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], Fnu 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: 3998

Abstract

Objective: The novel coronavirus disease (COVID-19), broke out in December 2019, is a global pandemic. Rapidly in the past few months, a large number of clinical studies have been initiated worldwide to find effective therapeutics, vaccines, and preventive strategies. In this study, we aim to understand the landscape of COVID-19 clinical research and identify the gaps and issues that may cause difficulty in recruitment and the lack of population representativeness.

Materials and Methods: We analyzed 2,034 COVID-19 studies registered in the largest public registry - ClinicalTrials.gov. Leveraging natural language processing, descriptive analysis, association analysis, and clustering analysis, we characterized COVID-19 clinical studies by phase and design features. Particularly, we analyzed their eligibility criteria to understand: (1) whether they considered the reported underlying health conditions that may lead to severe illnesses, and (2) if these studies excluded older adults, either explicitly or implicitly, which may reduce the generalizability of these studies in older adults.

Results: The 5 most frequently tested drugs are Hydroxychloroquine (N=148), Azithromycin (N=46), Tocilizumab (N=29), Lopinavir (N=20), and Ritonavir (N=20). Most trials did not have an upper age limit and did not exclude patients with common chronic conditions such as hypertension and diabetes that are prevalent in older adults. However, known risk factors that may lead to severe illnesses have not been adequately considered by existing studies.

Conclusions: A careful examination of the registered COVID-19 clinical studies can identify the research gaps and inform future COVID-19 trial design towards balanced internal validity and generalizability.

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

Introduction

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and the associated coronavirus disease (COVID-19) broke out in Wuhan, China in December 2019 and has quickly become a global pandemic with serious health and social consequences [1]. As of September 3, 2020, more than 26 million confirmed cases have been reported around the world and about one-fourth are from the U.S. [2]. Globally, more than 860,000 people have died due to COVID-19 and 186,000 in the U.S. alone. Since the novel coronavirus has only been discovered in the past few months, little is known about the mechanisms underlying the infection and progression of the disease. Neither an effective treatment nor a vaccine is yet available for COVID-19. In April 2020, the National Institutes of Health (NIH) launched the Accelerating COVID-19 Therapeutic Interventions and Vaccines (ACTIV) public-private partnership to develop a coordinated research strategy for prioritizing and speeding 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]. Typically, it may take many 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. So far, researchers are creating complex computer-generated models of SARS-Cov-2 and its biological processes to determine key interactions and pathways to target therapeutic development or developing 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].

Clinical studies, especially randomized controlled trials, are the gold standard for evaluating the efficacy and safety of a treatment. Regardless of the techniques used for drug discovery (e.g.

in vivo, in silico, in vitro), the therapeutics and vaccines have to go through three phases of clinical trials to evaluate their efficacy and safety before approvals of FDA can be granted for mass production and use in the population. Rapidly 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 [5]. Other issues such as population representation 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 [6]. A recent New York Times article conjectured that older adults are left out form COVID-19 trials [7]. 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, there is not a systematic analysis of COVID-19 clinical studies. Through analyzing the study phases, status, sponsors, types, interventions, purposes, and participant allocation of the clinical studies on COVID-19, we can gain a better understanding of the landscape of COVID-19 research. A number of other research questions are also worthy of investigation. For example, what eligibility criteria are used in COVID-19 clinical studies? Are these criteria too restrictive? 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) [8]. Have existing clinical studies sufficiently considered these known risk factors? 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, which may be the reason for their underrepresentation.

In this study, we conducted a systematic analysis of the registered clinical studies on COVID-19 from ClinicalTrials.gov (as of June 18, 2020) 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 [9]. 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 [10].

Dataset Processing

From ClinicalTrials.gov, we downloaded XML-format study records of 2,192 clinical studies that

are tagged with a condition "COVID-19" or "SARS-CoV-2" on 6/18/2020. We excluded 19 studies tagged with the study type "Expanded Access" and 139 studies that were tagged as patient registries, leaving 2,034 records that met our inclusion criteria. 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) [11] and Mayo Clinic [12] (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) in the QuickUMLS [13] 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 illness of COVID-19 patients and their associated UMLS CUIs. This list was used as a dictionary in QuickUMLS [13] to identify risk factors from trial description of the study records. 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 [14], which will be explained in the following subsection. 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.

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
Immunocompromised state (weakened immune system) from solid organ transplant	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, C2984299
Neurologic 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
Immunocompromised state (weakened immune system) from blood or bone marrow transplant, immune deficiencies, HIV, use of corticosteroids, or use of other immune weakening medicines	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 Valx tool [15] 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. We then analyzed the frequency of the quantitative criteria and the threshold values used for patient eligibility determination.

