Clinical Trials Data Analysis

CS:4470 Health Data Analytics

Project #1

Spring 2021

Group 4

Kelvin Encarnacao, Stella Lee, Zhao Zhang

Group 4 Work Distribution

Kelvin Encarnacao

1. Part A – Eligibility criteria, results availability, design aspects
2. Part B – 3
3. Part C – data collection, grouping and visualization.
4. Conclusion

Stella Lee

1. Data analysis + Report
    a. Dataset collection (filtering)
    b. Part A – Phase of study, Activity status, Type of trial, Study duration, Type of intervention, Enrollment
    c. Part B - 1
2. Report
    a. Creating the format and outline of the report
    b. Introduction
    c. Fixing the figures and table format

Zhao Zhang

1. post adverse events analysis for part B
2. search for the part c data

1. Introduction

ClinicalTrials.gov provides detailed descriptions of clinical trials online. It helps people to access which trials are happening, recruiting, and learning the recently available results. However, the database has a vast amount of data. Important aspects inside data might not be noticeable. In this report, we will analyze various characteristics of clinical trials related to COVID-19 and Hepatitis A. While COVID-19 was recently discovered, Hepatitis A could be prevented by vaccine since 1996 in the US. By extracting information from clinical trials, we will show the trend in clinical trials of two distinct diseases.

2. Method

2.1. Data Processing

We downloaded the CSV dataset from ClinicalTrials.gov. by searching for a specific condition or disease. The keyword "Covid19" and "Hepatitis A" was used to search related trials. The search result included 4724 trials for COVID-19 and 4083 trials for Hepatitis A as of February 12, 2021. However, we realized that the search result included some trials that do not relate to our target disease but similar ones. Therefore, we had to remove trials from the dataset. Since some trials included multiple conditions, we split the column value by "|" and check if the list contained the target disease. COVID-19 disease condition was stated in different ways such as covid-19, Covid19. Thus, we first changed conditions to lowercase and then used a regular expression to caption the optional "-" character between "covid" and "19" for filtering. For Hepatitis A, there were no other ways to mention the disease in the dataset. The analysis included clinical trials that at least included target disease in the condition.

2.2. Basic comparison analysis

We used Python libraries (seaborn, matplotlib) and Tableau to create the visualizations for basic comparison analysis. Since the difference between the number of trials for each disease was large, we compared the proportions of measures instead of count.

For analyzing intervention type, we filtered only interventional clinical trials. Then, selected the first available intervention type from each trial if there were multiple types in the same trial.

Since there are many different age ranges accepted to different trials, we chose to group them based on which age groups their range included for analyzing the age eligibility criteria. Child age indicates accepting some age in the range 0-17, adult 18-64, and older adult 65+. Some trials contain multiple ranges for example any that accept persons 18+ would be categorized as adult and older adult.

The study duration is defined as the duration between the completion date and start date. We first filtered trials lacking either start date or completion date. Since the degree of specifying the date was different by trials, if the trials did not have the exact date, we used 1 for the date's default value.

2.3. Most studied drugs that appeared in trials for each disease

We extracted drugs from the intervention description of the dataset. Then, we filtered drug types containing Placebo and Control. Some instances mentioned their drug as "best available treatment" which would not help our analysis. Since our purpose of the analysis is the type of drugs, we eliminated the dose of drugs and injection type (nasal or oral). Furthermore, there could be different ways to mention one drug in clinical trials. Thus, we used Medical Subject Headings (MeSH) term to correctly find what is the most studied drugs. Some drug terms were not yet included in MeSH especially for COVID-19. Therefore, we manually searched on the Internet to find either formula of the drug or added it to our drug list. In addition, there were several typing mistakes in the drug names in the dataset. For example, "Tociliuzumab" was corrected to "tocilizumab" by searching online.

2.4. Most serious adverse events that occurred in trials

Use java crawler to find keywords such as adverse events or outcomes. Collect and count the number and frequency of adverse events from xml files. Then look for keywords of some adverse events, record the number and frequency of occurrences, and make a table of them. The most serious adverse events happened in covid 19 is Coronavirus infections. For hepatitis A, the most serious adverse events are Hepatitis and Hepatitis A.

