COVID-19 Data Science Analysis

Mhealyssah Bustria & Anjelina Velazquez
Mentors: Dr. Niema Moshiri & Dr. Youwen Ouyang

UCSD DBMI Summer Internship 2020
August 12, 2020

This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

About Us

Mhealyssah (Mhea) Bustria

A Computer Science undergraduate at CSUSM. She is always happy to give back to her community, and she is especially-interested in advancing the fields of education and health.

Anjelina Velazquez

Currently a fourth year Computer Science student at CSUSM. Enjoys keeping busy by always learning new concepts and ideas. She is always encouraging others to do the same.

This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

Outline

BACKGROUND

- Motivation
- Our Project
- Methods

RESULTS

- Reading patient data
- Analyzing the dataset

CONCLUSIONS

- Lessons learned
- Next steps

This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

Background - Motivation

To combat the **COVID-19 pandemic**, researchers need to make inferences from data collected from patients.

This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

Background - Motivation

To combat the **COVID-19 pandemic**, researchers need to make inferences from data collected from patients.

Are these inferences generalizable?

This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

Background - Motivation

To combat the **COVID-19 pandemic**, researchers need to make inferences from data collected from patients.

Are these inferences generalizable?

Researchers need to account for...

potential errors and inaccuracies that are present in the data collected due to manual input

potential confounding factors, such as biases in patient sampling

This research was supported by the NLM training grant T15LM011271, the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

Background - Our project

Collection of over 75,000 SARS-CoV-2 Patient Records

Manually-entered data

Errors, inconsistencies, and missing data affects the data science analysis process.

Large dataset

What do the patient records in our dataset look like?

What are some potential confounding factors?

This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

Research Goals

Reading patient data

Demonstrate how manually-entered data affects data science analysis.

Analysis and visualization

Show how demographic information varies in our large sample.

This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

Methods

This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

Methods

Obtaining the dataset

Global initiative on sharing all influenza data (GISAID)

- Extracted records were stored in a gzipped JSON file

This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

Methods

Obtaining the dataset

Global initiative on sharing all influenza data (GISAID)

- Extracted records were stored in a gzipped JSON file

Data reading and analysis

Python

- Identify **what information can be found** in the records
- Decide **what information to use** for analysis and visualization

This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

Methods

Obtaining the dataset

Global initiative on sharing all influenza data (GISAID)

- Extracted records were stored in a gzipped JSON file

Data reading and analysis

Python

- Identify **what information can be found** in the records
- Decide **what information to use** for analysis and visualization

Data visualization

matplotlib Python library

- pie charts

seaborn Python library (based on matplotlib)

- bar plots
- count plots
- violin plots

This research was supported by the NLM training grant T15LM011271, the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

RESULTS

Reading Patient Data

This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

Results - Reading patient data

How were our analysis and visualization processes affected by manual data-entry practices?

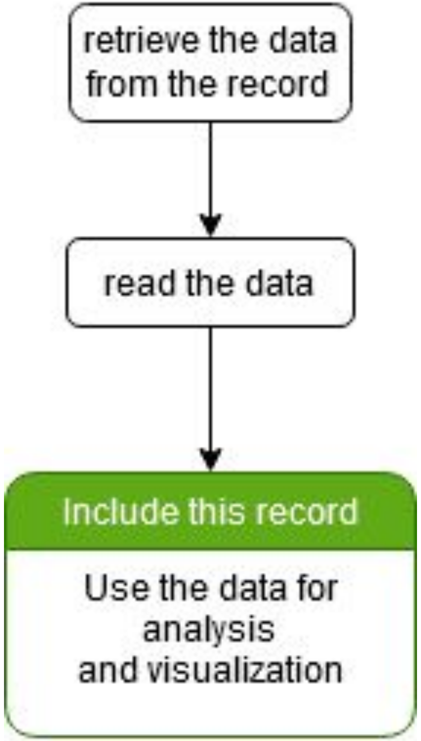
How did we modify our data-reading strategies to account for unusable data?

- data correction
- data exclusion

What are potential solutions to address the issues caused by manual data entry?

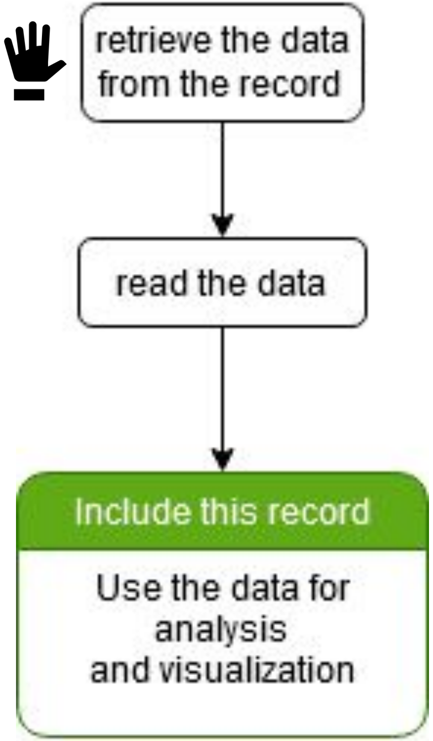
This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