Qualitative criteria: To extract the qualitative criteria from the eligibility criteria from COVID-19 studies, we used a recently published eligibility criteria parsing tool (presented in 2020 KDD Workshop on Applied Data Science for Healthcare) [14]. This new open-source tool consists of context-free grammar (CFG) and information extraction (IE) modules to transform free-text eligibility criteria to structured relations. 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 [14], its performance is competitive. As MeSH (27,000 concepts) is much smaller than ICD-9-CM (70,000

concepts) and SNOMED CT (350,000 concepts), it captures the most important concepts related to treatment and disease, which can adequately meet the needs of this work. In addition, as current version of MeSH has not added COVID-19 related concepts, it supplemented MeSH with a customized dictionary built specifically for COVID-19 clinical studies. We therefore adopted this new tool to parse eligibility criteria in this study. To evaluate its concept extraction accuracy, ZH manually reviewed a random sample of 300 rows of extracted results along with their original criteria. 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 manually corrected the parsing errors of the frequent concepts. 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 [15] 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 I. After the qualitative criteria of COVID-19 studies were parsed, we conducted three types of analyses: (1) frequency of the qualitative criteria; (2) clustering analysis of the clinical studies based on the parsed criteria; and (3) frequency of exclusion criteria 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 criteria. After the inclusion and exclusion criteria are parsed by the aforementioned tool [14], we utilized the parsed concepts as features to construct clinical study representation. For inclusion and exclusion criteria, 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 criteria. 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 [16]. K-means is rather easy to implement and apply on large and high dimensional data sets. 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 was used to measure the clustering results of K-means to determine the optimal number of clusters. The silhouette values measures 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 cluster distributions on a two-dimensional space, the t-Distributed Stochastic Neighbor Embedding (t-SNE) [17] was employed to project the high dimensional data into two-dimensional space. The t-SNE algorithm minimizes the sum of the KL divergences of all data points in the original dimensional space and the mapping space. The computational cost of t-SNE is high when the original dimensionality of the data is high. To speed up the process, Principle Component Analysis (PCA) was used to reduce the dimensionality before t-SNE technique was applied.

Exclusion criteria on chronic conditions and risk factors: First, we examined the upper limit and lower limit of the age eligibility criterion, which are structured data in the study summaries. Then, from the results of the criteria parsing tool [14], we examined the use of exclusion criteria of 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) [18]. 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 younger 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.

Results

Basic Characteristics of the COVID-19 Clinical Studies

Table 2 shows the basic characteristics of 2,034 COVID-19 clinical studies in ClinicalTrials.gov. Among 2,034 clinical studies included in this paper, a majority of them are interventional studies (clinical trials). Among those interventional studies, 24.2% are in Phase 2/3 or 3. Most of the studies (86.7%) are sponsored by hospitals, universities, research institutes, or individuals. Besides drugs, other interventions include biological (13.4%), device (6.3%), behaviors (5.4%), diagnostic test (3.4%), and others (14.1%, e.g., genetic, dietary supplements, radiation, and combination). The majority of the studies focused on treatment (43.4%) and prevention (7.7%). The 10 most frequently tested drugs in different clinical studies are Hydroxychloroquine (N=148), Azithromycin (N=46), Tocilizumab (N=29), Lopinavir (N=20), Ritonavir (N=20), Chloroquine (N=16), Favipiravir (N=16), Ivermectin (N=15), Ruxolitinib (N=14), and Remdesivir (N=14) (Figure 1).

Table 2. Basic characteristics of 2,034 COVID-19 clinical studies in ClinicalTrials.gov.

Characteristics	Number of studies	Percentage	Characteristics	Number of studies	Percentage
Study Type			**Sponsor**		
Interventional	1,245	61.2%	Industry	239	11.8%
Observational	789	38.8%	NIH	27	1.3%
Study Phase (Interventional Studies Only)			U.S. Federal Agencies	4	0.2%
Phase 1	82	6.6%	Other[1]	1,764	86.7%
Phase 1/Phase 2	62	5.0%	**Intervention Type (Interventional Studies Only)**		
Phase 2	336	27.0%	Drug	721	57.9%
Phase 2/Phase 3	89	7.2%	Procedure	36	2.9%
Phase 3	212	17.0%	Behavioral	70	5.6%
Phase 4	68	5.5%	Biological	173	13.9%
N/A	396	31.8%	Device	87	7.0%
Gender			Diagnostic test	45	3.6%
Female only	42	2.1%	Other[2]	350	28.1%
Male only	9	0.4%	**Primary Purpose**		
Both	1983	97.5%	Treatment	883	43.4%
Overall Status			Prevention	156	7.7%
Active, not recruiting	101	5.0%	Diagnostics	52	2.6%
Completed	134	6.6%	Supportive care	55	2.7%
Enrolling by invitation	55	2.7%	Other	99	4.9%
Not yet recruiting	687	33.8%	N/A	947	46.6%
Recruiting	1026	50.4%	**Allocation (Interventional Studies Only)**		
Suspended	12	0.6%	Randomized	900	72.3%
Terminated	4	0.2%	Non-randomized	109	8.8%
Withdrawn	15	0.7%	N/A	236	18.9%

