Comparing Clinical Trial Datasets for COVID-19 and Hepatitis A

Holly Bennett, Yoon Choi, Cami Reittinger

# Introduction

In the past year, the desire for knowledge regarding COVID-19 has exploded. With this quest for information came an increase in COVID-19 related clinical trials. On the other hand, Hepatitis A is an infection that has been around for decades. The differences between the two diseases are numerous, both with their relative importance in the world currently and how they affect the human body. COVID-19 is a respiratory infection while Hepatis A is a liver infection. Despite their differences, they are both infectious diseases that are vaccine preventable. Our objective in this study was to determine whether two very different diseases follow a similar format for analysis in clinical trials. We downloaded data from ClinicalTrials.gov for both diseases, extracted relevant information for both diseases on study design, trial information, and specific drug and adverse event data, and finally created charts to analyze the similarities between the two.

# Methods

The first step in analyzing the two conditions was to download the required data from ClinicalTrials.gov. Of the two options (downloading the data or parsing AACT), our group decided the most efficient method for analyzing the dataset was to download the data directly from ClinicalTrials.gov. On the site, there are multiple options for exporting data. Most of our analyses could be completed by downloading the table contents in a 1 row per study xml file. However, for more advanced portions of the project, such as querying adverse event data, we needed to download the full study XML records. This option exports a zip file containing a folder storing every trial in a separate XML file. For basic search operations this step is unnecessary, but for more in depth analyses this method is required.

Once the data was exported, the next step was to parse the files and extract all relevant information. We decided on using Python for our parsing language. Rather than parsing the XML contents directly, we decided on converting the data into a JSON file. This method was an extra step, but it allowed for easier data parsing which saved both time and effort. Utilizing Python’s json, os, and xmltodict libraries, the downloaded XML file was converted to JSON with a simple function.

For a given search term, ClinicalTrials.gov returns trials using an automated algorithm based on MeSH terms. This meant that some of the trials that were exported did not have the provided search query listed as a condition. As the focus of our study was comparing trials for COVID-19 versus Hepatitis A, we filtered the dataset even further to only contain trials where one of those two diseases was listed as a condition.

Our filtering process was a simple iterative function that took the data dictionary as an input and returned a dictionary with only the relevant results. The original search query was stored in the JSON object, so each study’s conditions were compared to the original search term. All studies that did not contain the search term were discarded.

# 2.1 Section A:

With the data stored in an efficient format for parsing, we began extracting relevant data for part A. Using the Python json library, we loaded the data file into our code. In total, 13 functions were created and utilized to extract the relevant data as outlined in the project specs file for part A. This included phase data, activity status, type of trial, eligibility criteria, result availability, design aspects, study duration, and our two features of choice, intervention data and location information. Our main function loaded the JSON file into a dictionary and called the remaining functions: getStudyDuration, getStudyResults, getAgeGroups, getMinMaxAge, getGender, getStudyType, getInterventionStatus, getActivityStatus, getStudyDesignData, getPhaseData, and getLocationCount. Each function had a single parameter input of the data object dictionary and returned a dictionary with the relevant extracted data. The general format of each procedure was as follows

1. Create a dictionary for storing the information.
2. Iterate over the dataset.
3. Extract the desired data from the current study.
4. If that item was not stored in the new dictionary, add the item as a key with a value of 1
5. If that item was stored in the new dictionary, add one to the value stored at the matching key.
6. Return the dictionary storing all the keys and corresponding value representing the data found in each study for a given procedure and the frequency at which they occurred.

Some of the functions required more advanced parsing methods.

