behaviors. *Health Educ. Behav.* **2015**, *42*, 32–45. [CrossRef] [PubMed]

110. Vidhya, K.; Soorya, R.; Saranavan, N.; Geetha, T.; Singaravelan, M. Entity resolution for symptom vs disease for top-K treatments. In Proceedings of the 2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 6–7 January 2017.

111. Byrne-Davis, L.; Wetherell, M.; Dieppe, P.; Weinman, J.; Byron, M.; Donovan, J.; Horne, R.; Brookes, S.; Vedhara, K. Emotional disclosure in rheumatoid arthritis: Participants' views on mechanisms. *Psychol. Health* **2006**, *21*, 667–682. [CrossRef]

112. Livas, C.; Delli, K.; Pandis, N. "My Invisalign experience": Content, metrics and comment sentiment analysis of the most popular patient testimonials on YouTube. *Prog. Orthod.* **2018**, *19*, 3. [CrossRef] [PubMed]

113. Bekhuis, T.; Demner-Fushman, D. Screening nonrandomized studies for medical systematic reviews: A comparative study of classifiers. *Artif. Intell. Med.* **2012**, *55*, 197–207. [CrossRef]

114. Li, Q.; Zhai, H.; Deleger, L.; Lingren, T.; Kaiser, M.; Stoutenborough, L.; Solti, I. A sequence labeling approach to link medications and their attributes in clinical notes and clinical trial announcements for information extraction. *J. Am. Med. Inform. Assoc.* **2012**, *20*, 915–921. [CrossRef]

115. Rajapaksha, P.; Weerasinghe, R. Identifying adverse drug reactions by analyzing Twitter messages. In Proceedings of the 2015 Fifteenth International Conference on Advances in ICT for Emerging Regions (ICTer), Colombo, Sri Lanka, 24–26 August 2015.

116. Bisgin, H.; Liu, Z.; Kelly, R.; Fang, H.; Xu, X.; Tong, W. Investigating drug repositioning opportunities in FDA drug labels through topic modeling. *BMC Bioinform.* **2012**, *13*, 1–9. [CrossRef]

117. Petkov, V.I.; Penberthy, L.T.; Dahman, B.A.; Poklepovic, A.; Gillam, C.W.; McDermott, J.H. Automated determination of metastases in unstructured radiology reports for eligibility screening in oncology clinical trials. *Exp. Biol. Med.* **2013**, *238*, 1370–1378. [CrossRef]

118. Huang, Z.; Ten Teije, A.; Van Harmelen, F. SemanticCT: A semantically-enabled system for clinical trials. In *Process Support and Knowledge Representation in Health Care*; Springer: Berlin/Heidelberg, Germany, 2013.

119. Oellrich, A.; Collier, N.; Smedley, D.; Groza, T. Generation of silver standard concept annotations from biomedical texts with special relevance to phenotypes. *PLoS ONE* **2015**, *10*, e0116040. [CrossRef]

120. Bean, D.M.; Wu, H.; Iqbal, E.; Dzahini, O.; Ibrahim, Z.M.; Broadbent, M.; Stewart, R.; Dobson, R.J. Knowledge graph prediction of unknown adverse drug reactions and validation in electronic health records. *Sci. Rep.* **2017**, *7*, 16416. [CrossRef]

121. Tarazona-Santabalbina, F.J.; Gómez-Cabrera, M.C.; Pérez-Ros, P.; Martínez-Arnau, F.M.; Cabo, H.; Tsaparas, K.; Salvador-Pascual, A.; Rodriguez-Manas, L.; Viña, J. A multicomponent exercise intervention that reverses frailty and improves cognition, emotion, and social networking in the community-dwelling frail elderly: A randomized clinical trial. *J. Am. Med. Dir. Assoc.* **2016**, *17*, 426–433. [CrossRef]

122. Segura-Bedmar, I.; Raez, P. Cohort selection for clinical trials using deep learning models. *J. Am. Med. Inform. Assoc.* **2019**, *26*, 1181–1188. [CrossRef] [PubMed]

123. Timimi, F.; Ray, S.; Jones, E.; Aase, L.; Hoffman, K. Patient-Reported Outcomes in Online Communications on Statins, Memory, and Cognition: Qualitative Analysis Using Online Communities. *J. Med. Internet Res.* **2019**, *21*, e14809. [CrossRef] [PubMed]

124. Stubbs, A.; Filannino, M.; Soysal, E.; Henry, S.; Uzuner, Ö. Cohort selection for clinical trials: n2c2 2018 shared task track 1. *J. Am. Med. Inform. Assoc.* **2019**, *26*, 1163–1171. [CrossRef] [PubMed]

125. Vydiswaran, V.; Strayhorn, A.; Zhao, X.; Robinson, P.; Agarwal, M.; Bagazinski, E.; Essiet, M.; Iott, B.E.; Joo, H.; Ko, P. Hybrid bag of approaches to characterize selection criteria for cohort identification. *J. Am. Med. Inform. Assoc.* **2019**, *26*, 1172–1180. [CrossRef]

126. Xiong, Y.; Shi, X.; Chen, S.; Jiang, D.; Tang, B.; Wang, X.; Chen, Q.; Yan, J. Cohort selection for clinical trials using hierarchical neural network. *J. Am. Med. Inform. Assoc.* **2019**, *26*, 1203–1208. [CrossRef]

127. Chen, L.; Gu, Y.; Ji, X.; Lou, C.; Sun, Z.; Li, H.; Gao, Y.; Huang, Y. Clinical trial cohort selection based on multi-level rule-based natural language processing system. *J. Am. Med. Inform. Assoc.* **2019**, *26*, 1218–1226. [CrossRef]

128. Chen, C.-J.; Warikoo, N.; Chang, Y.-C.; Chen, J.-H.; Hsu, W.-L. Medical knowledge infused convolutional neural networks for cohort selection in clinical trials. *J. Am. Med. Inform. Assoc.* **2019**, *26*, 1227–1236. [CrossRef]

129. Liu, C.; Yuan, C.; Butler, A.M.; Carvajal, R.D.; Li, Z.R.; Ta, C.N.; Weng, C. DQueST: Dynamic questionnaire for search of clinical trials. *J. Am. Med. Inform. Assoc.* **2019**, *26*, 1333–1343. [CrossRef]

130. Kehl, K.L.; Elmarakeby, H.; Nishino, M.; Van Allen, E.M.; Lepisto, E.M.; Hassett, M.J.; Johnson, B.E.; Schrag, D. Assessment of deep natural language processing in ascertaining oncologic outcomes from radiology reports. *JAMA Oncol.* **2019**, *5*, 1421–1429. [CrossRef]

131. Thompson, J.; Hu, J.; Mudaranthakam, D.P.; Streeter, D.; Neums, L.; Park, M.; Koestler, D.C.; Gajewski, B.; Jensen, R.; Mayo, M.S. Relevant Word Order Vectorization for Improved Natural Language Processing in Electronic Health Records. *Sci. Rep.* **2019**, *9*, 9253. [CrossRef]