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

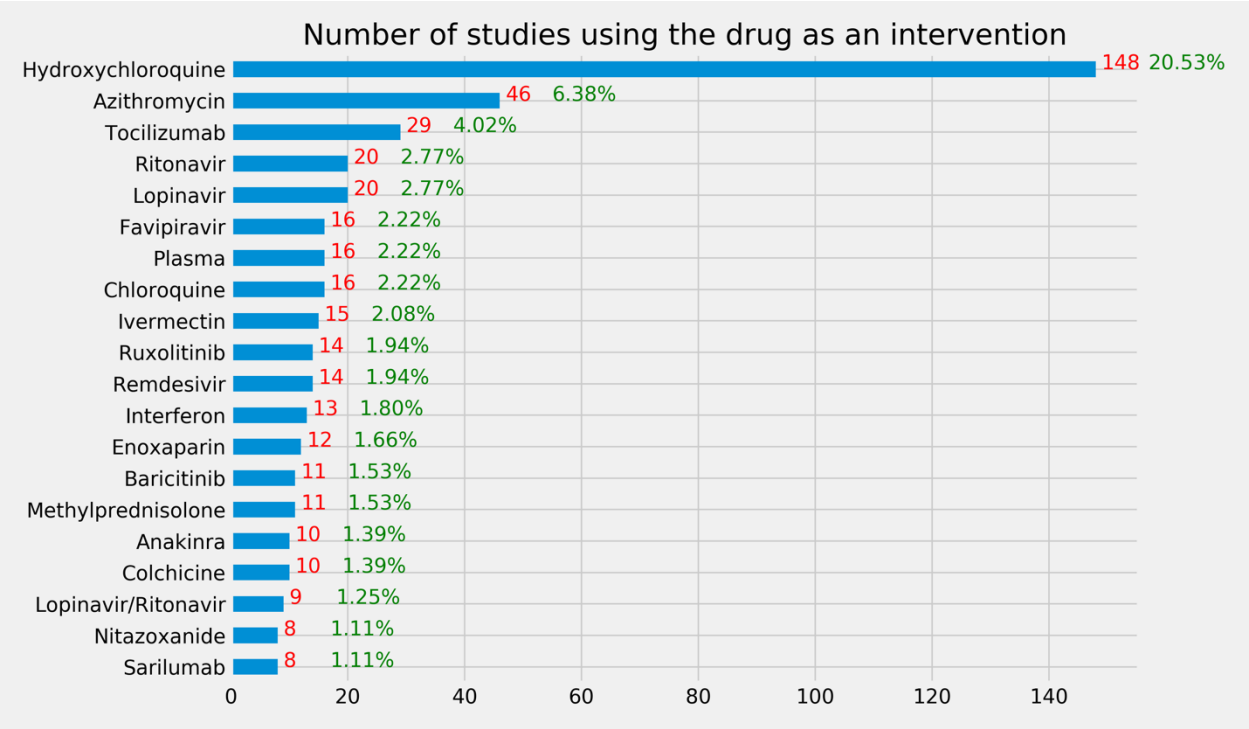

Figure 1. Number of interventional studies using a drug as an intervention. The denominator is the 721 interventional studies using drug as an intervention. Note that some studies tested multiple drugs.

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 old age, diabetes, hypertension, weakened immune system due to reasons other than solid organ transplant, and obesity. According to the t-test result, on average, interventional studies mentioned fewer risk factors in trial description than observational studies (mean value: 1.5 vs 1.8, $P = 0.006$, two-tailed t-test). There is no statistically significant association between the number of risk factor mentioned in trial description with the intervention type (P =

0.59, ANOVA) and primary purpose (P = 0.94, ANOVA).