* getLocationCount extracted data regarding the locations for each study. This data was stored in the dictionary object under the key ‘locations’. Some files had no location information available, and thus we had to filter through studies to add those with a NoneType value for the locations key to the dictionary key counting studies with 0 locations. For the remaining studies, locations could either be stored as a Unicode string or a list of Unicode strings. Those with a list had to be iterated over again to count the total number of locations found. Those with a Unicode string represented studies with only one location provided.
* getPhaseData extracted data regarding the phase of each study. Phases were either stored as a NoneType, meaning no data was available, a Unicode type meaning one phase was listed, or as a list which represented a trial with multiple phases provided. The NoneType and Unicode studies were simply added to the dictionary to count frequency. The list types were converted to a string joining the two phases together. For instance, [‘Phase 1’,’Phase 2’] was converted to ‘Phase 1/Phase 2’. Once converted, these objects were added to the dictionary containing the phase data.
* getStudyDesignData extracted data regarding the study design portion of each study. This section was not uniform among clinical trials. Some trials had no design data, some had design data stored as a string, and the majority of trials had a list containing the study design information. Within the lists were six possible categories of information: the primary study purpose, intervention model, allocation, masking information, time perspective, and observational model. We chose to extract information for all of these categories, as the study design of a trial is an important component in a clinical trial. The getStudyDesignData function returned 6 dictionaries containing study design and frequency information for each of the aforementioned categories.
* getInterventionStatus extracted data regarding the invention of a clinical trial. This information is interesting as the number of interventions per trial differs. Thus, we extracted both the types and frequencies of interventions as well as calculated the average number of interventions per trial.
* getStudyDuration extracted data regarding the length a trial was conducted. Some trials had their completion date listed under the ‘primary\_completion\_date’ tag while others stored the information under the ‘last\_update\_posted’ tag. The majority of trials used the former, and we utilized the Python library dateutil.parse to convert the date stored as strings into a datetime object. This library was extremely useful as many of the dates entered were saved under differing formats, and the dateutil library converted every date to a uniform format. We then calculated the difference in months between the start and end date and returned a dictionary containing the month data and their frequencies.

# 2.2 Section B:

We repeated section A in Section B. In total, 3 functions were created and utilized to extract the relevant data as outlined in the project specs file for part B. This included most studied drugs, frequent adverse events, and trends over the years in the number of trials for both drugs. Our main function loaded the JSON file into a dictionary and called the remaining functions: getMostStudiedDrugs, getAdverseEvents, and getTrends. Each function had a single parameter input of the data object dictionary and returned a dictionary with the relevant extracted data. In the visualization, we incorporated box chart to look at the median and general distribution of the trends over the years in part 3 of section B.

# 2.3 Section C:

We decided to compare the sponsors in the trials for each disease. We wanted to see who the lead sponsor was the most and who collaborated with others the most.

------

For each function returning a dictionary, the information then had to be saved to a CSV file to allow for diagram creation and more in-depth analysis. This required the use of the Python library, CSV. Each key value pair in each data dictionary was parsed and written to the corresponding CSV file.

Finally, data analysis was completed using Excel. Matching CSV files from both Hepatitis A and COVID-19 were opened in a single Excel workbook. This allowed us to use the ‘vlookup’ feature to combine data from both conditions into one table for a more direct analysis. Excel also has charting and graphing features that we used to visually analyze all data. Any calculations such as the average number of interventions per trial were calculated using Excel.

# Results

We exported Hepatitis A and COVID-19 trials from ClinicalTrials.gov on February 8th, 2021. At that time there were 4,686 trials pertaining to ‘COVID-19’ and 4,080 trials pertaining to ‘Hepatitis A’. After filtering the results according to the original search query, we were left with 1,536 COVID trials and 71 Hepatitis A trials. To account for this extreme difference in dataset size, all remaining analyses utilized percentages rather than total numbers.

# 3.1 Section A Results:

## 3.1.1 Phase of Study

Clinical trials can be classified into 5 different phases, 0 (or early phase 1) through 4. Phase 0 indicates a trial that is learning how a drug is processed by the body on a very small group of patients and Phase 4 indicates a trial that is testing FDA approved drugs in more participants. As the phase number increases, so does the number of participants and the information obtained about the drug being tested.

The majority of COVID-19 trials had no phase information provided, while the majority of Hepatitis A trials were in phase 4 at the time of analysis. It is most likely the case that trials with ‘None’ as the current phase were actually in Phase 0/Early Phase 1, meaning they are just starting out. This makes sense as the desire to learn more about COVID-19 is a new concept within the past year, indicating new clinical trials. Phase 4 clinical trials are in the last stages of drug testing after FDA approval, which aligns itself with an ‘older’ disease such as Hepatitis A.