2.5. Trends in the number of trials

Extract trial information using start date as grouping for each year. Filter out COVID-19 data that was incorrect for example some trials on COVID-19 had a listed start date of 2003 which is incorrect. Visualize in Tableau using the number of trials with a listed start date in each specific year and plot over time showing % of total trials done in each year since the clinical trials for that disease have existed.

2.6. Location

Using python libraries read dataset and grouped and counted each trial based on the text looking for specific countries. For each trial if a country or multiple country were listed in the location information then add that trial to the number of trials for each country. Visualize using Tableau symbol map to show percent of total trials located in each country and filter out any null information.

3. Result

3.1. Dataset

We extracted 4724 COVID-19 trials and 4083 Hepatitis A trials from ClinicalTrials.gov as of February 12, 2021. Nevertheless, the search result included trials only with similar conditions. After excluding unrelated trials, 2949 COVID-19 related clinical trials and 70 Hepatitis A related clinical trials were included in the analysis.

3.2. Basic comparison analysis

3.2.1. Phase of study

Based on Figure 1, Hepatitis A trials are concentrated on Phase 3 (30.65%) and Phase 4 (50%). However, COVID-19 related trials had majority on Phase 2 (25.47%) and Not applicable Phase (28.65%). It is notable that there is no Hepatitis A clinical trial that is currently on Early phase 1 to Phase 1|2.

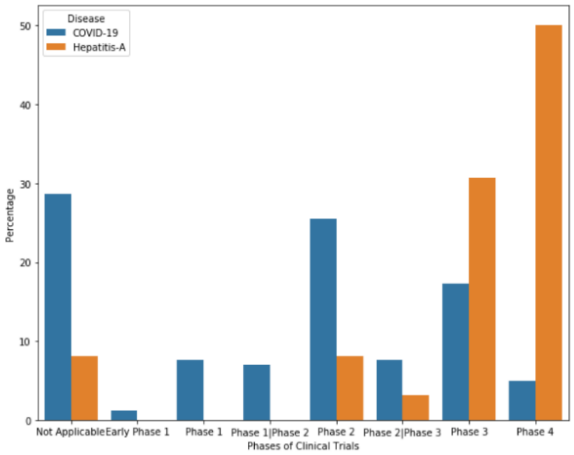

Figure 1: Phases of COVID-19 and Hepatitis A Clinical Trials

3.2.2. Activity status

The activity status indicates how clinical trials are progressing with the recruitment. According to Figure 2 and 3, 77.14% of Hepatitis A trials completed the recruitment but only 14.31% of COVID-19 trials are completed. For COVID-19, 50.9% was recruiting or not yet recruiting. (19.57%)

3.2.3. Study type of trial

Study type of trials are either interventional, observational, or expanded access according to ClinicalTrials.gov. Interventional clinical trials assign groups to participants while observational trials do not need to. Based on Figure 4, both disease trials have interventional type as the highest percentage. (59.65% for COVID-19 and 88.57% for Hepatitis A) Expanded access trials, which are using medical products are not approved by the U.S. Good and Drug Administration (FDA), only appears in COVID-19 trials. (0.75%)