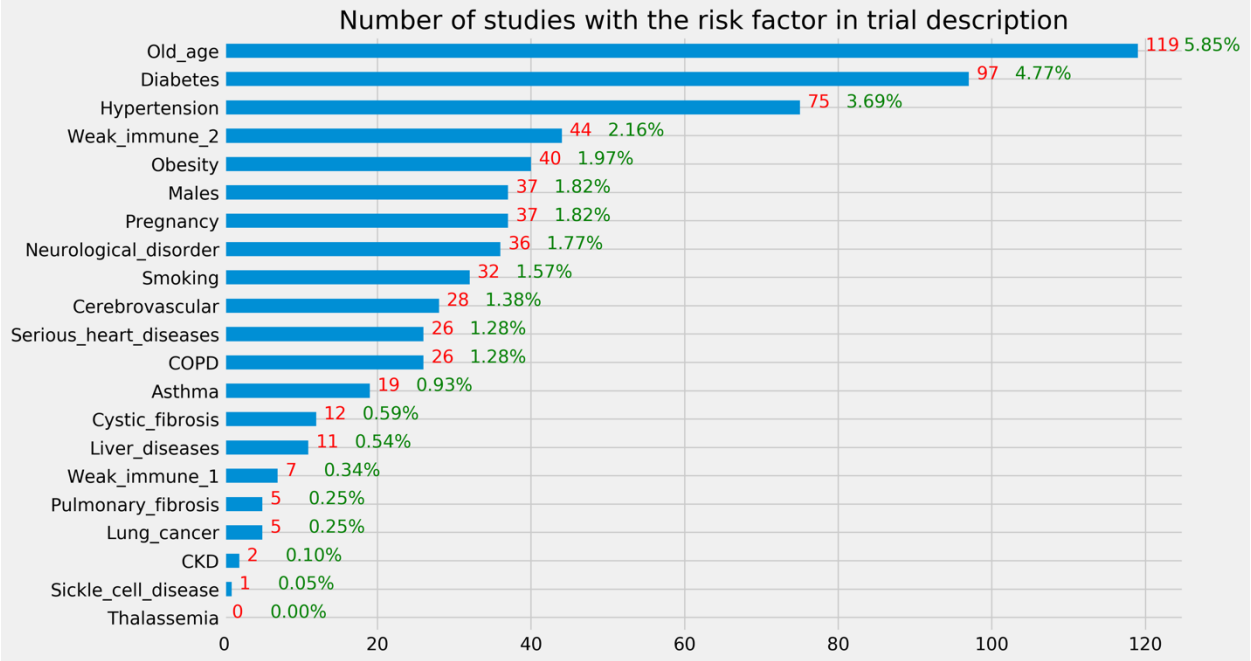

Figure 2. Number of studies with a risk factor for severe illness in the trial description. The denominator is the 2034 clinical studies included in this study.

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, 75.5% (N=1536) clinical studies do not have an upper age limit. For those that have an upper age limit, the most frequent limits are 80 (N=106), 75 (N=69), 100 (N=58), 65 (N=53), and 70 (N=49). Regarding the lower age limit, only 11.9% studies (N=241) do not have a lower age limit. Most frequently used lower age limits are 18 (N=1667), 16 (N=34), 20 (N=21), 12 (N=20), and 60 (N=17). 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 92% (18/185), 93% (75/185), or 94% (40/185) as threshold values.

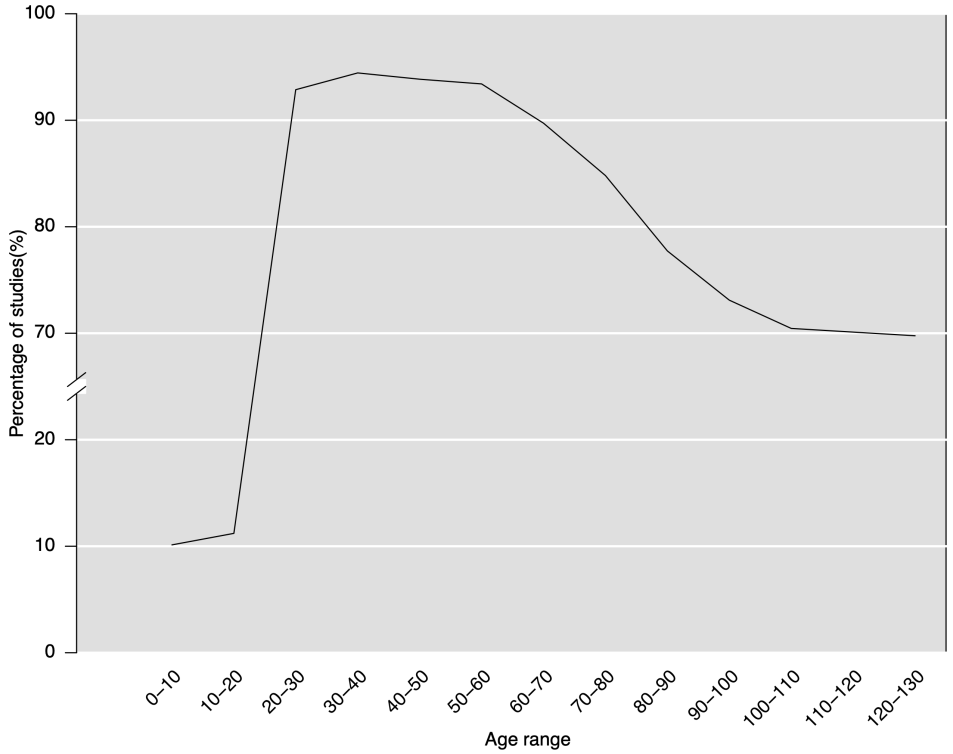

Figure 3. Percentage of COVID-19 clinical studies over permissible age ranges

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

Rank	Criteria	Frequency	Percentage	Rank	Criteria	Frequency	Percentage
1	Age	1,230	60.5%	11	ANC	62	3.1%
2	Oxygen saturation	185	9.1%	12	QTC interval	58	2.9%
3	Pao2/fio2	141	6.9%	13	Platelet count	61	3.0%
4	Respiratory rate	109	5.4%	14	Systolic blood pressure	57	2.8%
5	BMI	93	4.6%	15	Weight	51	2.5%
6	AST	93	4.6%	16	Diastolic blood pressure	37	1.8%
7	Creatinine clearance	75	3.7%	17	Heart rate	36	1.8%
8	EGFR	74	3.7%	18	Hemoglobin	26	1.3%
9	Temperature	66	3.2%	19	Creatinine	26	1.3%
10	ALT	64	3.2%	20	D-dimer	25	1.2%