Original strategy for reading the data



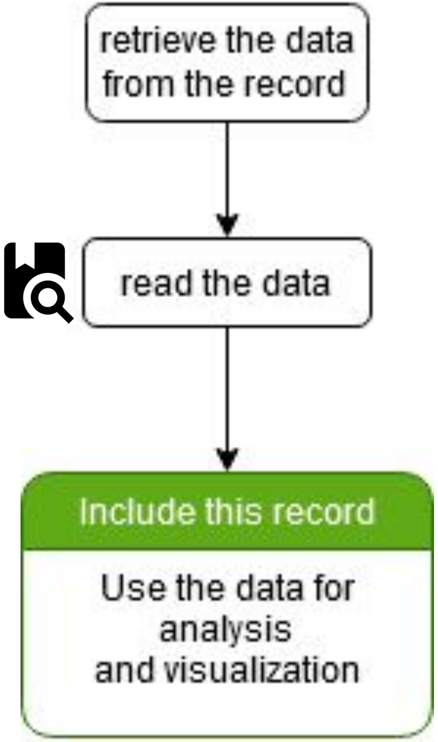
This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

Original strategy for reading the data



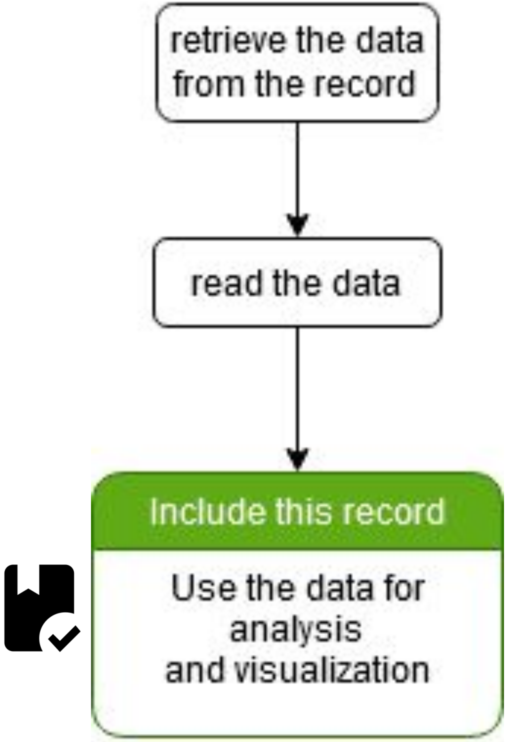
This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

Original strategy for reading the data



This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

Original strategy for reading the data



This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

Types of missing / unknown / invalid data

Type 1

Data for the attribute-of-interest was **not provided**.

This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

?Types of missing / unknown / invalid data

Type 1

Data for the attribute-of-interest was **not provided**.

Type 2

Data for the attribute-of-interest was provided, but was entered as some variation of **"unknown" or "not applicable"**.

This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

?Types of missing / unknown / invalid data

Type 1

Data for the attribute-of-interest was **not provided**.

Type 2

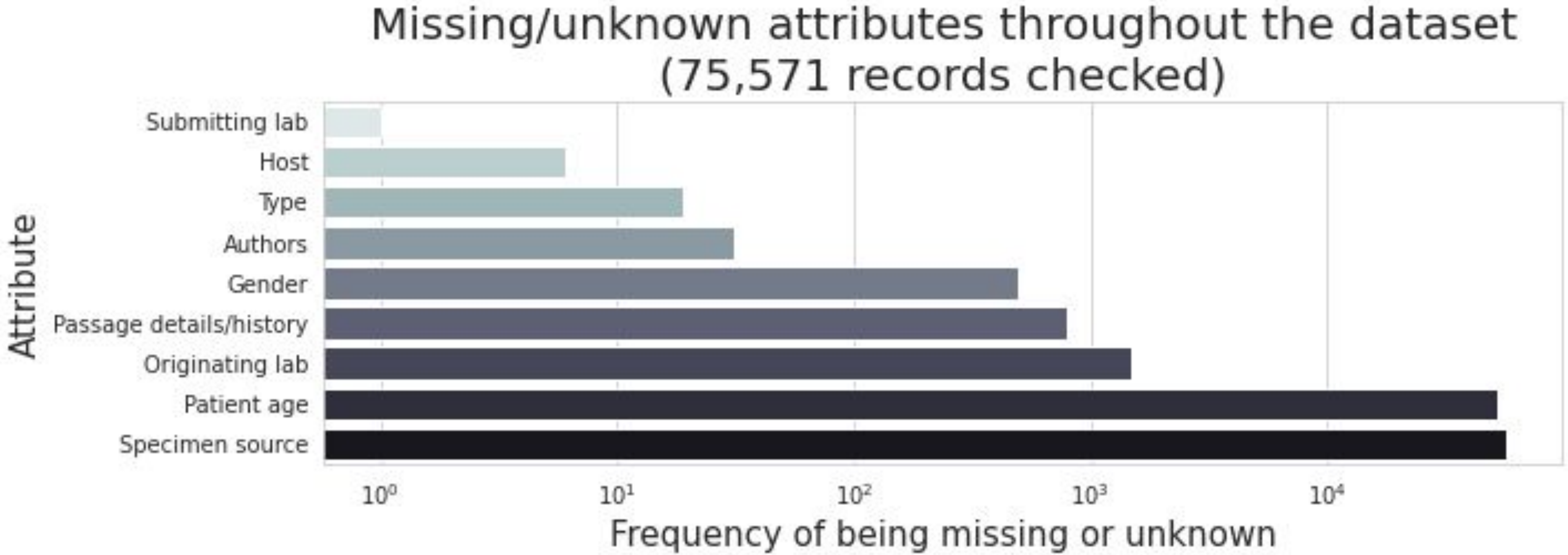
Data for the attribute-of-interest was provided, but was entered as some variation of **"unknown" or "not applicable"**.