## 3.1.2 Activity Status

Activity status can be generalized as open or not open, where a non-open trial is either completed, terminated, or not accepting participants. Of the 1,536 COVID-19 clinical trials, only 474 (31%) were not open, and for Hepatitis A, 70 of the 71 trials were not open. As mentioned earlier, COVID-19 trials are up and coming as this particular strain of the disease was first seen in 2019, so newer trials align with more open trials.

Other than just open or not open, a trial can be further classified based on its recruiting status. Hepatitis A had a miniscule one singular trial that was open and recruiting, while COVID-19 had 797 trials that were open and recruiting patients.

## 3.1.3 Type of Trial

Clinical trials can be one of three types: interventional, observational, or expanded access. Each of these types of trials include differing design aspects, which was made evident when looking at the design aspects. Despite the difference in total trials conducted for both Hepatitis A and COVID-19, the proportional breakdown for trial types was very similar among the two. Interventional trials making up the majority can be expected, as the goal of an interventional trial is to give some sort of intervention or treatment to evaluate outcome on participants. With two vaccine preventable diseases that are highly contagious, investigating intervention methods to prevent and or treat the diseases would be the most logical approach for a clinical trial.

## 3.1.4 Eligibility Criteria

Eligibility criteria is an important aspect in clinical trials as it lets potential participants and other researchers know who the object of the trial is. While there are more criteria than age and gender, we decided to focus on those two for our analysis. The vast majority of trials for both COVID-19 and Hepatitis A included all genders. Roughly one percent of COVID-19 trials were female only, and there was only one Hepatitis A study that did not include all genders. As for age groups, there are three categories, child (birth-17 years), adult (18-64 years), and older adult (65+). There is no limit on the number of age groups a trial can allow, so we differentiated data from a trial allowing only children from one allowing children and adults.

Overall, COVID-19 trials tended to include adults and older adults, with the older adults’ group eligible for almost 95% of the trials. On the other hand, children were eligible for only 15% of the COVID-19 trials. Hepatitis A was the opposite, with children being eligible for the majority of the trials and older adults eligible for only 35% of the trials.

This difference can be seen further in the average minimum and maximum age eligible for trials in each condition. For COVID, the average minimum eligibility age was 19 years old and the average maximum age was 76 years old. For Hepatitis A, the average minimum age was 10 years old and the average maximum was 20 years old. The maximum age for Hepatitis is almost equal to the minimum for COVID, showing the vast difference in the target audience for the two disease trials.

## 3.1.5 Result Availability

Of the 1,536 COVID-19 related trials, only 12 had results available, equating to less than 1%. Hepatitis A had roughly 37% of trials with available results and 63% with no results. Looking back at (FIGURE??) in the activity status section, Hepatitis A had significantly more trials that were completed than 37%, meaning some of the trials were finished but results were not available. **According to ClinicalTrials.gov**, potential reasonings behind no results available include the study not being subject to United States requirements to submit results, the results submission deadline has not been passed, or result submission has been delayed.

## 3.1.6 Design Aspects

Study design aspects relate to different types of trials. The design of interventional trials is different to that of an observational study, and thus the data available differs. As seen before in the trial type section, there were 532 observational COVID-19 trials, 989 interventional COVID-19 trials, 8 observational Hepatitis A trials, and 63 interventional Hepatitis A trials. The study design aspects we analyzed for observational studies were the time perspective and the observational model. For interventional studies we analyzed the primary purpose, masking, and allocation data available in the study design portion of the results.

### 3.1.6.1 Observational

Time perspective refers to which state of time researchers are examining. For prospective observation, the goal is to watch for developments over a longer time period to determine a future outcome. A retrospective study looks back in time to examine exposure or risk in relation to an outcome that was predetermined when the study was created. Cross-sectional studies look at data at a defined time period.

For both conditions, a prospective time perspective was the most common. Hepatitis A had no retrospective studies, while COVID-19 had 18% of its trials fall under this category. Understanding exposure and risk for a new disease strain like COVID-19 is an important part in determining how to respond to a disease outbreak. As such, it makes sense that COVID trials included some retrospective trials.