Qualitative Eligibility Criteria

Figure 4 illustrates frequent concepts extracted from inclusion and exclusion criteria of COVID-19 clinical studies. According to this results, COVID-19 studies often included patients with COVID-19 diagnosis, polymerase chain reaction, pneumonia, on ventilation, with multiple endocrine neoplasia and excluded patients who are pregnant, with cancer, use hydroxychloroquine, with HIV, myocardial infarction, mycoses, glucosephosphate dehydrogenase deficiency, liver diseases, immunosuppression, hepatitis B, on ECMO, and with chronic kidney diseases. The parsing results are provided in the Supplementary Material II.

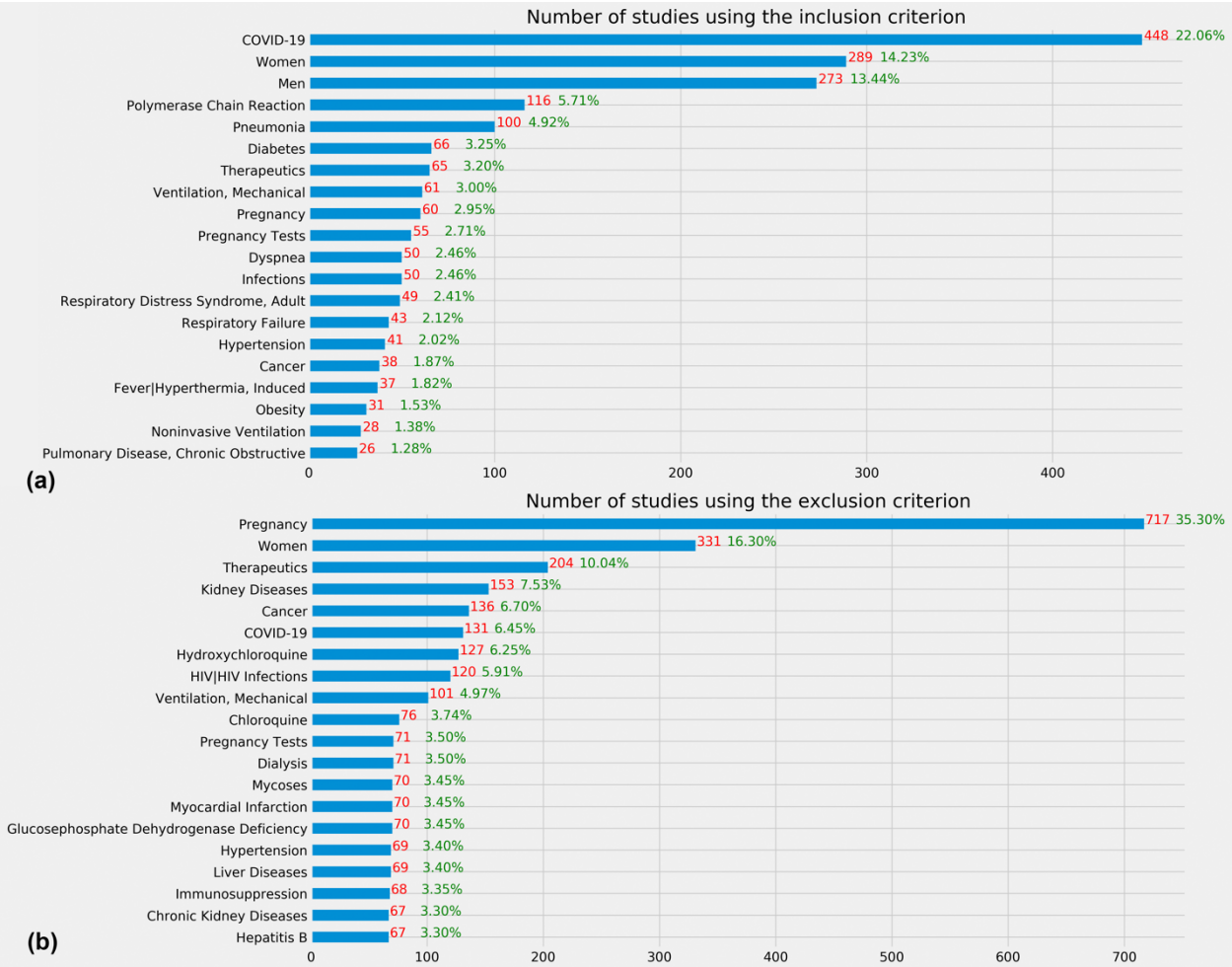

Figure 4. Frequent eligibility criteria of COVID-19 clinical studies. The denominator is the 2034 clinical studies included in this study.