Type 3

Data for the attribute-of-interest contained an error such as **formatting inconsistencies or misspellings**.

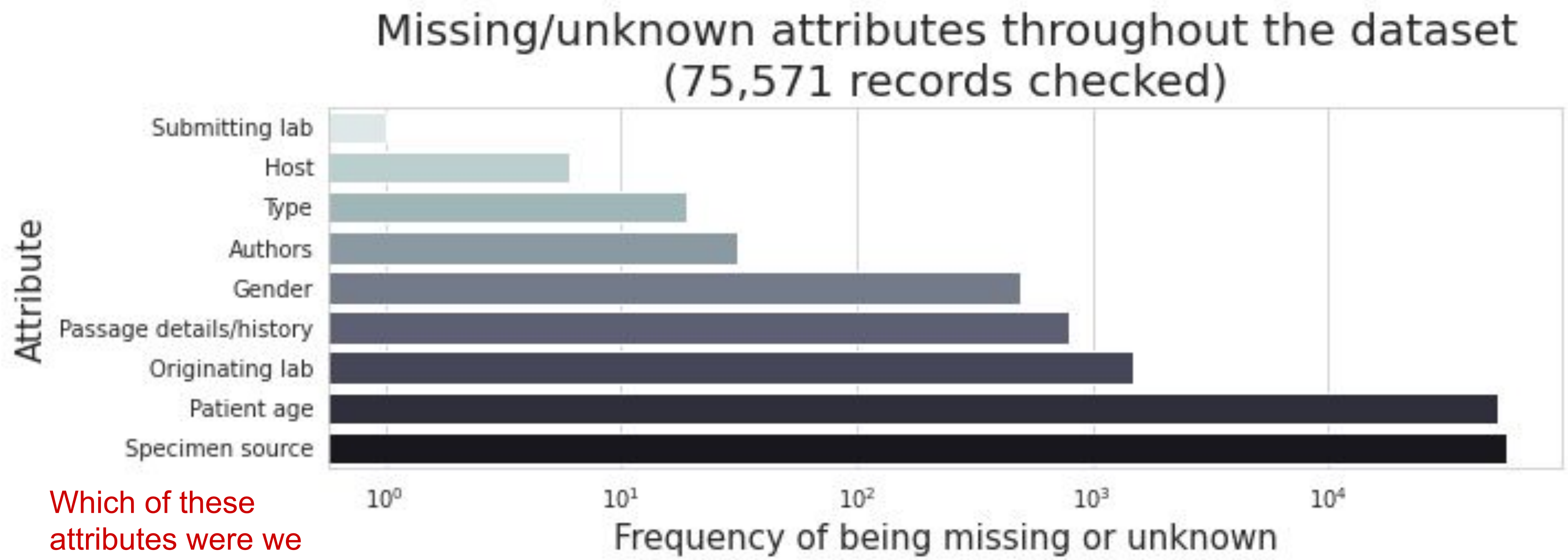
This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

?How often was attribute data missing or invalid/unknown?



