How the clinical research community responded to the COVID-19 pandemic: An analysis of the COVID-19 clinical studies in ClinicalTrials.gov

Zhe He, PhD[1], Arslan Erdengasileng, MS[2], Xiao Luo, PhD[3], Aiwen Xing, MS[2], Neil Charness, PhD[4], Jiang Bian, PhD[5]

[1]School of Information, Florida State University, Tallahassee, Florida, USA;
[2]Department of Statistics, Florida State University, Tallahassee, Florida, USA;
[3]School of Engineering and Technology, Indiana University–Purdue University Indianapolis, Indianapolis, Indiana, USA;
[4]Department of Psychology, Florida State University, Tallahassee, Florida, USA;
[5]Department of Health Outcomes and Biomedical Informatics, University of Florida,Gainesville, Florida, USA

Corresponding Author:

Zhe He, PhD
School of Information
College of Communication and Information
Florida State University
142 Collegiate Loop
Tallahassee, Florida 32306-2100
zhe@fsu.edu
Phone: 001(850)644-5775

Word count: 4498

Abstract

Objective: In the past few months, a large number of clinical studies on the novel coronavirus disease (COVID-19) have been initiated worldwide to find effective therapeutics, vaccines, and preventive strategies for COVID-19. In this study, we aim to understand the landscape of COVID-19 clinical research and identify the issues that may cause recruitment difficulty or reduce study generalizability.

Methods: We analyzed 3,765 COVID-19 studies registered in the largest public registry - ClinicalTrials.gov, leveraging natural language processing and using descriptive, association, and clustering analyses. We first characterized COVID-19 studies by study features such as phase and tested intervention. We then took a deep dive and analyzed their eligibility criteria to understand whether these studies: (1) considered the reported underlying health conditions that may lead to severe illnesses, and (2) excluded older adults, either explicitly or implicitly, which may reduce the generalizability of these studies to the older adults population.

Results: Our analysis included 2295 interventional studies and 1470 observational studies. Most trials did not explicitly exclude older adults with common chronic conditions. However, known risk factors such as diabetes and hypertension were considered by less than 5% of trials based on their trial description. Pregnant women were excluded by 34.9% of the studies.

Conclusions: Most COVID-19 clinical studies included both genders and older adults. However, risk factors such as diabetes, hypertension, and pregnancy were under-represented, likely skewing the population that was sampled. A careful examination of existing COVID-19 studies can inform future COVID-19 trial design towards balanced internal validity and generalizability.

Keywords: COVID-19, clinical trial, eligibility criteria, natural language processing

Lay Summary

Since the outbreak of COVID-19 in early 2020, thousands of clinical studies have been conducted to evaluate the efficacy and safety of various types of treatments and vaccines in human. COVID-19 clinical studies play a crucial role in controlling the virus. Yet it is unclear what types of patients were considered by these studies. This study analyzed 3,765 COVID-19 clinical study summaries downloaded from a major clinical trial registry ClinicalTrials.gov. We employed natural language processing techniques to parse the study description and eligibility criteria of these studies and then performed descriptive and clustering analysis on the parsing results. We found that older adults were not systematically excluded but pregnant women were often excluded. It was also found that the known risk factors such as diabetes, hypertension, obesity, and asthma, which may lead to serious illnesses, were considered by less than 5% of the studies according to their study description and eligibility criteria. This study provides an evidence that natural language processing can be applied to examine the design of clinical studies and identify issues that may cause delays in patient recruitment and the lack of real-world population representativeness.

Introduction

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and the associated coronavirus disease (COVID-19) broke out in December 2019 and has quickly become a global pandemic with serious health and social consequences [1]. As of February 11, 2021, more than 107 million confirmed cases have been reported around the world and about one-fourth are from the U.S. [2]. Globally, more than 2 million people have died due to COVID-19 and 472,000 in the U.S. alone. In April 2020, the National Institutes of Health (NIH) launched the Accelerating COVID-19 Therapeutic Interventions and Vaccines (ACTIV) public-private partnership to prioritize and speed up the clinical evaluation of the most promising treatments and vaccines [3]. In

July 2020, NIH released its strategic plan for COVID-19 research to speed up the development of treatments, vaccines, and diagnostics [4]. Traditionally, it may take years to discover, develop, and evaluate a therapeutic agent; nevertheless, for COVID-19, the goal has been to compress the timeline to months while continuing to apply rigorous standards to ensure safety and efficacy. Strategies such as applying complex computer-generated models of SARS-Cov-2 and its biological processes to determine key interactions and pathways have been applied to develop therapeutic agents and vaccines (e.g., monoclonal antibodies) to neutralize the virus. A significant efforts have also been made to screen existing drugs approved for other indications to treat COVID-19 [4]. Recently, the COVID-19 vaccines developed and manufactured by Pfizer-BioNTech and Moderna have been approved to be administered in a few countries including the United Kingdom [5], Canada [6], and the U.S. [7].

Clinical studies, especially randomized controlled trials (RCT), are the gold standard for evaluating the efficacy and safety of a treatment. Regardless of the techniques (e.g., in vivo, in silico, or in vitro) used for drug discovery and development, the therapeutics and vaccines have to go through three phases of clinical trials to evaluate their efficacy and safety before approvals (e.g., by the U.S. Food and Drug Administration [FDA]) can be granted for mass production and use in the general population. In the past few months, many COVID-19 clinical studies have been launched around the world, leading to situations where studies have to compete for participants from the same pool of eligible participants. Trials such as those for the promising drug – *Remdesivir* – were suspended due to the lack of trial participants in China [8]. Other issues such as population representativeness are also critical. In the past, older adults are often excluded from clinical trials with overly restrictive exclusion criteria, which lead to concerns on the generalizability of those clinical studies across many disease domains [9]. A recent New York

Times article conjectured that older adults are left out from COVID-19 trials [10]. It is therefore important to understand the landscape of COVID-19 clinical research and further identify the gaps and issues that may cause delays in patient recruitment and the lack of real-world population representativeness, especially for older adults.

To date, 12 other studies have analyzed registered COVID-19 clinical studies [11-22] (see Supplementary Material I for details). For example, Wang et al. analyzed the basic characteristics and the drug interventions of 306 COVID-19 trials from ClinicalTrials.gov as of April 3, 2020 [11]. Paudi et al. analyzed 1,551 COVID-19 studies registered between March 1, 2020 and May 19, 2020 in ClinicalTrials.gov and focused on basic characteristics, primary and secondary outcomes, and study design [12]. Kim et al. evaluated the impact of frequently used quantitative eligibility criteria in 288 COVID-19 studies (as of June 18, 2020) on the recruitment and clinical outcomes using the EHR data of COVID-19 patients in Columbia University Irving Medical Center [22]. While most of these studies analyzed the basic characteristics of the included trials, some studies further analyzed the interventions [18-21], locations [13,14,18], data monitoring characteristics [14], timing of the registration and enrollment [15,21], outcomes [19,20], risk of biases [16], Medical Subject Heading (MeSH) keywords (Words or phrases that best describe the protocol) [19], sample size [17], and a subset of eligibility criteria [16,22]. Nonetheless, existing studies have not comprehensively investigated the categorical eligibility criteria and the consideration of known risk factors of severe illnesses in COVID-19 clinical studies. To do so, advanced approaches such as natural language processing (NLP) are necessary. As such, we can gain a better understanding of the landscape of COVID-19 research and answer a number of important research questions: (1) What eligibility criteria are used in COVID-19 clinical studies? (2) Further, as more COVID-19 cases have been identified and treated in the past 8 months, we have accumulated important knowledge on the underlying health conditions and other risk factors that may cause severe illness among COVID-19 patients (e.g., hypertension and diabetes) [23]. Have existing clinical studies considered these known risk factors? (3) Last but not the least, because of the concerns on study generalizability in older adults, it is of interest to assess whether the COVID-19 clinical studies excluded subjects with common chronic conditions that are prevalent in older adults, potentially leading to their underrepresentation.