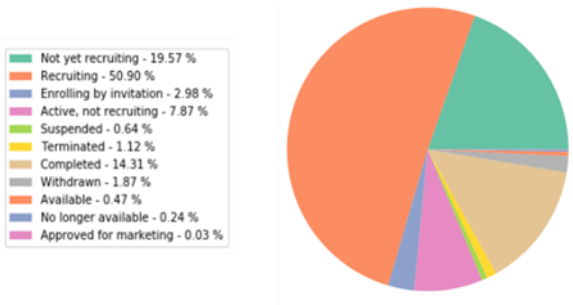

Figure 2: Activity Status of COVID-19 Trials

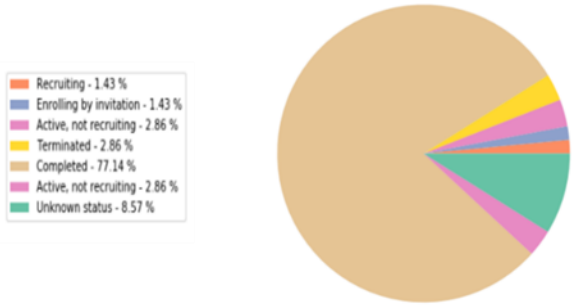

Figure 3: Activity Status of Hepatitis A Trials

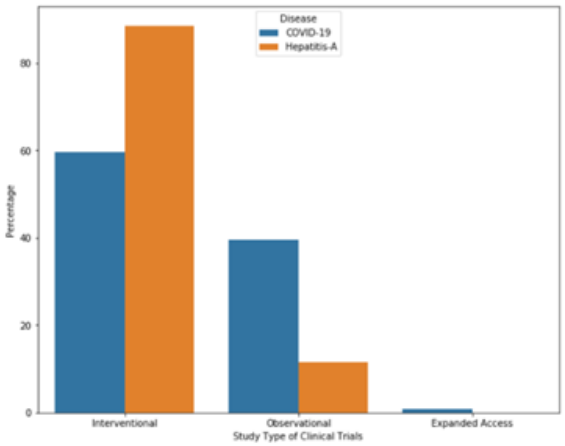

Figure 4: Study Type of COVID-19 and Hepatitis A

3.2.4. Eligibility criteria (age and gender)

For eligibility criteria, we compared the listed and accepted age groups and gender for both the extracted COVID-19 and Hepatitis A trials. Looking at Figure 5, the age groups for COVID-19 are predominantly only accepting adults and older adults, with 86.22%. This is most likely because the disease is new and there are still many unknown factors that could be have unknown effects on children. With Hepatitis A, child only trials were the majority with 47.14% most likely due to Hepatitis A being a more studied disease that is proven to be mostly found in children. With regards to gender, the results were similar. Both groups of trials accepted both gender in most of their trials with COVID-19 having 97.50%, and Hepatitis A having 98.57%