This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

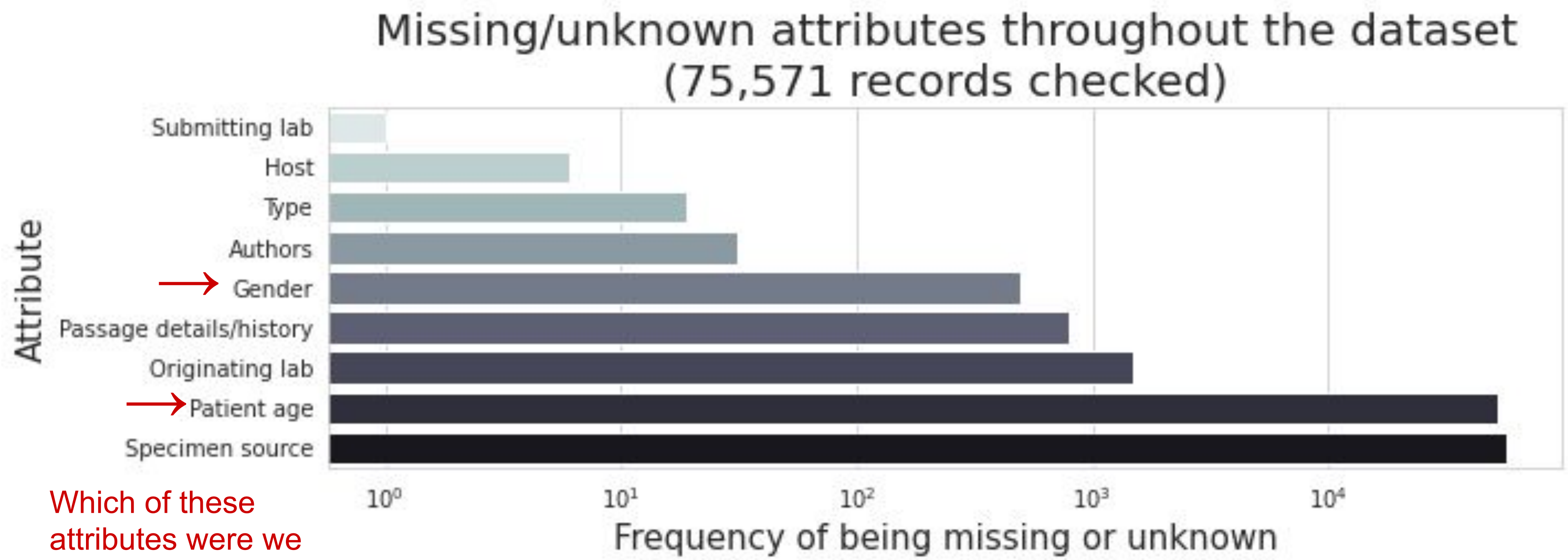
?How often was attribute data missing or invalid/unknown?



Which of these attributes were we most interested in?

This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

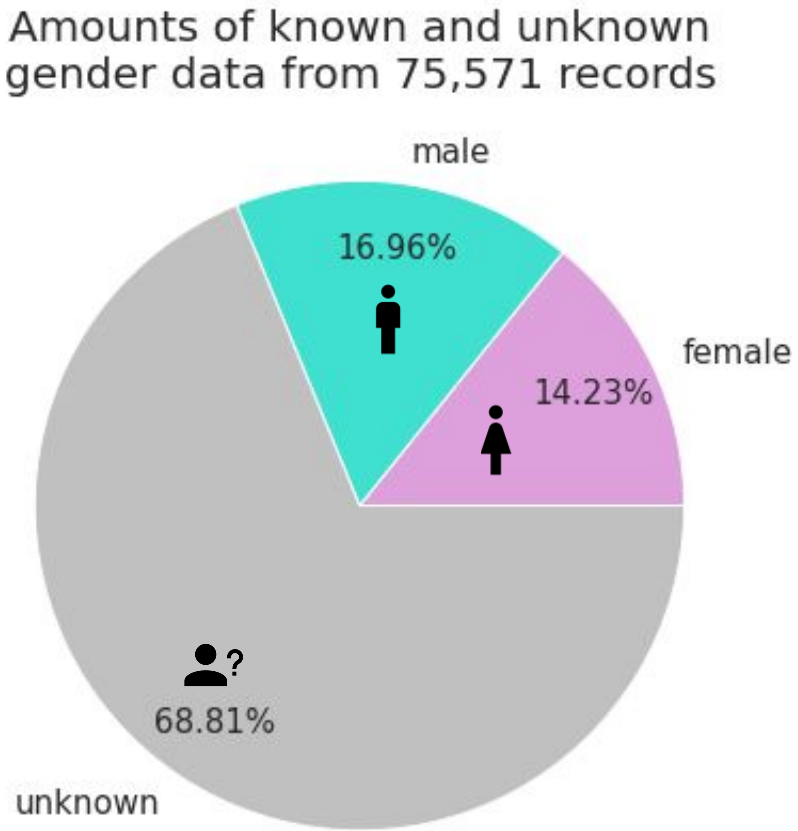
?How often was attribute data missing or invalid/unknown?



Which of these attributes were we most interested in?