In this study, we conducted a systematic analysis of the registered clinical studies on COVID-19 (as of November 27, 2020) from ClinicalTrials.gov to answer the aforementioned research questions. The contribution of this paper is multi-fold: (1) it systematically summarizes various important aspects of the COVID-19 clinical studies; (2) it identifies the research gaps on the risk factors related to serious illness caused by COVID-19; (3) it groups COVID-19 studies based on their eligibility criteria, and (4) it identifies salient exclusion criteria that may implicitly exclude older adults, who are most vulnerable and should be studied when evaluating the efficacy and safety of COVID-19 treatments and vaccines. Our findings could inform future trials designed for COVID-19 treatment and prevention and identify strategies to rapidly but appropriately stand up a large number of clinical studies for future pandemics similar to COVID-19.

Materials and Methods

Data Source

ClinicalTrials.gov, built and maintained by the U.S. National Library of Medicine, is the largest clinical study registry in the world [24]. In the U.S., all drugs and devices regulated by the U.S. Food and Drug Administration (FDA) are required to be registered on the ClinicalTrials.gov.

ClinicalTrials.gov is thus considered as the most comprehensive trial registry in the world and has been widely used for secondary analysis [25].

Dataset Acquisition and Processing

From ClinicalTrials.gov, we downloaded records of 4,028 clinical studies that are tagged with a condition of "COVID-19" or "SARS-CoV-2" on November 27, 2020. We excluded studies tagged with the study type "Expanded Access" and studies that were tagged as patient registries, leaving 3,765 records that met our inclusion criteria. The reason for excluding "Expanded Access" studies is that the original studies have been included. The reason for excluding patient registries is that they do not use selective eligibility criteria and are probably biased in different ways than clinical trials. We extracted the NCTID (an unique identifier of a study record), conditions, agency, agency class, brief summary, detailed summary, status, start date, eligibility criteria, enrollment, study phase, study type, intervention type, intervention name, study design (i.e., allocation, masking, observation model, time perspective), primary purpose, and endpoint classification. We split the eligibility criteria into inclusion criteria and exclusion criteria and further extracted individual criteria for natural language processing. To identify the top frequently tested drugs, we extracted the drugs information from the "intervention" field from the study record. We used QuickUMLS to normalize the drug names and removed the dosage information before analyzing their frequencies.

Consideration of Risk Factors in COVID-19 Clinical Studies

We first identified known risk factors of COVID-19 from online resources such as the Centers for Disease Control and Prevention (CDC) [26] and Mayo Clinic [27] (as of July 17, 2020). Then, we coded the risk factors with the concepts from Unified Medical Language System (UMLS). To do so, we used the risk factor terms as the input and identified their corresponding Concept Unique Identifiers (CUIs) using QuickUMLS [28] with the default setting (Jaccard similarity threshold > 0.8, all semantic types included). As a concept of the UMLS is associated with its synonyms from UMLS source ontologies, we were able to unify all the terms mentioned in the text. ***Table 1*** lists the risk factors that may lead to severe COVID-19 and their associated UMLS CUIs. This list was used as a dictionary in QuickUMLS [28] to identify risk factors from the study description. As reported in Soldaini et al. [28], QuickUMLS achieved better performance than MetaMap and cTAKES on a number of benchmark corpora. Nevertheless, we manually reviewed the risk factors extracted by QuickUMLS on a random sample of 100 clinical studies; and QuickUMLS achieved a precision of 91%[1]. We also identified the studies that used these risk factors in the inclusion/exclusion criteria using the parsing results of the eligibility criteria parsing tool [29] described below. T-test and analysis of variance (ANOVA) were employed to assess the association between the number of risk factors in the trial descriptions with the study type, intervention type, and primary purpose.

[1] We did not evaluate the recall due to lack of a benchmark dataset for trial description with all the risk factors annotated manually.

Table 1. UMLS CUIs for the risk factors for severe illness among COVID-19 patients reported by CDC and Mayo Clinic websites

Risk Factors	**UMLS CUIs**
Old age	C0231337, C1999167
Males	C0086582
Chronic kidney disease	C1561643, C4075517, C4553188, C4075526
COPD	C0024117
Lung cancer	C0684249, C0242379, C1306460
Weakened immune system from solid organ transplant (Weakened immune system 1)	C0029216, C0524930
Obesity	C0028754, C1963185
Serious heart conditions, such as heart failure, coronary artery disease, or cardiomyopathies	C0018802, C4554158, C0018801, C0010054, C1956346, C0878544, C0796094, C0020542
Sickle cell disease	C0002895
Asthma	C0004096
Neurological conditions, such as dementia	C0002395, C0011265, C0497327, C0014544, C0026769, C0455388, C1417325, C0030567, C0036572
Cerebrovascular disease (affects blood vessels and blood supply to the brain) such as stroke	C0678234, C1961121, C0549207, C1261287, C1522213, C0524466, C0038454, C0007282, C0595850, C0158570, C0002940
Cystic fibrosis	C0010674
Hypertension	C0020538, C1963138
Weakened immune system from blood or bone marrow transplant, immune deficiencies, HIV, use of corticosteroids, or use of other immune weakening medicines (Weakened immune system 2)	C0005961, C3540726, C3540727, C0021051, C0279026, C3539185, C3540725, C0001617, C1955133
Pregnancy	C0032961
Liver diseases	C0023895
Pulmonary fibrosis (having damaged or scarred lung tissues)	C4553408, C0034069
Smoking	C1881674, C1548578, C0037369,C0453996
Diabetes	C0011847, C0011849
Thalassemia	C0039730, C0002312

Analysis of Eligibility Criteria

Quantitative criteria: We used the open-source Valx tool [30] to extract and standardize the quantitative eligibility criteria from the COVID-19 studies. Valx is a system that can extract numeric expressions from free-text eligibility criteria and standardize them into a structured format. For example, from the inclusion criterion “BMI > 25 kg/m^2”, the variable name “BMI”, the comparison operator “>”, the threshold value “25”, and the measurement unit “kg/m^2” were extracted into 4 discrete fields. Valx is also able to recognize synonyms of a variable and convert the units to standard ones. According to [30], the F-scores for structuring HbA1c and glucose comparison statements for Type 2 diabetes trials are 97.8% and 92.3%, respectively. We then analyzed the frequency of the quantitative criteria and the threshold values used for patient eligibility determination.

Categorical eligibility features: To extract the categorical eligibility features from COVID-19 studies, we used a new eligibility criteria parsing tool [29], which consists of a context-free grammar (CFG) and an information extraction (IE) modules to transform free-text eligibility criteria to structured formats for downstream analysis. The CFG module uses a lexer to divide criteria into tokens and a modified Cocke-Younger-Kasami algorithm to build parse trees from tokens, which are subsequently analyzed by removing duplicates and subtrees. The IE module uses an attention-based bidirectional long short-term memory with a conditional random field layer for named entity recognition to extract MeSH terms from criteria text. Based on the evaluation in [29], its performance is competitive in entity recognition, entity linking, and attribute linking. As it only extracted MeSH concepts from eligibility criteria but not their temporal constraints and other qualifiers, we called them “eligibility features” in this paper. To evaluate its concept extraction accuracy, we manually reviewed a random sample of 300 rows of extracted results along with their original criteria. In the extracted results, there are cases where a term was identified but was not matched to a MeSH concept. In such cases, we considered them to be false negatives. The precision is 98.9%. The recall is 81.1%. The false negative ones were mostly quantitative criteria (29.3%) or due to missing concepts in MeSH (48.3%). We also identified the parsing errors of the frequent concepts and used Python to correct them in the whole parsing result. For example, we corrected the parsing results of the criterion “men”, which was parsed as “multiple endocrine neoplasia”. It is fine to miss some quantitative criteria as they were extracted by Valx [30] with a high sensitivity and specificity. We also merged similar concepts in the parsing results based on the analysis needs. Detailed information about the merging of extracted concepts can be found in the Supplementary Material II. To evaluate the accuracy of extracting known risk factors from eligibility criteria using the tool [29], we reviewed a random sample of 300 rows of extracted risk factors; the precision is 100%. Regarding the recall, we took a random sample of 200 unique criteria. 72 of them contain a risk factor and the program extracted 61 of them, making the recall to be 84.7%. After the categorical eligibility features of COVID-19 studies were parsed, we conducted three types of analyses: (1) frequency of the categorical eligibility features; (2) clustering analysis of the clinical studies based on the parsed eligibility features; and (3) frequency of exclusion eligibility features on chronic conditions and risk factors. Since (1) is intuitive, we explain the process of (2) and (3) in details as follows.