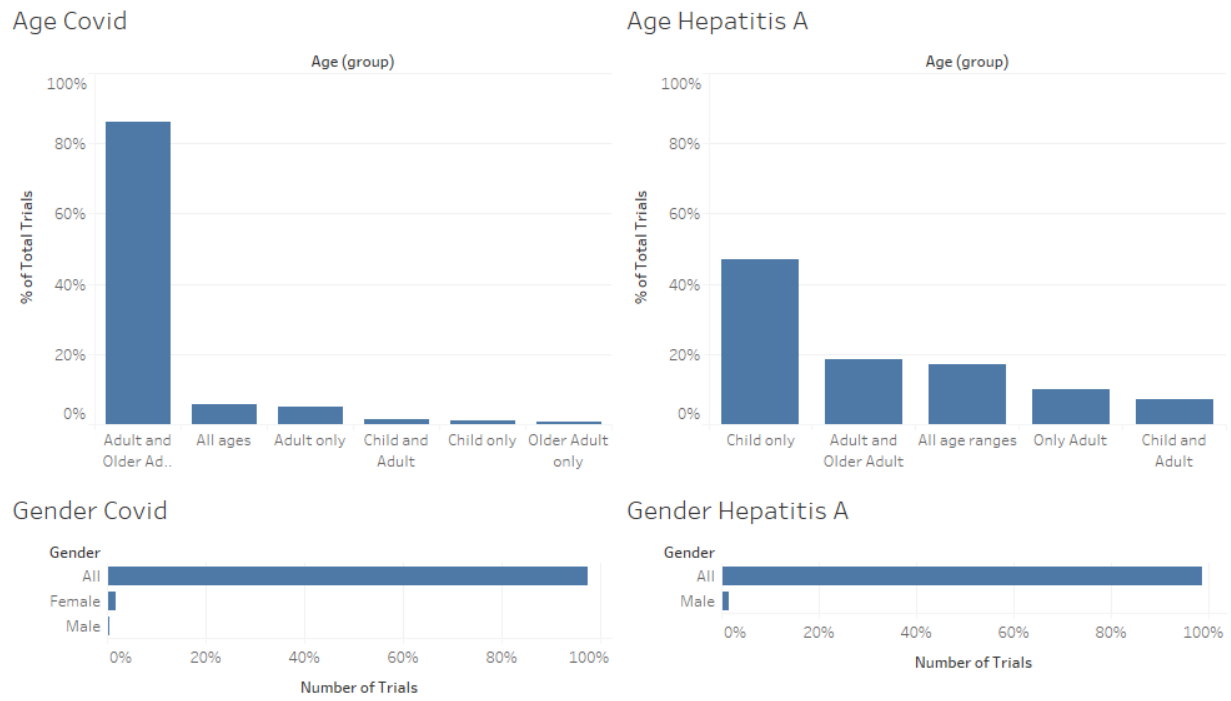

Figure 5: Age and Gender distribution of COVID-19 and Hepatitis A trials

of trials accepting all genders. In the 70 trials from Hepatitis A there were no female only trials while the COVID-19 set contained 1.88% but this can be attributed to the lack of data from Hepatitis A.

3.2.5. *Results availability*

We looked at the availability of results for both COVID-19 trials and Hepatitis A trials categorized as either having results or having no results available. Based on table 1 we can see that almost all the COVID-19 trials have no results available with 99.40% not having results. In contrast, Hepatis A trials as seen in the table have 37.14% trials with available results with much higher than COVID-19 most likely attributed to the timeline of Hepatitis-A being much longer than COVID-19 which was only recently discovered in the past years.

	Has Results	No Results
COVID-19	0.60%	99.40%
Hepatitis A	37.14%	62.86%

Table 1: Results Availability for COVID-19 and Hepatitis A

3.2.6. *Design aspects*

For design aspects we looked at both allocation which tells us whether the trial was randomized or not, and we looked at the interventional model for the trial (single group, parallel, etc.). Looking at Figure 6, we can see that for both COVID-19 and Hepatitis A parallel assignment was the most common interventional model use followed by single group. The other interventional models followed a similar pattern of total % but Hepatitis A did not contain any trials with sequential assignment. For allocation, the majority for both COVID-19 and Hepatitis A were randomized and N/A (allocation not applicable most likely in case of single arm trials). Hepatitis A contained 65.71% randomized while COVID-19 contained 44.57% randomized. This shows good design aspects used in trials which can be best seen when comparing randomized to non-randomized and seeing in COVID-19 there are almost 10x as many randomized trials as non-randomized and for Hepatitis A the number is almost 7x.

3.2.7. *Study duration*

We analyzed the duration of study for COVID-19 and Hepatitis A. There were 22 COVID-19 clinical trials without start date or completion date information. The mean study duration of COVID-19 was 13.46 months with 25.87.5 standard deviation. The mean duration of Hepatitis A trials was 40.07 months with 34.45 standard deviation. Based on Table #, study duration of trials have extreme outliers. There exist 47 COVID-19 clinical trials with same start date and completion date.

	min	25%	50%	75%	max
COVID-19	0.0	4.39	9.75	15.21	1039.21
Hepatitis A	1.07	13.04	30.43	59.23	151.14

Table 2: Descriptive Statistics of Study Duration of COVID-19 and Hepatitis A Trials in months