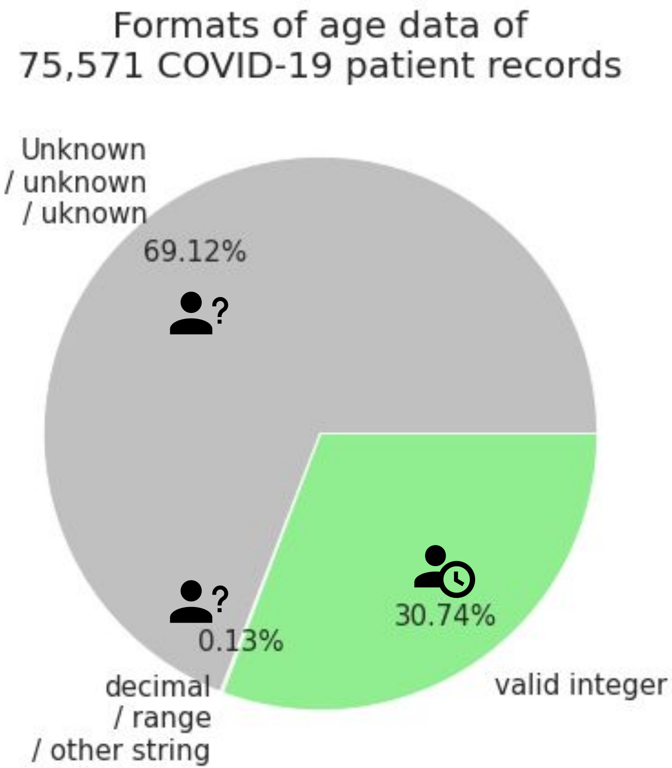
This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

How often did we encounter unknown gender data?



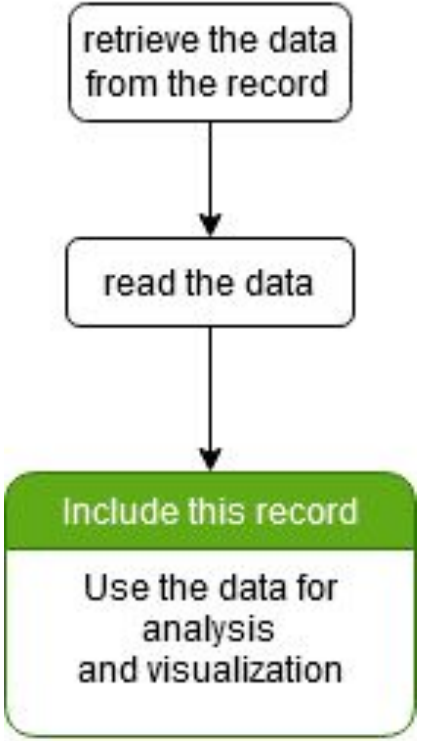
This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

How often did we encounter unknown or unusable age data?



This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

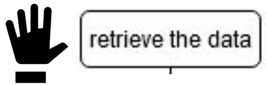
Original strategy for reading the data



Throughout the study, how did we revise our strategies to handle the issues involved with manually-entered data?

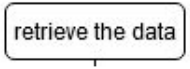
This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

Handling manually-entered data



This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

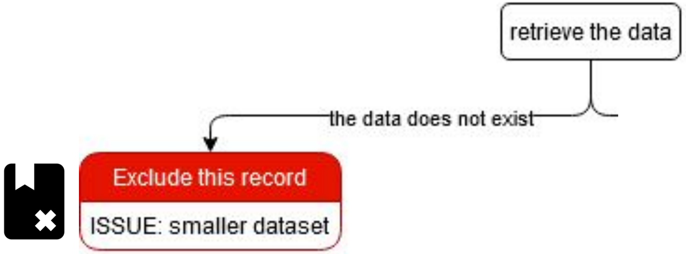
Handling manually-entered data



the data may or may not
be missing

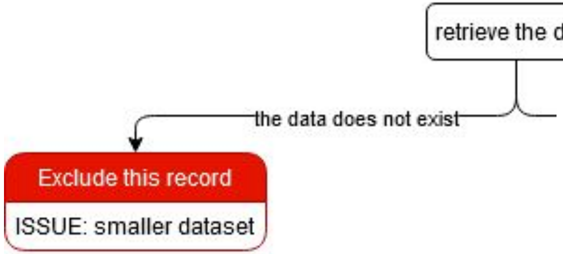
This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

Handling manually-entered data



This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

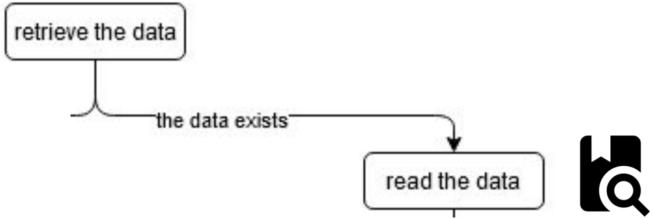
Handling manually-entered data - example of data exclusion