Clustering analysis of clinical studies: We used the clustering analysis to group the clinical studies based on their eligibility features. After the inclusion and exclusion criteria are parsed by the aforementioned tool [29], we utilized the parsed concepts as features to construct clinical study representation. For inclusion and exclusion eligibility features, we first removed the duplicated concepts for each clinical study. For example, if “pregnancy women" is mentioned multiple times in the exclusion criteria, only one was kept. Then, we append the prefixes ‘inc’ or ‘exc’ to the concepts extracted from inclusion or exclusion criteria respectively to differentiate them. After data preprocessing, we constructed the data representations by treating each clinical study as a text document that contains concepts from inclusion and exclusion eligibility features. The Term Frequency-Inverse Document Frequency (TF-IDF) weighting scheme was employed to construct the feature vectors to feed to the K-means clustering algorithm [31]. K-means is straightforward and can be applied to analyze large and high dimensional datasets. The algorithm assigns the instance to one of the clusters. The objective is to minimize the sum of the distances of the instances within the cluster to the cluster centroid. The silhouette value and CHindex were jointly used to measure the clustering results of K-means to determine the optimal number of clusters. The silhouette values measure similarity of an instance to its own cluster compared to other clusters. In this research, we experimented with k values from 2 to 50 for k-means. The optimal k was chosen when the silhouette value average of all instances is high and there are at least 20 instances for each cluster. To visualize the clustering result, we employed the uniform manifold approximation and projection (UMAP) [32] to project the high dimensional data into two-dimensional space for visualization. The UMAP reduces the dimensions by estimating the topology of the high dimensional data. It considers the local relationships within groups and global relationships between groups. UMAP can be applied directly to sparse matrices. In addition, we also clustered only the interventional studies considering both the extracted eligibility features and the enrollment values. The reason we included only interventional studies in this analysis is that observational studies often have a huge enrollment value, thereby dominating the clustering results. In addition, the eligibility criteria of observational studies are usually broad and unrestrictive. We used Principal Component Analysis (PCA) to reduce the dimensionality of eligibility features (weighted by TF-IDF) to k and then added two more features: the enrollment value (normalized to 0-1) and the intervention type, due to their importance deemed by the study team. The original dimension is 2608 presenting the total number of eligibility features. Since we add two study features, in order to avoid biases and preserve the data distribution on eligibility features, we experimented with k value from 10 to 15. There is no significant difference between the clustering results based on the optimal silhouette and the cluster distribution. Hence, k value was set to 10 in this study.

Exclusion eligibility features on chronic conditions and risk factors: First, we examined the upper limit and lower limit of the age criterion, which are structured data in the study summaries. Then, from the results of the criteria parsing tool [29], we examined the use of exclusion eligibility features for 15 most prevalent chronic conditions among older adults in the National Inpatient Sample of the Healthcare Cost and Utilization Project (HCUP) (appearing in over 6% of the older adults in NIS) [33]. These conditions include hypertension, hyperlipidemia, ischemic heart disease, diabetes, anemia, chronic kidney disease, atrial fibrillation, heart failure, chronic obstructive pulmonary disease and bronchiectasis, rheumatoid arthritis or osteoarthritis, acquired hypothyroidism, Alzheimer disease and related disorders or senile dementia, depression, osteoporosis, and asthma. In addition, we also considered three chronic conditions that are prevalent in all adults: cancer, stroke, and high cholesterol. We then analyzed the use of risk factors that may lead to serious illnesses in the eligibility criteria.

All the data and codes pertaining to this project have been deposited to GitHub: https://github.com/ctgatecci/Covid19-clinical-trials-11-27-2020.

Results

Basic Characteristics of the COVID-19 Clinical Studies

Table 2 shows the basic characteristics of 3,765 COVID-19 clinical studies in ClinicalTrials.gov. Among 3,765 clinical studies included in this paper, a majority of them are interventional studies (clinical trials). Among those interventional studies, 43.4% are in Phase 2 or 3. Most of the studies (83.19%) are sponsored by hospitals, universities, research institutes, or individuals. Besides drugs, other interventions include biological (15.9%), behaviors (7.06%), device (6.01%), diagnostic test (3.57%), and others (10.72%, e.g., genetic, dietary supplements, radiation, and combination). The majority of the studies focused on treatment (40.88%) and prevention (8.95%). The 10 most frequently tested drugs in different clinical studies are Hydroxychloroquine (N=162), Azithromycin (N=56), Tocilizumab (N=36), Ivermectin (N=31), Favipiravir (N=28), Remdesivir (N=28), Ritonavir (N=21), Lopinavir (N=20), Interferon (N=20), Plasma (N=19) (Figure 1).

Table 2. Basic characteristics of 3,765 COVID-19 clinical studies in ClinicalTrials.gov.

Characteristics	Number of studies	Percentage
Study type		
Interventional	2295	60.95%
Observational	1470	39.05%
Study Phase (Interventional study only)		
Phase 2	559	24.36%
Phase 3	350	15.25%
Phase 1	171	7.45%
Phase 2/Phase 3	156	6.80%
Phase 1/Phase 2	138	6.01%
Phase 4	108	4.71%
Early Phase 1	38	1.66%
N/A	775	33.77%
Gender		
Female only	69	1.83%
Male only	29	0.77%
Both	3677	97.40%
Overall status		
Recruiting	1910	50.73%

Not yet recruiting	851	22.60%
Completed	479	12.72%
Active, not recruiting	280	7.44%
Enrolling by invitation	121	3.21%
Withdrawn	64	1.70%
Terminated	34	0.90%
Suspended	26	0.69%
Sponsor		
Industry	578	15.35%
NIH	46	1.22%
U.S. Federal Agencies	9	0.24%
Other[1]	3132	83.19%
Intervention Type (Interventional Studies Only)		
Drug	1156	50.37%
Biological	365	15.90%
Other[2]	246	10.72%
Behavioral	162	7.06%
Device	138	6.01%
Diagnostic Test	82	3.57%
Dietary Supplement	62	2.70%
Procedure	44	1.92%
Combination Product	23	1.00%
Radiation	16	0.70%
Genetic	1	0.04%
Primary Purpose		
Treatment	1539	40.88%
Prevention	337	8.95%
Other	110	2.92%
Supportive Care	106	2.82%
Diagnostic	101	2.68%
Health Services Research	47	1.25%
Basic Science	26	0.69%
Screening	22	0.58%
Device Feasibility	7	0.19%
N/A	1470	39.04%
Allocation (Interventional Studies Only)		
Randomized	1688	73.55%
Non-Randomized	191	8.32%
N/A	416	18.13%
Intervention Model		
Parallel Assignment	1632	43.35%

Single Group Assignment	450	11.95%
Sequential Assignment	122	3.24%
Crossover Assignment	57	1.51%
Factorial Assignment	34	0.90%
N/A	1470	39.04%

[1]”Other” includes hospitals, universities, research institutes, and individuals
[2]”Other” includes dietary supplements, genetic, radiation, and combination product

[Figure 1. Number of interventional studies using a drug as an intervention. The denominator is the 1218 interventional studies using drug as an intervention. Note that some studies tested multiple drugs.]