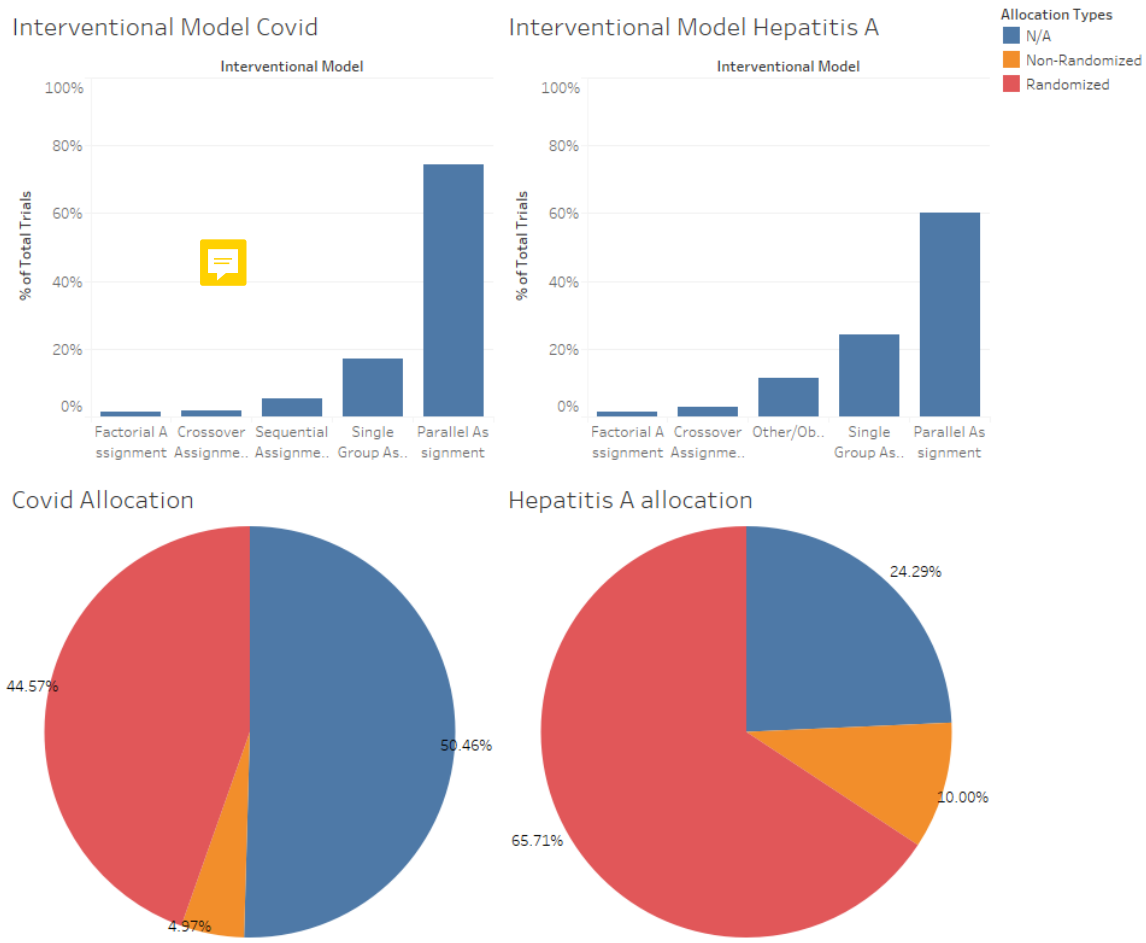

Figure 6: Allocation and Interventional Model Distribution for COVID-19 and Hepatitis A

3.2.8. Type of Intervention

We compared the types of intervention for each disease. As we mentioned earlier in section 3.2.3, 59.65% of COVID-19 trials and 88.57% of Hepatitis A trials were interventional. Based on Figure 7, we could observe that Hepatitis A interventional trials are focused on biological (48.57%). While COVID-19 trials tend to be more evenly distributed, drug had highest proportion of interventional trials. (34.18%)

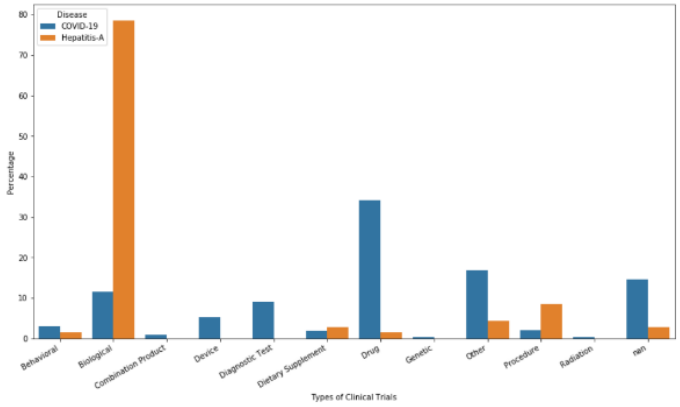

Figure 7: Intervention Types of COVID-19 and Hepatitis A Trials

3.2.9 Enrollment

We analyzed the number of enrollments of each disease trials. The mean enrollment of COVID-19 was 29109.13 with 553706.5 standard deviation. The mean enrollment of Hepatitis A was 1185.15 patients with 4454.28 standard deviation. Based on Table 3, the number of enrollments of trials have extreme outliers such as "The Doctors for Coronavirus Prevention Project Thanksgiving Messaging Campaign" trials with 20000000 enrollments.

	min	25%	50%	75%	max
COVID-19	0.0	60.0	174.0	561.0	20000000.0
Hepatitis A	0.0	89.75	277.0	583.5	35000.0

Table 3: Descriptive Statistics of Enrollment number of each disease trials

3.3. Most studied drugs that appeared in trials for each disease

We analyzed what is most studied drugs for each disease. For Hepatitis A clinical trials, only one drug appeared: Vaqta injectable product. This drug is Hepatitis A vaccine in MeSH term. For COVID-19, we could observe that Hydroxychloroquine was most frequently studied drugs in the clinical trials.