If age is an attribute of interest but there is no age provided in that record → **data exclusion**

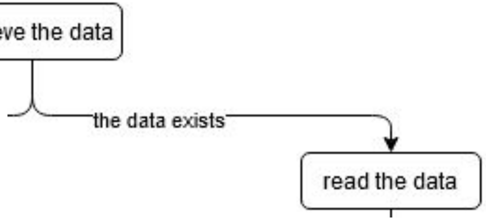
This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

Handling manually-entered data



This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

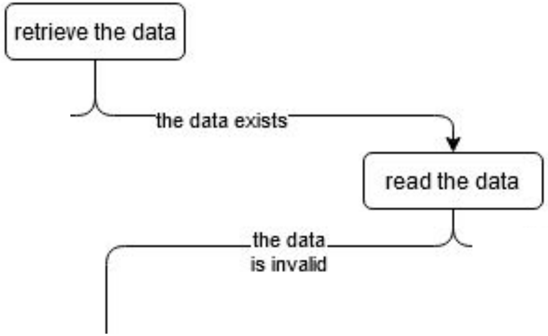
Handling manually-entered data



a data-entry error
may or may not
be discovered

This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

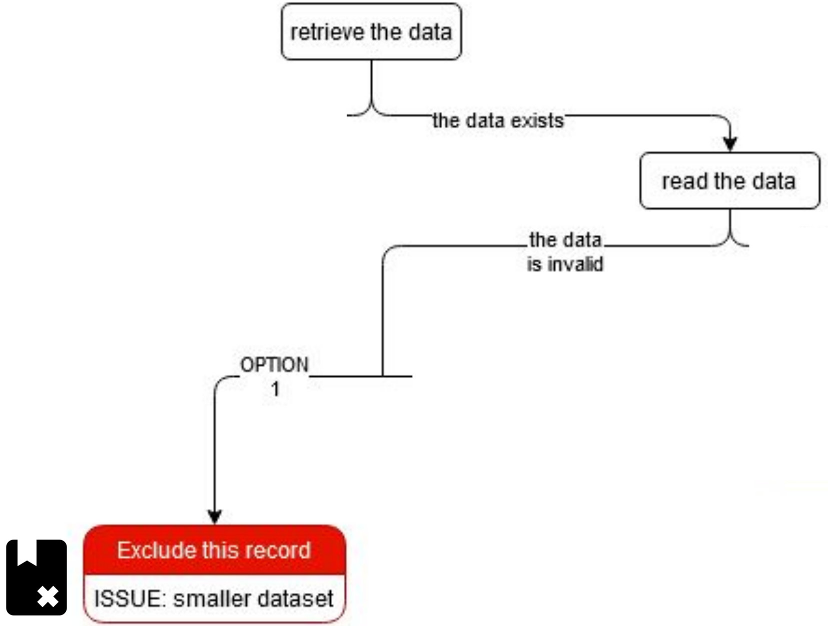
Handling manually-entered data



OPTIONS:
data-exclusion
or
data-correction

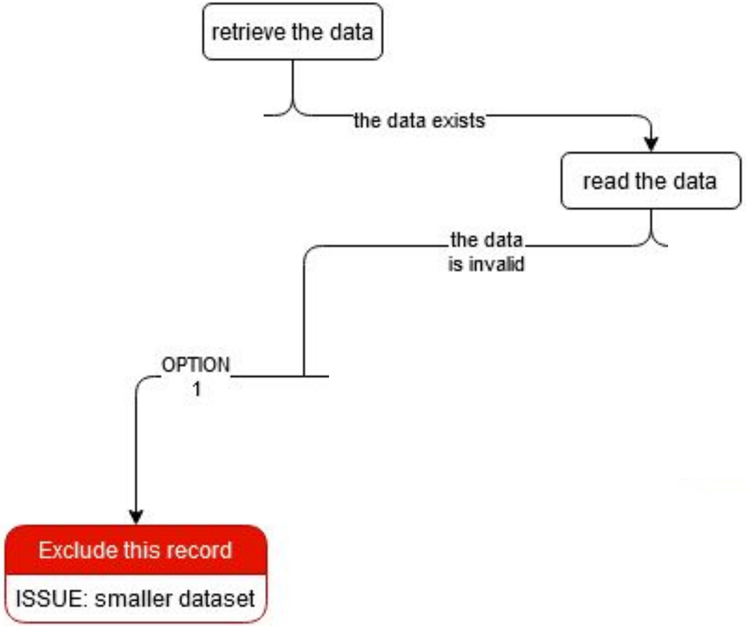
This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