Risk Factors in Trial Description

Figure 2 illustrates the occurrences of the risk factors in the study description of the included studies. We merged the brief summary and detailed description. “Weak immune 1” corresponds to immunocompromised state from solid organ transplant and “weak immune 2” corresponds to immunocompromised state from blood or bone marrow transplant, immune deficiencies, HIV, use of corticosteroids, or use of other immune-weakening medicines. The top 5 risk factors mentioned in trial description are diabetes, hypertension, weak immune 2, obesity, and pregnancy. According to the t-test result, on average, interventional studies mentioned fewer risk factors in trial description than observational studies (mean value: 0.2 vs 0.27, $P = 0.002$, two-tailed t-test). The number of risk factors mentioned in trial description was significantly associated with the intervention type ($p < 0.001$, ANOVA), while no statistically significant association between it and the primary purpose ($P = 0.078$, ANOVA).

[Figure 2. Number of studies with a risk factor for severe illness in the trial description. The denominator is the 3765 clinical studies included in this study.]

Quantitative criteria

Table 3 lists the top 20 frequently used quantitative criteria in COVID-19 clinical studies. Note that the "age" criterion is also a structured field in the study records. Based on the analysis of upper age limit, 67.3% (N= 2534) clinical studies do not have an upper age limit. For those that have an upper age limit, the most frequent limits are 80 (N=191), 75 (N=117), 65 (N=108), 100 (N=99), and 70 (N=94). Regarding the lower age limit, only 9.8% studies (N=369) do not have a lower age limit. Most frequently used lower age limits are 18 (N=2856), 16 (N=59), 20 (N=49), 19 (N=32), and 50 (N=31). Figure 3 illustrates the percentage of COVID-19 clinical studies that consider each age range. In general, patients who are over 18 years old are considered while those over 70 years old are less considered than 18-70 years old. Regarding oxygen saturation, most studies use 93% (N=114), 94% (N=58), or 90% (N=29) as threshold values.

Table 3. Top 20 frequently used quantitative criteria in COVID-19 clinical studies.

Rank	**Criteria**	**Frequency**	**Percentage**	**Rank**	**Criteria**	**Frequency**	**Percentage**
1	Age	2255	70.4%	11	Platelet count	108	3.4%
2	Oxygen saturation	229	7.2%	12	ANC	106	3.3%
3	Pao2/fio2	223	7.0%	13	Creatinine clearance	105	3.3%
4	BMI	222	6.9%	14	Heart rate	75	2.3%
5	Respiratory rate	190	6.0%	15	Diastolic blood pressure	64	2.0%
6	AST	185	5.8%	16	QTC	63	2.0%
7	EGFR	145	4.5%	17	Total bilirubin level	56	1.8%
8	Temperature	131	4.1%	18	Pulse rate	52	1.6%
9	Systolic blood pressure	113	3.5%	19	Hemoglobin	48	1.5%
10	ALT	109	3.4%	20	Creatinine	46	1.4%

[**Figure 3.** Percentage of COVID-19 clinical studies allowing age ranges]

Categorical Eligibility Features

Figure 4 illustrates frequent concepts extracted from inclusion and exclusion criteria of COVID-19 clinical studies. According to these results, COVID-19 diagnosis, polymerase chain reaction, pneumonia, diabetes, therapeutics, mechanical ventilation were often used eligibility features in the inclusion criteria, whereas pregnancy, therapeutics, kidney diseases, cancer, HIV, mechanical ventilation, hydroxychloroquine, hepatitis C were often used eligibility features in the exclusion criteria.

[**Figure 4**. Frequent eligibility features of COVID-19 clinical studies. The denominator is the 3765 clinical studies included in this study.]

Figure 5 shows the number of studies that used an exclusion eligibility feature about a common chronic condition prevalent among older adults in the included studies as well as the risk factors for serious illnesses in the eligibility criteria. Even though a majority of studies did not exclude patients with these chronic conditions, some highly prevalent chronic conditions such as cancer, heart failure, hypertension, diabetes, chronic kidney disease, and COPD are among the most frequently used exclusion criteria in 3.64% - 9.99% studies. Few studies purposely included patients with a risk factor that may lead to serious illnesses, but many studies explicitly excluded them, especially pregnant women. According to the results of the statistical tests, on average, interventional studies used more risk factors in eligibility criteria than observational studies (mean: 1.19 vs. 0.22, p<0.001, two-tailed t-test). There is a statistically significant association between the number of risk factors used in eligibility criteria and the intervention type (p < 0.001, ANOVA), and primary purpose of the studies (P< 0.001, ANOVA).

[**Figure 5.** (a) Number of studies using a prevalent chronic condition among the older adults in exclusion criteria. ** represents the conditions that are not in the list of top 15 prevalent conditions among older adults but prevalent in younger adults. (b) Number of studies with the risk factor in inclusion criteria (c) Number of studies with the risk factor in exclusion criteria. The denominator of these three figures is the 3765 clinical studies included in this study.]

Table 4 shows the top 10 frequent inclusion and exclusion features used in the studies in each of the 7 clusters resulting from the clustering analysis with eligibility features only. Pregnancy is the most frequent exclusion criterion in all the 7 clusters. Cluster #0 and #1 are the largest clusters with low silhouette value and the trials in these clusters often excluded patients with different diseases. Studies in Cluster #0 often included patients with pneumonia and excluded patients with cognition/cognitive behavioral therapy/cognitive dysfunction. Studies in Cluster #1 often excluded patients who are on therapeutics, kidney diseases, and cancer. Studies in Cluster #2 often included patients with polymerase chain reaction. All the trials in Cluster #3 excluded pregnant women. Studies in Cluster #5 often excluded HIV/HIV Infections, Hepatitis B, Hepatitis C, and Cancer. Cluster #6 has the highest silhouette value which indicates that the studies in it are more cohesive than those in other clusters. Most studies in Cluster #6 have no exclusion criteria. Through the clustering analysis, we can help classify studies on relationships not available a priori and identify sets of clinical studies focusing on different study population. Figure 6 shows the visualization of these 7 clusters using UMAP. The detailed results of the clustering analysis of the COVID-19 clinical studies are provided in the Supplementary Material III. The results of the clustering analysis of the interventional studies when considering the eligibility features, the enrollment, and the intervention type are provided in the Supplementary Material IV (table and figure) and V (detailed results).

Table 4. Top 10 frequently used concepts in inclusion criteria and exclusion criteria of the studies in each cluster of the clustering analysis with eligibility features.