Rank	Drug Name	Frequency
1	Hydroxychloroquine	138
2	Azithromycin	41
3	Ivermectin	38
4	Tocilizumab	37
5	Remdesivir	33
6	Favipiravir	31
7	Heparin	29
8	Ritonavir	27
9	Lopinavir	25
10	Enoxaparin	24
11	Dexamethasone Isonicotinate	20
12	Saline Solution	19
13	Interleukin 1 Receptor Antagonist Protein	17
14	Colchicine	16
14	Nitazoxanide	16
14	Chloroquine	16

Table 4: 15 Most Studied Drugs that Appeared in COVID-19 Trials

3.4. Most serious adverse events that occurred in trials

For the most serious adverse event of the Covid19, the Coronavirus Infections are most common appear through all the clinical trials, there are 6 times out of total trials. For the Hepatitis A, the most common adverse events are Hepatitis and Hepatitis A, they appeared both 484 time each. Most of the adverse events of the Hepatitis A are close, it shows that the adverse events always appeared together.

Adverse Events	Frequency
Communicable Diseases	4
Coronaviridae Infections	3
Coronavirus Infections	6
Infection	4
Lung Diseases	2
Nidovirales Infections	3
Pneumonia	2
Respiratory Aspiration	1
Respiratory Tract Diseases	3
Respiratory Tract Infections	3

Table 5: Adverse Events of COVID-19

Adverse Events	Frequency
Hepatitis	484
Hepatitis A	484
Communicable Diseases	480
Infection	480
Virus Diseases	477
Digestive System Diseases	470
Gastrointestinal Diseases	470
Liver Diseases	470
Hepatitis, Viral, Human	467
RNA Virus Infections	456

Table 6: Adverse Events of Hepatitis A

3.5. Trends in the number of trials over time

For trends in trials over time we tried to figure out any significant patterns of when higher numbers of trials occurred on a given year for COVID-19 or Hepatitis A. To do this we looked at the start date for each trial and based on that assigned each trial to a specific year. For COVID-19 this information was less useful because as the disease is much newer than Hepatitis A and trials did not begin until (late) 2019 when the coronavirus was first discovered. Using figure 7 we can see that looking at the years 2020 had the most trials with 88.67% of trials having a start date in 2020. Currently 2021 has listed 10.04% of trials for COVID-19 which is sure to go up at new trials are still being conducted at a fast rate. For Hepatitis A the trials per year were much steadier with most years sitting around 3-7% of total trials done. There were however some spikes in the graph that occurred in 2003 with 10% of trials, 2004 with 11.43%, 2008 with 11.43% and 2017 with 10%. Over the 24-year span of collected trials for Hepatitis A, a relatively even distribution among all years can be seen with some years having a bit more.