Handling manually-entered data



This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

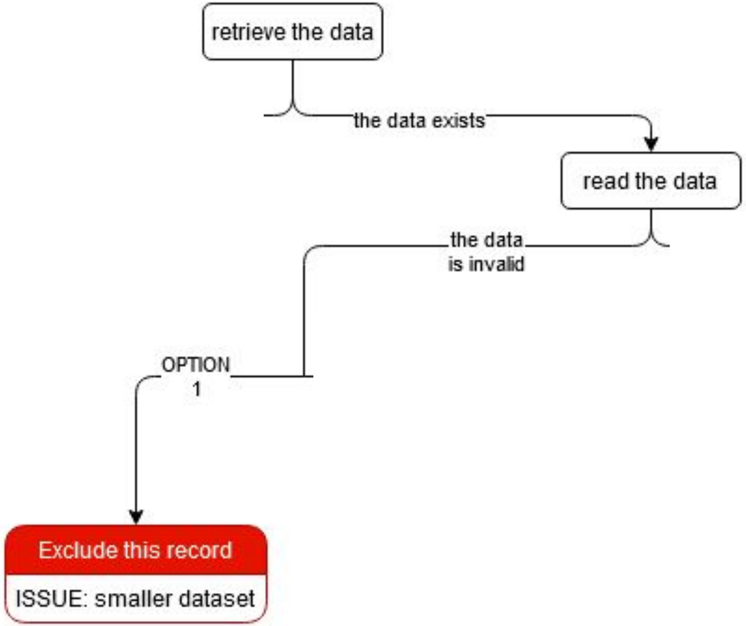
Handling manually-entered data - example of data exclusion



"North America / USA / LA"
could refer to Los Angeles or Louisiana. It is not specified whether the third piece is the state or the city.
ambiguous → exclude

This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

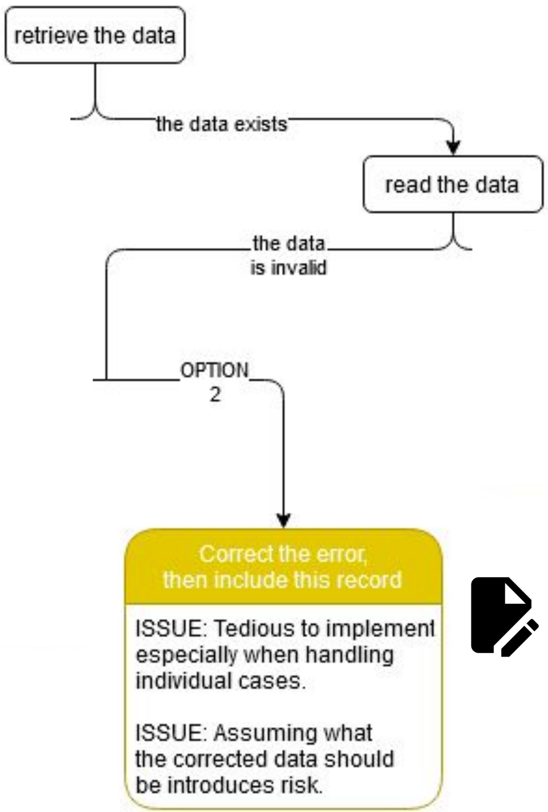
Handling manually-entered data - example of data exclusion



Age data is present in the record, but a value such as "unknown" or "not applicable" was provided.
→ **exclude**

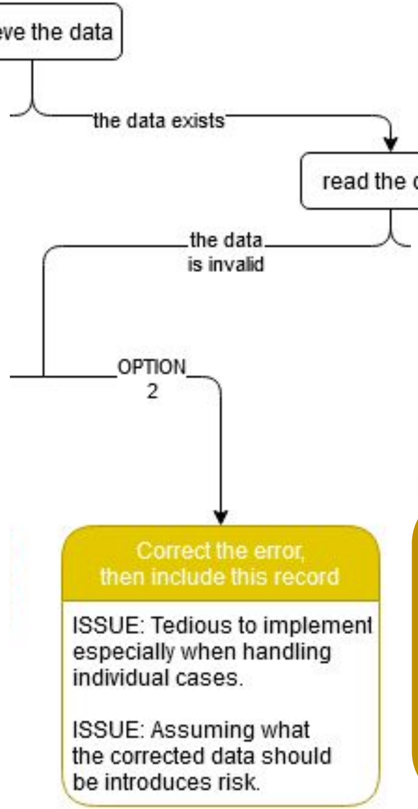
This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