The observational study model is the general design for identifying patients and following up during an observational study trial. The different models include cohort, case-only, case-crossover, case-control, family-based, ecologic/community, and other. With a very limited number of Hepatitis A observational studies ,analyzing the model proportions against that of COVID-19 is not very accurate. However, both had the same majority model of cohort, which is a study design that studies a group for biomedical outcomes. The only other remaining model seen in Hepatitis A trials was case-only, which includes only patients with the condition being studied.

### 3.1.6.2 Interventional

For interventional models, we first looked at the primary purpose. Over 90% of the Hepatitis A interventional trials were focused on prevention, while COVID interventional trials focused heavily on treatment. The second highest category for COVID-19 was prevention, which could indicate that newer disease trials focus on treatment, and then slowly turn into prevention studies.

Another strategy of interventional clinical trials is masking, which describes when a party involved in the trial is unaware of which participants have received which intervention. Both COVID-19 and Hepatitis A had the majority of their trials qualified as ‘None’ for masking, which would indicate an open label. This means all parties are aware of which participant received which treatment. The remainder of the trials for each disease were fairly evenly distributed amongst single, double, triple, and quadruple masking. These strategies define the number of parties who are blinded. Some trials only mask the investigators, while others mask everyone involved. Overall, it can be seen that the most popular interventional masking strategy was an open label for COVID-19 and Hepatitis A.

The final strategy we analyzed for interventional trials was allocation, or how participants are assigned to an arm of the trial. There are two types of allocation, randomized or non-randomized, as well as the occasional trial with ‘N/A’ listed. Once again, despite the difference in total studies for both diseases, the proportion of studies assigned to each allocation type was very similar, with the majority being randomized.

## 3.1.7 Study Duration

Study duration describes the time between the study first posted date and the completion date or last update posted date, whichever a study provided. The completion date can be in the future, so some study durations are not exact but an approximation of what the researchers expect. The duration calculated is equal to the difference in months between the date the first participant was enrolled and the date the last participant was studied.

Despite there being a large difference in the first cases of COVID-19 and Hepatitis A, the longest study duration for both diseases was greater than 120 months, or 10 years.

## 3.1.8 2 Additional Features

### 3.1.8.1 Intervention Information

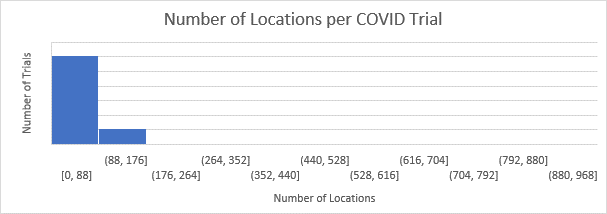
Interventional clinical trials can have multiple interventions of differing types. After calculating the proportion of trials and how many interventions occurred per study, the data for both COVID-19 and Hepatitis A followed a similar plotted line as seen in FIGURE X. The extreme majority of trials had only one or two interventions, and the percentage of trials decreased as the number of interventions increased.

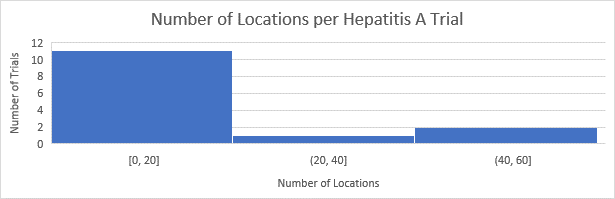
Where the two diseases differed was in the type of intervention. For COVID, 44% of the trials had a drug intervention method, while the majority of Hepatitis A trials, roughly 85%, had a biological intervention method. There were some COVID intervention types that were not found in any of the Hepatitis trials, including radiation, diagnostic test, genetic intervention, and device.

### 3.1.8.2 Locations

The final feature we analyzed for both conditions were the locations. COVID-19 trials had much more locations per trial on average than Hepatitis A. The maximum number of locations for a Hepatitis trial, 42, was nowhere near the maximum number of locations found in a single COVID trial, 882.

We also looked at how many locations were found in multiple trials. Many of the Hepatitis A locations that were found in multiple trials were different GSK Investigational sites found in different cities across the world. The maximum number of trials found at one singular location was a GSK Investigational site in Wilrijk, Belgium. For COVID-19, GSK investigational had no locations. The maximum number of trials found for one location for COVID was 19 trials at Regeneron Study Site, New York, New York, United States. Overall, there were 1,211 COVID locations that had multiple trials, and only 29 Hepatitis A locations with more than one trial.