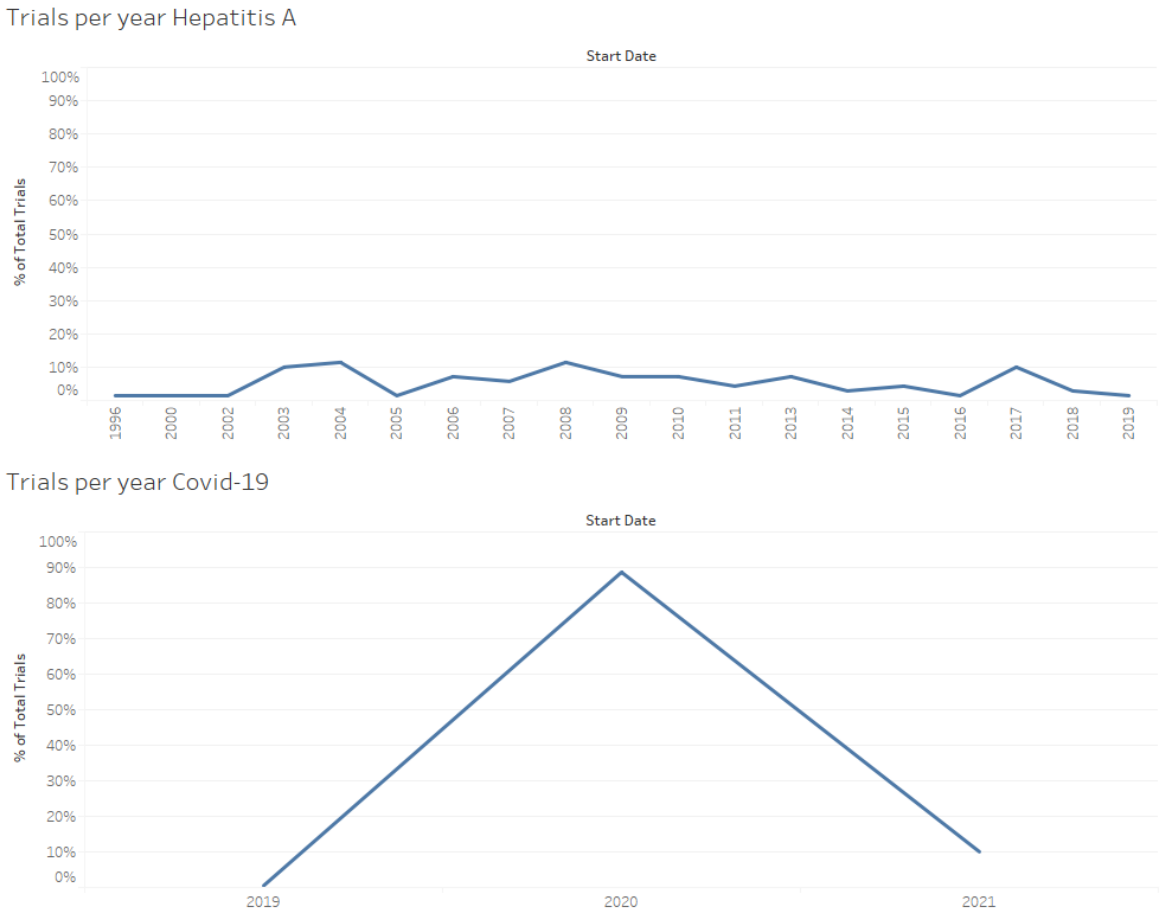

Figure 8:Trials Yearly Distribution for COVID-19 and Hepatitis A

3.6. Location

Another question we raised about the two data sets was whether we could find a pattern in the frequency of trials done in specific locations around the world. To answer this question, we went through the location data from the trials and grouped them by country. Many of the trials were done simultaneously together in multiple countries so to account for this we added trials multiple times to multiple groups for each different country to not take away from one country or another when grouping the trials. After grouping each trial and filtering out null values we used a symbol map in order to plot dots for each country with size varying based on the percentage of total trials located in each specific country. Looking at figures 9 and 10 we can see cluster patterns of where many trials are held and the diversity of locations for each set of trials. For the COVID-19 data we can see the highest percent of trials held in the US with 20.53% followed by other world superpowers like France with 10.92%, the UK with 4.08% and China with 3.24%. Another thing to note from figure # is the spread of dots on the map showing the diversity of locations of the trials for COVID-19 as we can see less but still a substantial number of trials in South America, Africa, and Eastern Europe, this is most likely due to COVID-19 being so scary and life-threatening to the world that every country is taking resources in order to help prevent this pandemic from getting worse. Looking at the map for Hepatitis A in figure # we can see similar patterns of the world richer and bigger countries holding most trials as Belgium has 20%, the US has 10%, and China has 14.29%. Figure # shows on the other hand that for the limited number of trial data for Hepatitis A, there is very little being done in Africa, South America, and Eastern European countries, this is most likely due to Hepatitis A being something we already have a vaccine for and less important that poorer smaller countries would not want to waste money and resources conducting clinical trials for.