Figure 5 shows the number of studies that used a common chronic condition prevalent among older adults in eligibility criteria of the included studies. 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, and chronic kidney disease, COPD, and diabetes are among the most frequently used exclusion criteria in 2.76% - 8.42% studies. Few studies purposely included patients with a risk factor that may lead to serious illnesses but few studies explicitly excluded them except for 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: 1.38 vs. 0.25, $p<0.001$, 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).

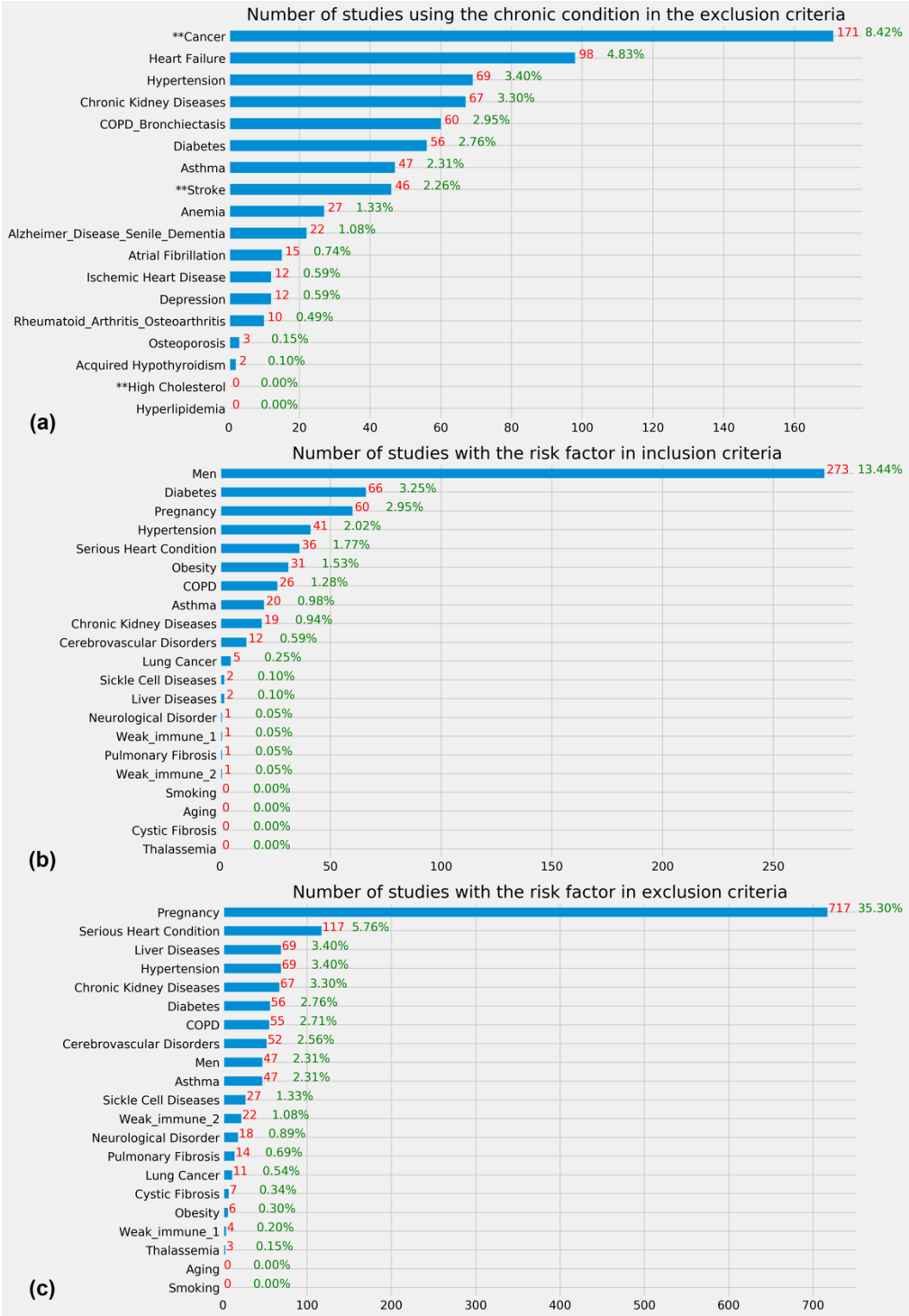

Figure 5. (a) Number of studies using a prevalent chronic condition among the older adults in exclusion criteria. ** represents the conditions that are not in the list of top 15 prevalent conditions among older adults but prevalent in younger adults. (b) Number of studies with the risk factor in inclusion criteria (c) Number of studies with the risk factor in exclusion criteria. The denominator of these three figures is the 2034 clinical studies included in this study.