# 3.2 Section B Results:

## 3.2.1 Most Studied Drug

We saw that Havrix 720 Junior and Vaqta were the most studied drugs for Hepatitis A. After further investigation, we found that Havrix 720 is a Hepatitis A vaccine manufactured by GSK Investigational in Belgium. As we saw in section 3.1.8.2, that location was the most frequent site found for Hepatitis A trials.

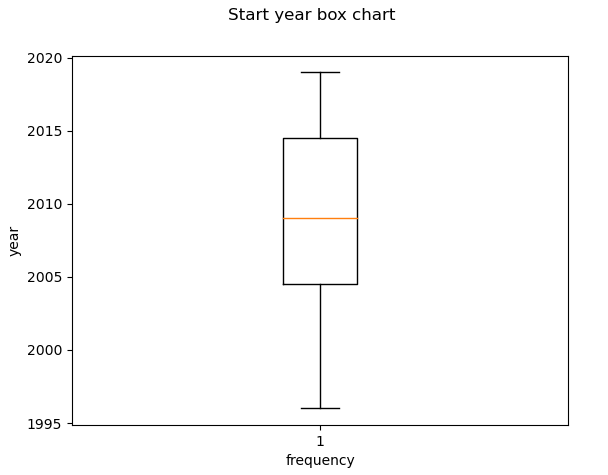
For COVID-19, the most studied drug was Hydroxychloroquine, found in over 50 trials. This is an FDA approved drug for the treatment of malaria and certain autoimmune conditions such as rheumatoid arthritis. Hydroxychloroquine was tested for effectiveness in COVID-19 as a repositioned drug.

No overlap was found between Hepatitis A drugs and COVID-19 drugs, which makes sense due to the difference in the diseases.

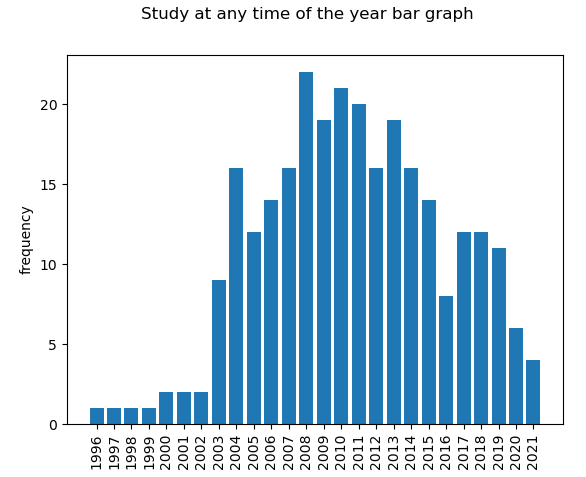
## 3.2.2 Adverse Events

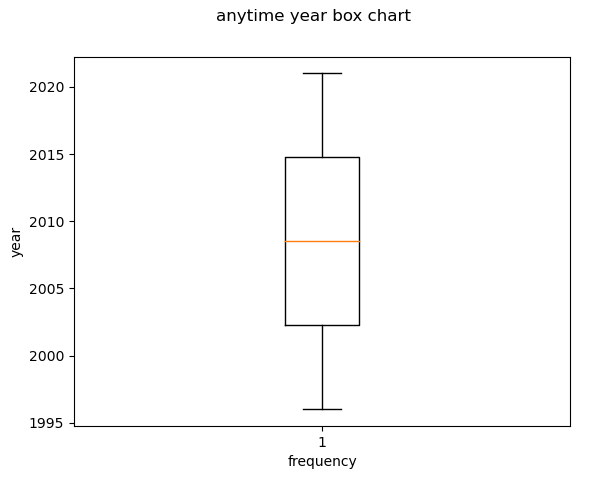
## 3.2.3 Trends over the Years

Following is the chart of when Hep A study started over the years.