Cluster Number	**Number of Studies**	**Total Enrollment**	**Silhouette scores**	**Inclusion Criteria**	**Exclusion Criteria**
0	124	38351	0.0112027	Pneumonia (N=90), COVID-19 (N=49), Women (N=21), Men (N=19),	Pregnancy (N=77), Cognition/Cognitive behavioral therapy/Cognitive dysfunction (N=35), Women (N=30),

					Therapeutics (N=19), HIV/HIV infection (N=16), Kidney diseases (N=15)
1	975	853187	-0.019703	COVID-19 (N=199), Women (N=101), Men (N=86),	Pregnancy (N=427), Women (N=160), Therapeutics (N=149), Kidney diseases (N=134), Cancer (N=100), Hypertension (N=79), Liver diseases (N=79)
2	216	71812	0.0103414	Polymerase chain reaction (N=52), COVID-19 (N=37), Diabetes (N=36), Hypertension (N=22), Obesity (N=19)	Pregnancy (N=198), Women (N=124), Therapeutics (N=23), Pregnancy tests (N=20), Hydroxychloroquine (N=19),
3	73	16039	0.0848274	COVID-19 (N=18), Women (N=7), Men (N=6),	Pregnant women (N=73), Women (N=15), Cancer (N=10), Asthma (N=7), Kidney diseases (N=6), Diabetes (N=6), Heart failure (N=6)
4	252	107769	0.0076462	Men (N=232), Women (N=219), COVID-19 (N=72), Pregnancy tests (N=37),	Pregnancy (N=165), Women (N=90), Therapeutics (N=61), Dialysis (N=38), Ventilation mechanical (N=36), COVID-19 (N=30)
5	297	502661	0.0052849	Men (N=128), Women (N=121)	Pregnancy (N=221), HIV/HIV infections (N=159). Women (N=125), Hepatitis B (N=109), Hepatitis C (N=104), Cancer (N=101), Therapeutics (N=97), COVID-19 (N=81)
6	105	103814	0.2595492	COVID-19 (N=96), Women (N=7), Men (N=5)	Pregnancy (N=33), COVID-19 (N=30), Women (N=14), Therapeutics (N=7), Cancer (N=5), Kidney diseases (N=4), Therapies investigational (N=4)

[**Figure 6**. Visualization of the 7 clusters using UMAP]

Discussion

As the novel coronavirus COVID-19 has significantly impacted our lives and even taken lives of almost 3 millions of people so far, we must quickly identify repurposed drugs or develop new drugs and vaccines to safely and effectively control the spread of the virus and save lives. Clinical studies, especially randomized controlled trials, are a fundamental tool used to evaluate the efficacy and safety of new medical interventions for disease prevention or treatment. Many clinical studies are being conducted to find safe and effective treatments and vaccines. Thus far, significant efforts have been devoted to repurposing existing FDA-approved drugs including immunosuppression (e.g., Hydroxychloroquine, Tocilizumab), anti-virus (e.g., Fevipiravir, Lopinavir/Ritonavir), anti-parasite (e.g., Ivermectin, Nitazoxanide), antibiotics (e.g., Azithromycin), and anticoagulant (e.g., Enoxaparin). In our analysis of the COVID-19 clinical studies, we found that the use of eligibility criteria and consideration of risk factors in these studies did not change much from June 18, 2020 to November 27, 2020 even though the number of COVID-19 studies in ClinicalTrials.gov grew from 2,192 to 4,028.

To transform clinical trials and lower their cost, a notion of "digital clinical trial" was created to leverage digital technology to improve important aspects such as patient access, engagement, and trial measurement [34]. The US National Institutes of Health and the National Science Foundation held a workshop in April 2019 about the implementation of digital technologies in clinical trials, in which "defining and outlining the composition and elements of digital trials" and "elucidating digital analytics and data science approaches" were identified as two of the five top priorities. As COVID-19 is a major health crisis that impacts people regardless of their age, gender, and race/ethnicity, it is in our interest to understand if clinical studies on COVID-19 adequately considered the representation of real-world populations. Based on our analysis, most clinical studies consider both genders (97.4%, N=3,667), do not have an upper age limit 67.3% (N= 2,534), and have a lower age limit of 18 (75.9%, N=2,856). The exclusion of children in these studies may be due to lower susceptibility and lower rates of mortality and hospitalization for children with COVID-19 compared to adults [35]. As serious illnesses of COVID-19 mostly occurred in older adults with underlying health conditions, it is not surprising that they are in general considered by most COVID-19 studies, based on our analysis of their eligibility criteria. Most studies did not set an upper age limit (67.3%, N= 2534) and did not exclude older adults with common chronic conditions. This is contrary to the recent New York Times articles conjecturing that older people are left out form COVID-19 trials [10]. As older adults are the most likely to be hospitalized due to COVID-19, clinicians may be more likely to choose to include them to fulfill the sample size requirement of the trials. Nonetheless, conducting COVID-19 clinical studies could still be challenging in the traditional clinical trial eco-system, where patient accrual is often delayed due to logistical constraints [36]. The generalizability of the study results to the real-world population should be evaluated with state-of-the-art techniques [9]. Older adults could have still been underrepresented in COVID-19 clinical studies due to logistical reasons, which can only be assessed with the published results after the completion of the studies [37]. In addition, pregnant women are often excluded in COVID-19 studies. Even though pregnant women are in general excluded in most clinical trials due to the potential risks to both the women and the unborn babies, observational studies should carefully evaluate the vertical transmission of the virus and negative impact of COVID-19 on the well-being of mothers and infants [38]. Recently, Director of the National Institute of Child Health and Human Development published a viewpoint article in JAMA to call for greater inclusion of pregnant and lactating women in COVID-19 vaccine clinical research [39]. Clinical studies should adequately evaluate the efficacy and safety of treatments and vaccines on vulnerable population groups.

Limitations

A few limitations should be noted. First, some data in ClinicalTrials.gov are missing. For example, 33.8% (N=775) of the interventional studies miss study phase information. 39% (N=1,470) of studies do not have primary purpose information. Second, we relied on the search function of ClinicalTrials.gov when retrieving COVID-19 studies. There may be study indexing errors, but the scale should be minimal and would not impact the findings. Third, we used the QuickUMLS and the new eligibility criteria parsing tool [29] to extract risk factors, chronic conditions, disorders, and procedures from study records. Thus, the sensitivity and specificity of the term extraction and normalization are dependent on the quality of the UMLS Metathesaurus and the eligibility criteria parsing tool. Nonetheless, we have carefully curated the term extraction results to ensure that our results are as accurate as possible.

Conclusions and Future Work

In this paper, we systematically analyzed COVID-19 clinical study summaries in ClinicalTrials.gov using natural language processing. Specifically, we analyzed whether these clinical studies considered the underlying health conditions (and other risk factors) that may increase the severity of the COVID-19 illness. Given the ongoing nature of this pandemic, it is inevitable that early trials will start with different knowledge of risk factors than later trials. In future work, we will perform a longitudinal analysis of COVID-19 studies to assess the changes in the use of eligibility criteria and consideration of risk factors for severe illness in COVID-19 patients. As results of COVID-19 studies become available, we will be able to assess the extent to which the trial design and eligibility criteria in particular would impact the findings as well as the real-world population representativeness of these studies using generalizability assessment methods [9].

Acknowledgments

We would like to sincerely thank Markku Salkola for his generous help with the eligibility criteria parsing tool.

Data Availability

The data underlying this article were accessed from ClinicalTrials.gov. The derived data and code generated in this research are publicly available at https://github.com/ctgatecci/Covid19-clinical-trials-11-27-2020.

Funding

This study was partially supported by the National Institute on Aging (NIA) of the National Institutes of Health (NIH) under Award Number R21AG061431; and in part by Florida State University-University of Florida Clinical and Translational Science Award funded by National Center for Advancing Translational Sciences under Award Number UL1TR001427. The development of Valx was supported by National Library of Medicine Grant R01LM009886. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Author Contribution

ZH conceived, designed, guided, and coordinated the study and the writing. ZH collected the data from ClinicalTrials.gov, performed the data analyses, interpreted the results, and drafted the manuscript. AE performed the natural language processing of the clinical study records. XL performed the clustering analysis. AX performed statistical tests to assess the association between the occurrences of risk factors and the study characteristics. All the authors edited the manuscript thoroughly. The submitted manuscript has been approved by all the authors.

Conflict of Interest

None

References

[1] Koopmans M. The Novel Coronavirus Outbreak: What We Know and What We Don't. Cell. 2020;180.