Table 4 shows the top 10 frequent concepts in inclusion criteria and exclusion criteria used in the studies in each of the 11 clusters. Studies in Cluster #0 mostly excluded patients with HIV/HIV infection and pregnant women. Studies in Cluster #1 all recruited patietns on mechanical ventilation. Studies in Cluster #2, #4, and #6 often excluded pregnant women. Studies in Cluster #6 also often excluded patients with myocoses. Studies in Cluster #3 all excluded patients with COVID-19 diagnosis. Studies in Cluster #5 mainly recruited women and patients with multiple endocrine neoplasis. Studies in Cluster #7 all included patients with polymerase chain reaction. Studies in Cluster #8 mostly excluded patients who used a therapeutics. Studies in Cluster #9 all recruited patients with COVID-19 diagnosis. Studies in Cluster #10 have almost no inclusion criteria and mostly excluded patients who used Hydroxychloroquine. Figure 6 is the two-dimensional visualization of the 11 clusters using t-SNE. The detailed result of the clustering analysis of the COVID-19 clinical studies is provided in the Supplementary Material II.

Table 4. Top 10 frequently used concepts in inclusion criteria and exclusion criteria of the studies in each cluster.

Cluster Number	Number of Studies	Inclusion Criteria	Exclusion Criteria
0	129	COVID-19 (43), Women (42), Men (38)	HIV/HIV infection (106), Pregnancy (92), Hepatitis B (51), Women (46), Therapeutics (31), Cancer (31), Immunosuppression (22)
1	48	Ventilation mechanical (48), COVID-19 (17), Men (10), Women (9), Respiratory failure (7)	Pregnancy (22), Chronic kidney disease (7), Dialysis (7), Therapeutics (7), Ventilation mechanical (7)
2	819	COVID-19 (173), Women (85), Men (54), Pneumonia (50), Therapeutics (41)	Pregnancy (306), Women (146), Cancer (67), Ventilation mechanical (48), Pregnant women (46)
3	48	COVID-19 (19), Women (3), Fever (2)	COVID-19 (48), Pregnancy (15), Cancer (6), Women (4), Kidney failure (2), liver failure (2), Hydroxychloroquine (2)
4	27	Pneumonia (3)	Pregnancy (27), Women (7), Extracorporeal membrane oxygenation (1)
5	105	Women (87), Men (66), Multiple endocrine neoplasia (40), COVID-19 (23), Pregnancy tests (8)	Pregnancy (51), Women (27), Ventilation mechanical (10), COVID-19 (9), Therapeutics (9)
6	88	Women (29), Men (21), COVID-	Pregnancy (61), Mycoses (54), Virus

		19 (18)	diseases (37), Bacterial infection (35), Therapeutics (34), Infections (28), Women (21), Tuberculosis (18)
7	51	Polymerase chain reaction (51), COVID-19 (7), Pneumonia (3), Cancer (3), Men (2)	Pregnancy (20), Women (11), Liver diseases (4), Pneumonia (4), Pregnancy tests (4)
8	77	COVID-19 (19), Women (10), Men (8)	Therapeutics (64), Pregnancy (33), Women (17), Pregnancy tests (11), Bipolar disorder (10), Liver diseases (9), Language fluency English (7)
9	98	COVID-19 (98), Non-invasive ventilation (2), Women (2), Hospitalization (1),	Pregnancy (13), Women (6), Pregnant women (2), Pharmaceutical preparation (2), Cancer (2), Pneumonia (1)
10	115	COVID-19 (31)	Hydroxychloroquine (87), Pregnancy (74), Glucosephosphate dehydrogenase deficiency (58), Chloroquine (57), Women (43), Porphyrias (39), Azithromycin (30), Psoriasis (28), Epilepsy (26)