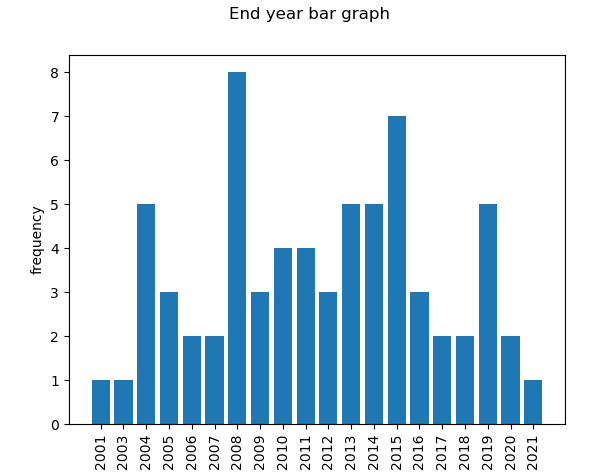


Following shows distribution of the Hep A study start year. Oldest one started in 1990s and the most recent one started in 2010s. Median is 2000s.

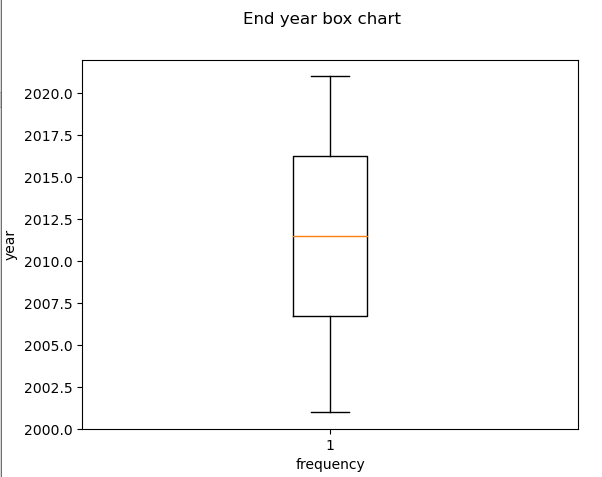
 Following is ongoing Hep A study at any particular year. We can see that Hep A study keeps around 2000s and there are not a lot of study going around left end of the tail of the graph. The downtrend suggests that studies on Hep A are almost on completion.



Following shows distribution of the Hep A study going on at any time of the year. Few are ongoing since a few years ago, but most seem to have been going on slightly before and after 2010s.

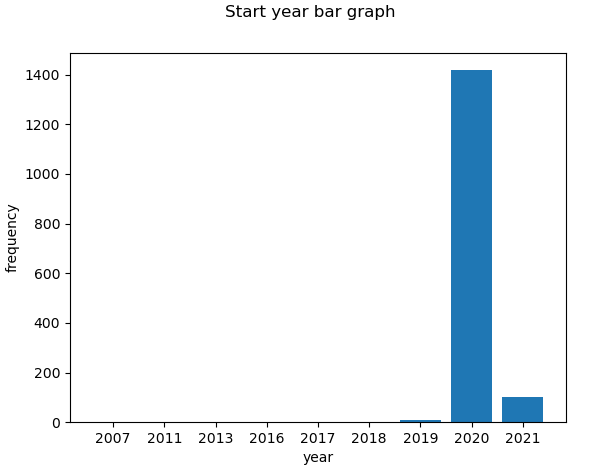


Following is the chart of when Hep A study ended over the years. It seems the conclusion year is concentrated towards the middle of the graph.