[2] COVID-19 Map - Johns Hopkins Coronavirus Resource Center 2020. Available from: https://coronavirus.jhu.edu/map.html.

[3] Collins FS, Stoffels P. Accelerating COVID-19 Therapeutic Interventions and Vaccines (ACTIV): An Unprecedented Partnership for Unprecedented Times. JAMA. 2020.

[4] NIH-Wide Strategic Plan for COVID-19 Research 2020. Available from: https://www.nih.gov/sites/default/files/research-training/initiatives/covid-19-strategic-plan/coronavirus-strategic-plan-20200713.pdf.

[5] Ledford H, Cyranoski D, Van Noorden R. The UK has approved a COVID vaccine — here's what scientists now want to know 2020 [December 14, 2020]. Available from: https://www.nature.com/articles/d41586-020-03441-8.

[6] BBC. Covid: Canada latest country to approve Pfizer-BioNTech vaccine 2020 [December 14, 2020]. Available from: https://www.bbc.com/news/world-us-canada-55251830.

[7] Flaherty A, Salzman S, Strauss EM. FDA authorizes 1st COVID-19 vaccine in United States 2020 [December 14, 2020]. Available from: https://abcnews.go.com/Health/fda-authorizes-1st-covid-19-vaccine-united-states/story?id=74665712.

[8] Gilead suspension of China Covid-19 trials should serve as bellwether 2020 [07/14/2020]. Available from: https://www.clinicaltrialsarena.com/comment/gilead-remdesivir-covid-19-china-trials.

[9] He Z, Tang X, Yang X, Guo Y, George TJ, Charness N, Quan Hem KB, Hogan W, Bian J. Clinical Trial Generalizability Assessment in the Big Data Era: A Review. Clinical and Translational Science. 2020.

[10] Span P. Older Adults May Be Left Out of Some Covid-19 Trials. The New York Times. 2020.

[11] Wang Y, Zhou Q, Xu M, Kang J, Chen Y. Characteristics of Clinical Trials relating to COVID-19

registered at ClinicalTrials.gov. J Clin Pharm Ther. 2020.

[12] Pundi K, Perino AC, Harrington RA, Krumholz HM, Turakhia MP. Characteristics and Strength of Evidence of COVID-19 Studies Registered on ClinicalTrials.gov. JAMA Intern Med. 2020.

[13] Gianola S, Jesus TS, Bargeri S, Castellini G. Characteristics of academic publications, preprints, and registered clinical trials on the COVID-19 pandemic. PLoS One. 2020;15(10):e0240123.

[14] Ma LL, Yin X, Li BH, Yang JY, Jin YH, Huang D, Deng T, Wang YY, Ren XQ, Ji J, Zeng XT. Coronavirus Disease 2019 Related Clinical Studies: A Cross-Sectional Analysis. Front Pharmacol. 2020;11:540187.

[15] Jones CW, Woodford AL, Platts-Mills TF. Characteristics of COVID-19 clinical trials registered with ClinicalTrials.gov: cross-sectional analysis. BMJ Open. 2020;10(9):e041276.

[16] Zhu RF, Gao YL, Robert SH, Gao JP, Yang SG, Zhu CT. Systematic review of the registered clinical trials for coronavirus disease 2019 (COVID-19). J Transl Med. 2020;18(1):274.

[17] Nasrallah AA, Farran SH, Nasrallah ZA, Chahrour MA, Salhab HA, Fares MY, Khachfe HH, Akl EA. A large number of COVID-19 interventional clinical trials were registered soon after the pandemic onset: a descriptive analysis. J Clin Epidemiol. 2020;125:170-8.

[18] Rabby MII, Hossain F. Study of ongoing registered clinical trials on COVID-19: a narrative review. Sao Paulo Med J. 2020.

[19] Alag S. Analysis of COVID-19 clinical trials: A data-driven, ontology-based, and natural language processing approach. PLoS One. 2020;15(9):e0239694.

[20] Fragkou PC, Belhadi D, Peiffer-Smadja N, Moschopoulos CD, Lescure FX, Janocha H, Karofylakis E, Yazdanpanah Y, Mentre F, Skevaki C, Laouenan C, Tsiodras S, Viruses ESGfR. Review of trials currently testing treatment and prevention of COVID-19. Clin Microbiol Infect. 2020;26(8):988-98.

[21] Lu L, Li F, Wen H, Ge S, Zeng J, Luo W, Wang L, Tang C, Xu N. An evidence mapping and analysis of registered COVID-19 clinical trials in China. BMC Med. 2020;18(1):167.

[22] Kim JH, Ta CN, Liu C, Sung C, Butler AM, Stewart LA, Ena L, Rogers JR, Lee J, Ostropolets A, Ryan PB, Liu H, Lee SM, Elkind MSV, Weng C. Towards clinical data-driven eligibility criteria optimization for interventional COVID-19 clinical trials. J Am Med Inform Assoc. 2020.

[23] Zheng Z, Peng F, Xu B, Zhao J, Liu H, Peng J, Li Q, Jiang C, Zhou Y, Liu S, Ye C, Zhang P, Xing Y, Guo H, Tang W. Risk factors of critical & mortal COVID-19 cases: A systematic literature review and meta-analysis. J Infect. 2020;81(2):e16-e25.

[24] ClinicalTrials.gov. History, Policies, and Laws - ClinicalTrials.gov 2020 [7/10/2020]. Available from: https://clinicaltrials.gov/ct2/about-site/historyNPRM.

[25] Schwartz LM, Woloshin S, Zheng E, Tse T, Zarin DA. ClinicalTrials. gov and Drugs@ FDA: a comparison of results reporting for new drug approval trials. Annals of internal medicine. 2016;165(6):421-30.

[26] CDC. Evidence used to update the list of underlying medical conditions that increase a person's risk of severe illness from COVID-19 2020 [7/14/2020]. Available from: https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/evidence-table.html.

[27] Clinic M. COVID-19: Who's at higher risk of serious symptoms? 2020 [07/14/2020].

[28] Soldaini L, Goharian N, editors. QuickUMLS: a fast, unsupervised approach for medical concept extraction. MedIR workshop, sigir; 2016.

[29] Tseo Y, Salkola M, Mohamed A, Kumar A, Abnousi F. Information Extraction of Clinical Trial Eligibility Criteria. arXiv preprint arXiv:200607296. 2020.

[30] Hao T, Liu H, Weng C. Valx: a system for extracting and structuring numeric lab test comparison statements from text. Methods of information in medicine. 2016;55(3):266.

[31] Hartigan JA, Wong MA. Algorithm AS 136: A k-means clustering algorithm. Journal of the royal statistical society series c (applied statistics). 1979;28(1):100-8.

[32] McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:180203426. 2018.

[33] He Z, Bian J, Carretta HJ, Lee J, Hogan WR, Shenkman E, Charness N. Prevalence of Multiple Chronic Conditions Among Older Adults in Florida and the United States: Comparative Analysis of

the OneFlorida Data Trust and National Inpatient Sample. J Med Internet Res. 2018;20(4):e137.

[34] Inan O, Tenaerts P, Prindiville S, Reynolds H, Dizon D, Cooper-Arnold K, Turakhia M, Pletcher M, Preston K, Krumholz H. Digitizing clinical trials. npj Digital Medicine. 2020;3(1):1-7.

[35] Nicholas GD, Petra K, Yang L, Kiesha P, Mark J, Rosalind M, group CC-w. Age-dependent Effects in the Transmission and Control of COVID-19 Epidemics. Nature medicine.

[36] Howard SC, Algra A, Warlow CP, Rothwell PM. Potential consequences for recruitment, power, and external validity of requirements for additional risk factors for eligibility in randomized controlled trials in secondary prevention of stroke. Stroke. 2006;37(1):209-15.