Figure 9: Trial Distribution by Country Hepatitis A

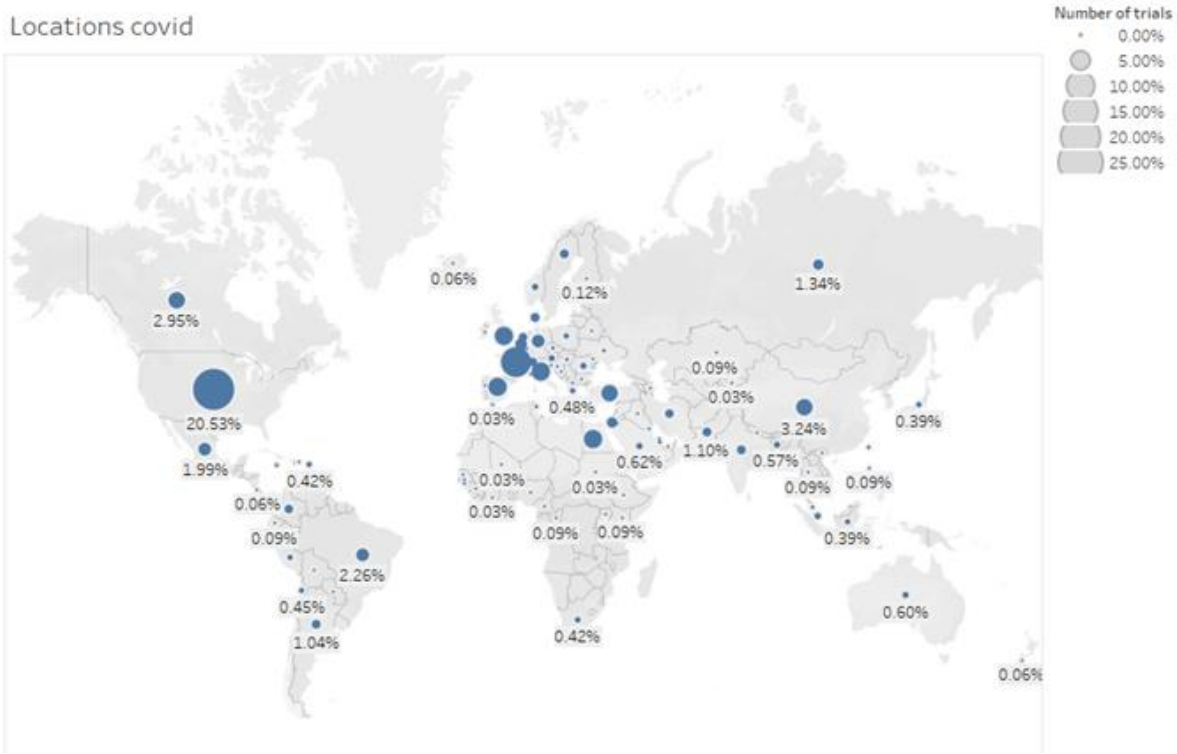

Figure 10:Trial Distribution by Country COVID-19

4. Limitation

We did not expect the number of clinical trials related to Hepatitis A (70 trials) would be small compared to COVID-19. The recency of COVID-19 also attributed to some problems with the data specifically those looking at trends over time since there is not much of a timeline and really only one year of data to look at. Recency also can be used to explain some values in the table such as % of COVID-19 trials with published results as most trials are still ongoing.

5. Conclusion

Comparing the two datasets there were many insights and conclusions we could draw about the data and the diseases themselves with regards to clinical trials. For eligibility criteria looking at the COVID-19 data in comparison with the Hepatitis A data we can see a more willingness to experiment and try different criteria with regards to age and gender because it's a new disease, while Hepatitis A is more predictable with children being the main target for trial testing. For design aspects we can see that good trial practices are being practiced with both disease trials as there is a good distribution for both multiple types of interventional trials as well as good use of randomization when applicable. Looking at trends over time there was not much we could learn from this based on the recency of the COVID data, but we can see a good steady distribution for the Hepatitis A trials over time. Finally looking at trends by location there were some interesting things we could conclude from the symbol map. One can see the significance COVID-19 has had on the world with the very broad distribution across all continents and the time in effort being put into learning more about this disease while Hepatitis A can predictable only be seen in trials in 1st world bigger countries that have the time and resources to do research on something that is not effecting the entire population.