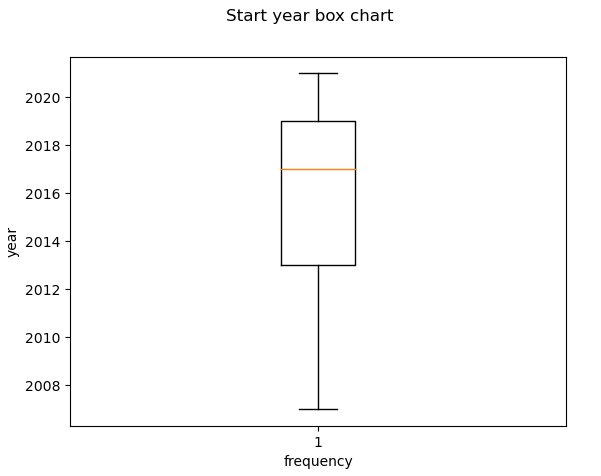


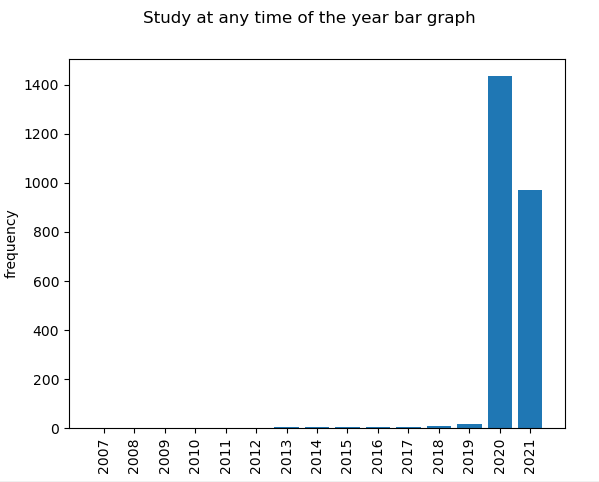
Following shows the distribution of end years for hep A study. Most seem to have ended between 2007 and 2016.

**COVID trends over the years**



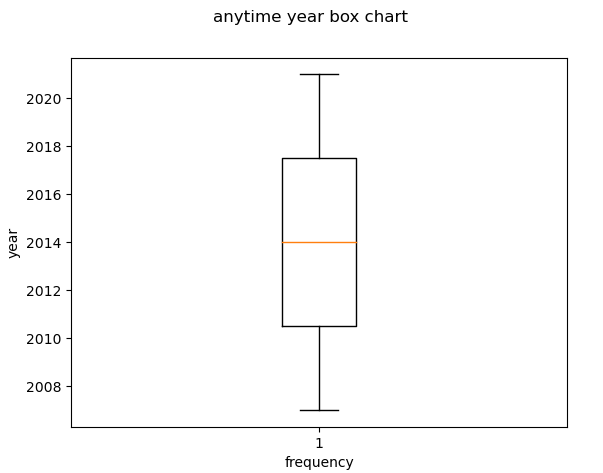
- there are a lot of COVID-19 studies that opened in 2020 because of the pandemic

- median is 2016+s, and it is skewed to more recent years, because of a lot of studies that started in 2020

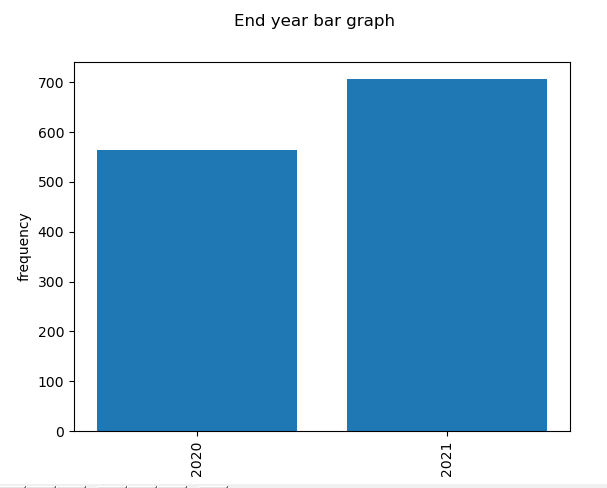


This is a bar graph of COVID study at any time. Most are 2020 and 2021.

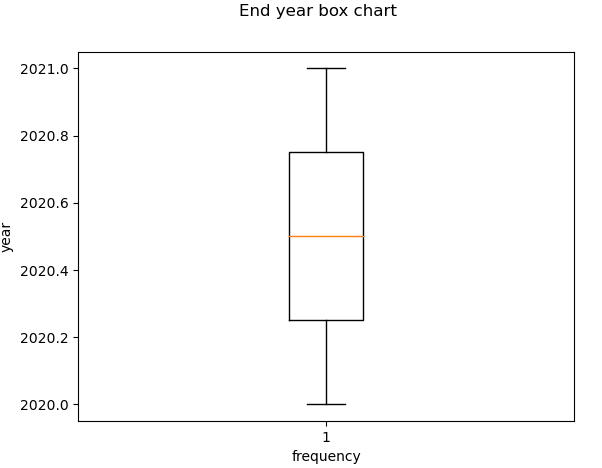
s



Following is the distribution of studies opened on COVID-19. Most are concentrated between 2010 and 2018.