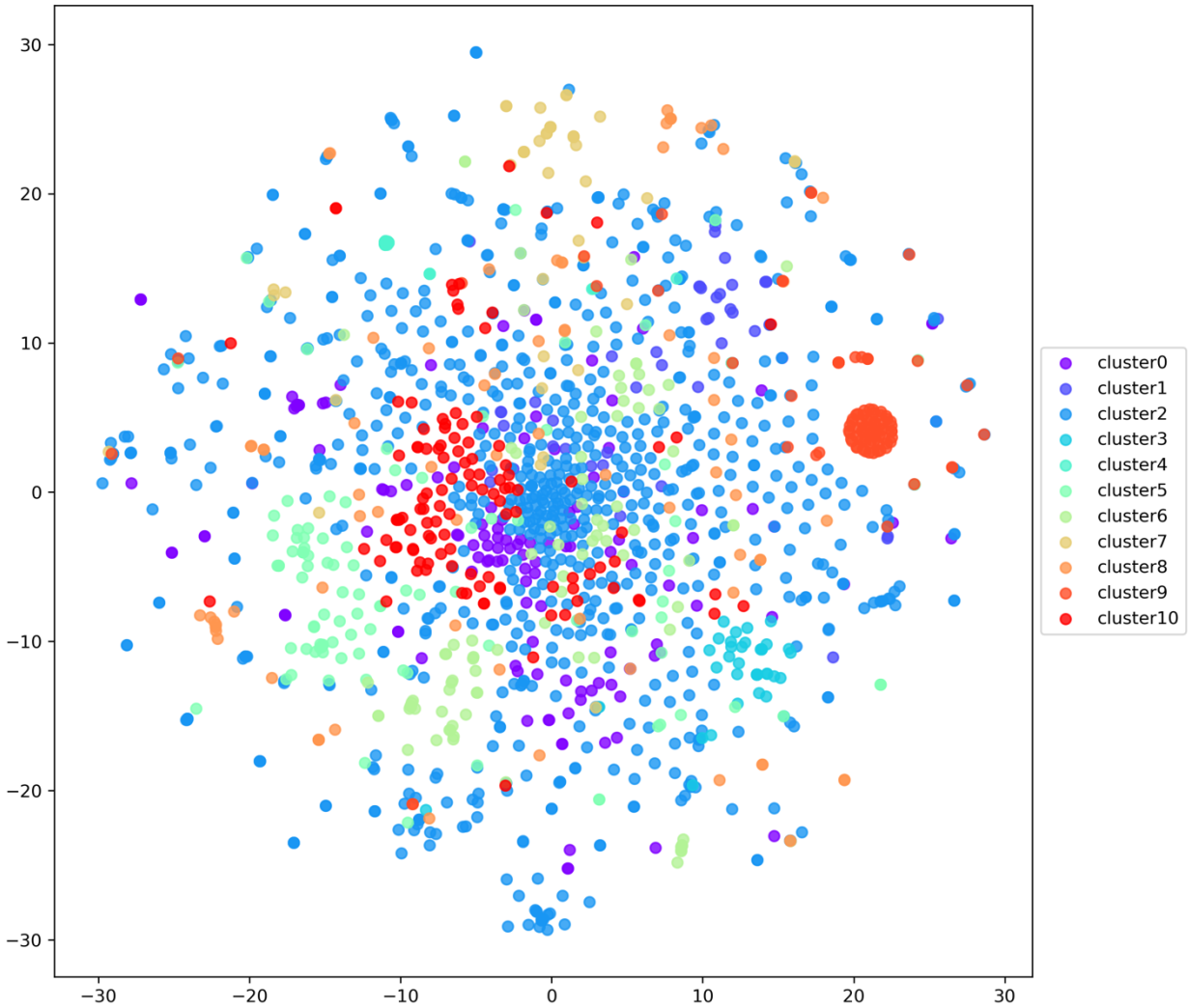

Figure 6. Visualization of the 11 clusters using t-SNE.

Discussion

As the novel coronavirus COVID-19 have significantly impacted our lives and even taken lives of hundreds of thousands of people in the past 8 months, 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).

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 [19]. 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. This study is a necessary step towards data-driven understanding of the research gaps and clinical trial design issues.

As COVID-19 is a major health crisis that impacts people regardless of their age, gender, and race/ethnicity, it is our interest to understand if clinical studies on COVID-19 adequately considered the representation of real-world population. Based on our analysis, most clinical studies consider both genders (97.5%, N=1,983), do not have an upper age limit (75.5%, N=1,536), and have a lower age limit of 18 (81.9%, N=1,667). 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 [20]. 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 (75.5%, N=1536) 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 [7]. 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 [21]. The generalizability of the study results to the real-world population should be evaluated with state-of-the-art techniques [6]. 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 [22]. 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 [23]. Clinical studies should adequately evaluate the efficacy and safety of treatments and vaccines on vulnerable population groups.

Limitation

A few limitations should be noted. First, some data in ClinicalTrials.gov are missing. For example, 31.8% (N=396) of the interventional studies miss study phase information. 46.6% (N=947) 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 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 [6].

Acknowledgments

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

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 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. FE 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] 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.
[6] 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.
[7] Span P. Older Adults May Be Left Out of Some Covid-19 Trials. The New York Times. 2020.
[8] 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.
[9] ClinicalTrials.gov. History, Policies, and Laws - ClinicalTrials.gov 2020 [7/10/2020]. Available from: https://clinicaltrials.gov/ct2/about-site/historyNPRM.
[10] 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.
[11] 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.
[12] Clinic M. COVID-19: Who's at higher risk of serious symptoms? 2020 [07/14/2020].
[13] Soldaini L, Goharian N, editors. QuickUMLS: a fast, unsupervised approach for medical concept extraction. MedIR workshop, sigir; 2016.
[14] Tseo Y, Salkola M, Mohamed A, Kumar A, Abnousi F. Information Extraction of Clinical Trial Eligibility Criteria. arXiv preprint arXiv:200607296. 2020.
[15] 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.
[16] 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.
[17] Maaten Lvd, Hinton G. Visualizing data using t-SNE. Journal of machine learning research. 2008;9(Nov):2579-605.
[18] 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.
[19] 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.
[20] 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.
[21] 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.
[22] 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.
[23] 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.