[37] He Z, Gonzalez-Izquierdo A, Denaxas S, Sura A, Guo Y, Hogan WR, Shenkman E, Bian J. Comparing and Contrasting A Priori and A Posteriori Generalizability Assessment of Clinical Trials on Type 2 Diabetes Mellitus. AMIA Annu Symp Proc. 2017;2017:849-58.

[38] Liu H, Wang L-L, Zhao S-J, Kwak-Kim J, Mor G, Liao A-H. Why are pregnant women susceptible to viral infection: an immunological viewpoint? Journal of reproductive immunology. 2020:103122.

[39] Bianchi DW, Kaeser L, Cernich AN. Involving Pregnant Individuals in Clinical Research on COVID-19 Vaccines. JAMA. 2021.

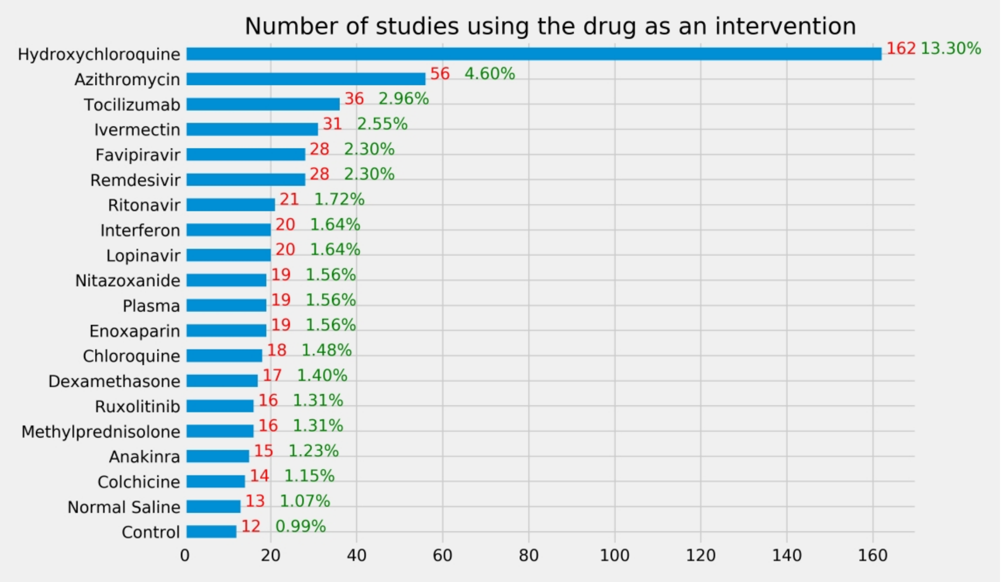

Figure 1. Number of interventional studies using a drug as an intervention. The denominator is the 1218 interventional studies using drug as an intervention. Note that some studies tested multiple drugs.

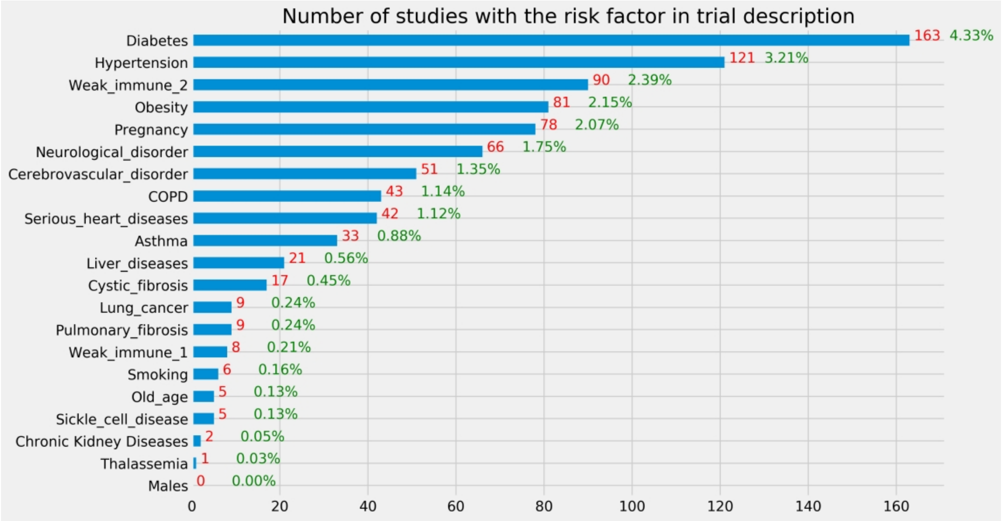

Figure 2. Number of studies with a risk factor for severe illness in the trial description. The denominator is the 3765 clinical studies included in this study.]