All ended in 2020 and 2021.



The following is a distribution of studies ending year for COVID.

# 3.3 Section C Results:

There are more total number of collaborators for studying COVID than Hep A.

The following graphs are top collaborators for each disease by their frequency.

The following graphs are top lead sponsors for each disease by their frequency.

The following graphs compare the total number of lead sponsor and collaborator in each disease.

# Limitations

Both Hepatitis A and COVID-19 started out with a similar sized dataset, but because trials were included without matching conditions, we had to filter the data. This resulted in a very different number of trials for the two diseases. We had to analyze based on proportion of the whole instead of total numbers to account for this difference. Any analyses thus might not hold true if the datasets were of the same size.

We also filtered based on original search query, so if a trial had the condition ‘Coronavirus infection’ instead of ‘COVID-19’, it was excluded. This could mean that we were missing some relevant trials, however, Coronavirus is an infection that has been around since the mid 1900’s, and COVID-19 is a specific strain of the infection. We were only curious about COVID-19, and our dataset was large enough just filtering by that, so we are confident that we had enough data for COVID to make a proper analysis.

# Conclusions

COVID-19 and Hepatitis A are two very different conditions, the former is a new strain of a contagious respiratory infection, while the latter is a long-established condition affecting the liver. Despite these differences, the general study design of the two diseases was fairly similar. As both diseases are contagious and vaccine-preventable, it seems that this connector between the two resulted in similar trial designs. Proportionately, the trials were similar in masking and allocation strategy, interventional and observational models, type of trial, and gender.

The trials do differ in many sections, but most of these differences can be explained by the data obtained in part 3.3. A new disease strain that is being studied will have different phase data and result availability than a disease that has been studied for decades. COVID-19 studies exploded in 2020, and thus its result availability and higher phase trial count is much less than for Hepatitis A. Also, eligibility criteria, drugs, and adverse events are different. The diseases affect different parts of the body and different groups of people, so the drugs created and the primary population differs.

The general study design and strategy is similar for two very different diseases because of their similarity in being contagious and vaccine preventable. These two features of COVID-19 and Hepatitis A caused many similarities among the two trial datasets. The remaining differences can be explained by the newness of COVID-19 and how they differ in affecting people.

# Bibliography

1. “Glossary of Common Site Terms,” *ClinicalTrials.gov*. [Online]. Available: https://clinicaltrials.gov/ct2/about-studies/glossary. [Accessed: 26-Feb-2021].
2. “Havrix™ 720 Junior, one dose vial,” *World Health Organization*, 20-Mar-2015. [Online]. Available: https://www.who.int/immunization\_standards/vaccine\_quality/pq\_267\_hepa\_1dose\_gsk/en/. [Accessed: 27-Feb-2021].
3. “json - JSON encoder and decoder¶,” *json - JSON encoder and decoder - Python 3.9.2 documentation*. [Online]. Available: https://docs.python.org/3/library/json.html. [Accessed: 26-Feb-2021].
4. Martinblech, “xmltodict,” *GitHub*. [Online]. Available: https://github.com/martinblech/xmltodict. [Accessed: 26-Feb-2021].
5. “National Comprehensive Cancer Network,” *Phases of clinical trials*. [Online]. Available: https://www.nccn.org/patients/resources/clinical\_trials/phases.aspx. [Accessed: 26-Feb-2021].
6. “Our COVID-19 Rapid Response Program,” *Regeneron's COVID-19 Response Efforts*. [Online]. Available: https://www.regeneron.com/covid19. [Accessed: 26-Feb-2021].
7. Python, “python/cpython,” *GitHub*, 07-Apr-2020. [Online]. Available: https://github.com/python/cpython/blob/3.9/Lib/os.py. [Accessed: 26-Feb-2021].
8. “Welcome to GSK Clinical Trials,” *GSK Study Register*. [Online]. Available: https://www.gsk-studyregister.com/en/. [Accessed: 26-Feb-2021].