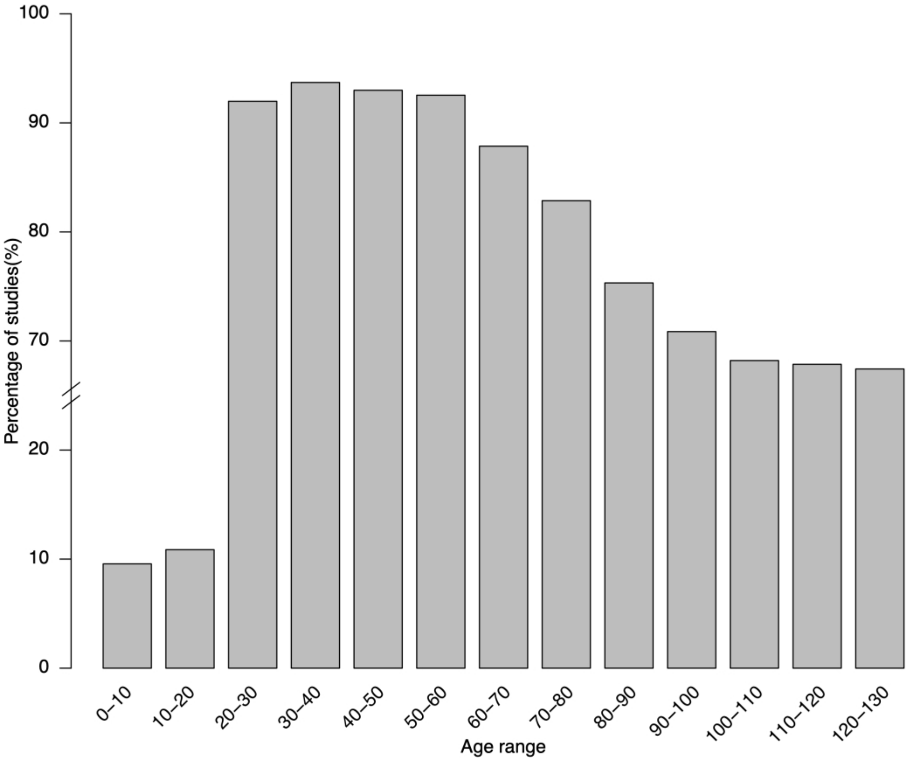

Figure 3. Percentage of COVID-19 clinical studies allowing age ranges

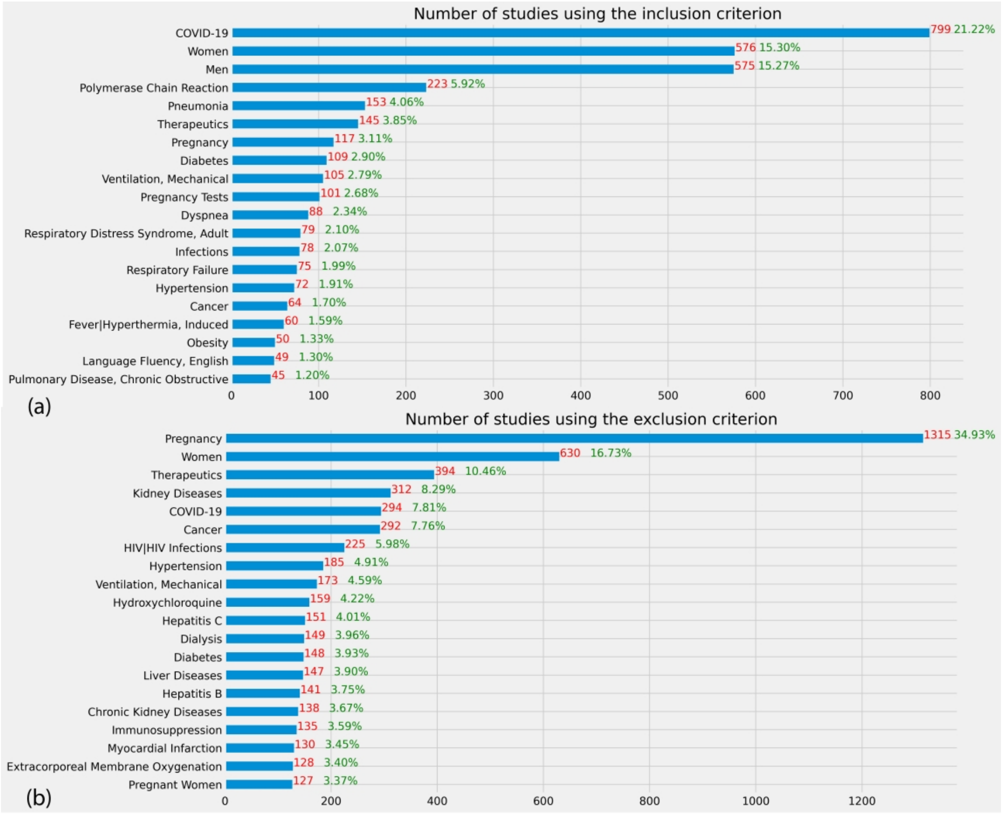

Figure 4. Frequent eligibility features of COVID-19 clinical studies. The denominator is the 3765 clinical studies included in this study.

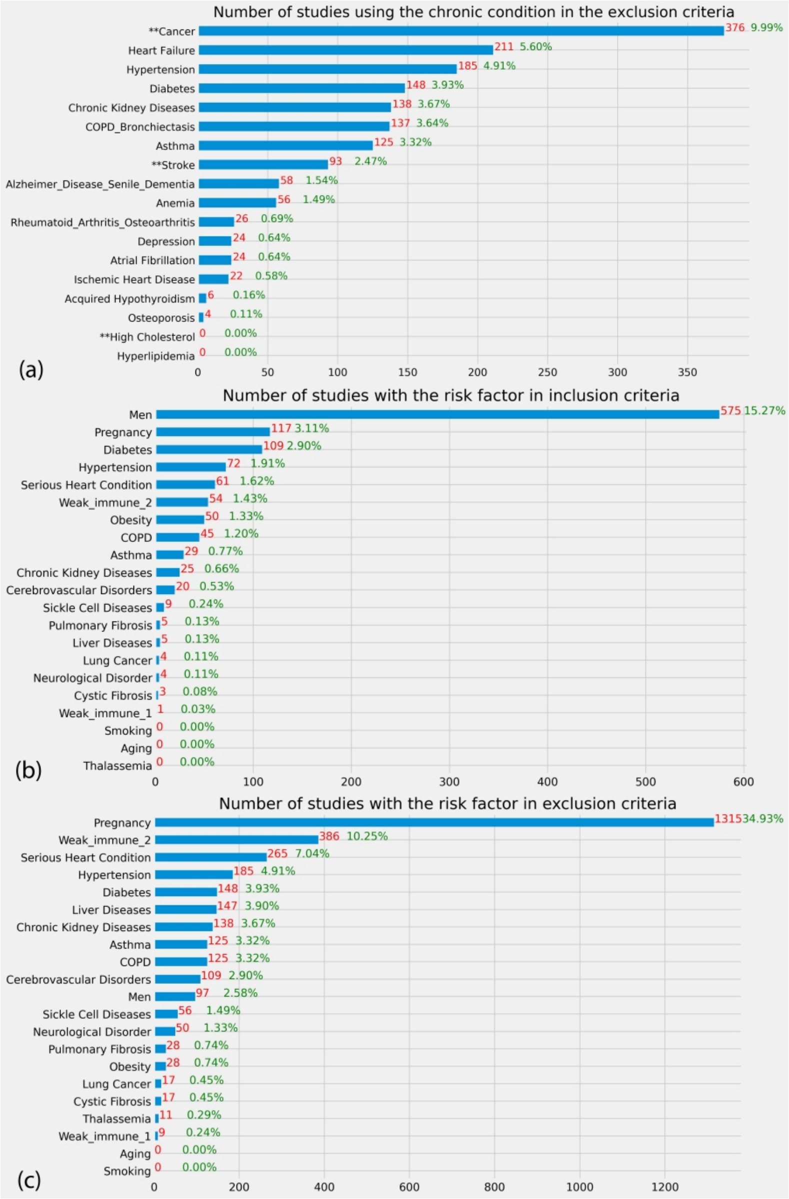

Figure 5. (a) Number of studies using a prevalent chronic condition among the older adults in exclusion criteria. ** represents the conditions that are not in the list of top 15 prevalent conditions among older adults but prevalent in younger adults. (b) Number of studies with the risk factor in inclusion criteria (c) Number of studies with the risk factor in exclusion criteria. The denominator of these three figures is the 3765 clinical studies included in this study.]

1411x2144mm (72 x 72 DPI)

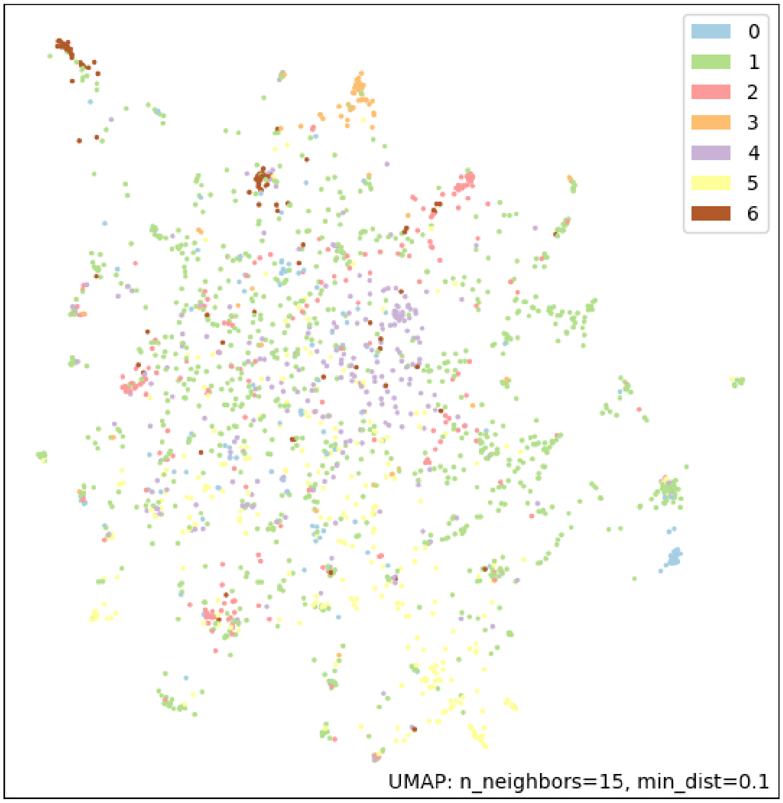

Figure 6. Visualization of the 7 clusters using UMAP