Handling manually-entered data



This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

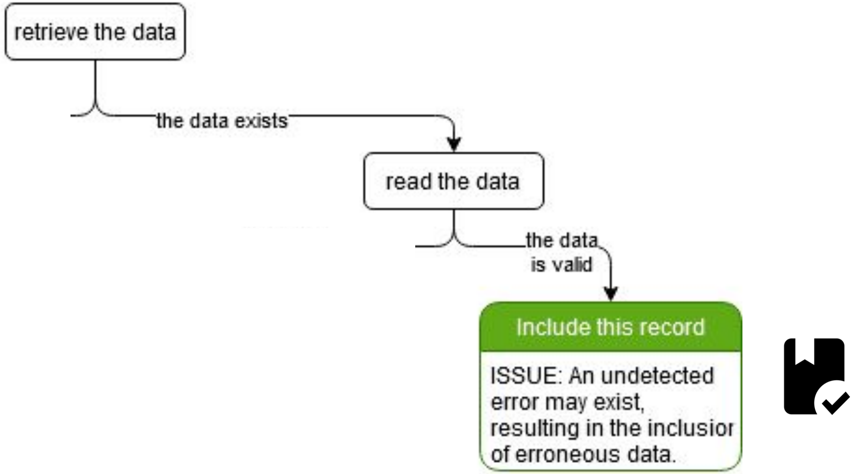
Handling manually-entered data - example of data correction



Location is provided as **"Oceania / Australia / NSW" → assume "NSW" means "New South Wales", correct the value and include the data**

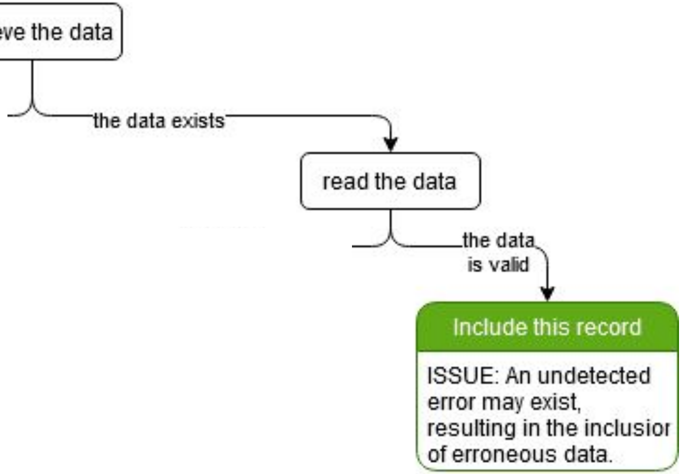
This research was supported by the NLM training grant T15LM011271, the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

Handling manually-entered data



This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

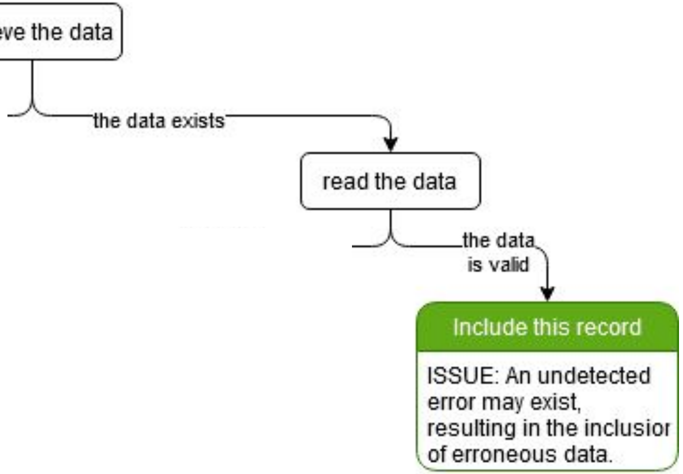
Handling manually-entered data - example of data inclusion



"Asia / China / Wuhan"
Wuhan gets categorized as a state in China, but is really a city in Hubei, China.
→ incorrect data was included due to unclear formatting

This research was supported by the NLM training grant T15LM011271, the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

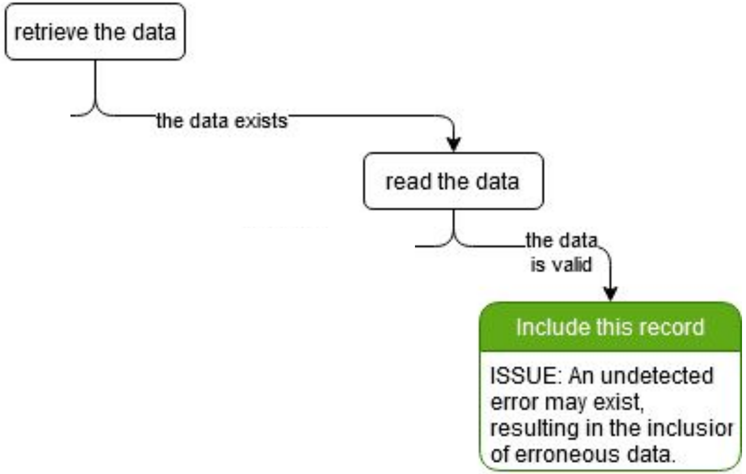
Handling manually-entered data - example of data inclusion



"Washington DC" and "District of Columbia" are stored as two separate categories.
→ **dataset becomes misleading due to inconsistent naming conventions**

This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

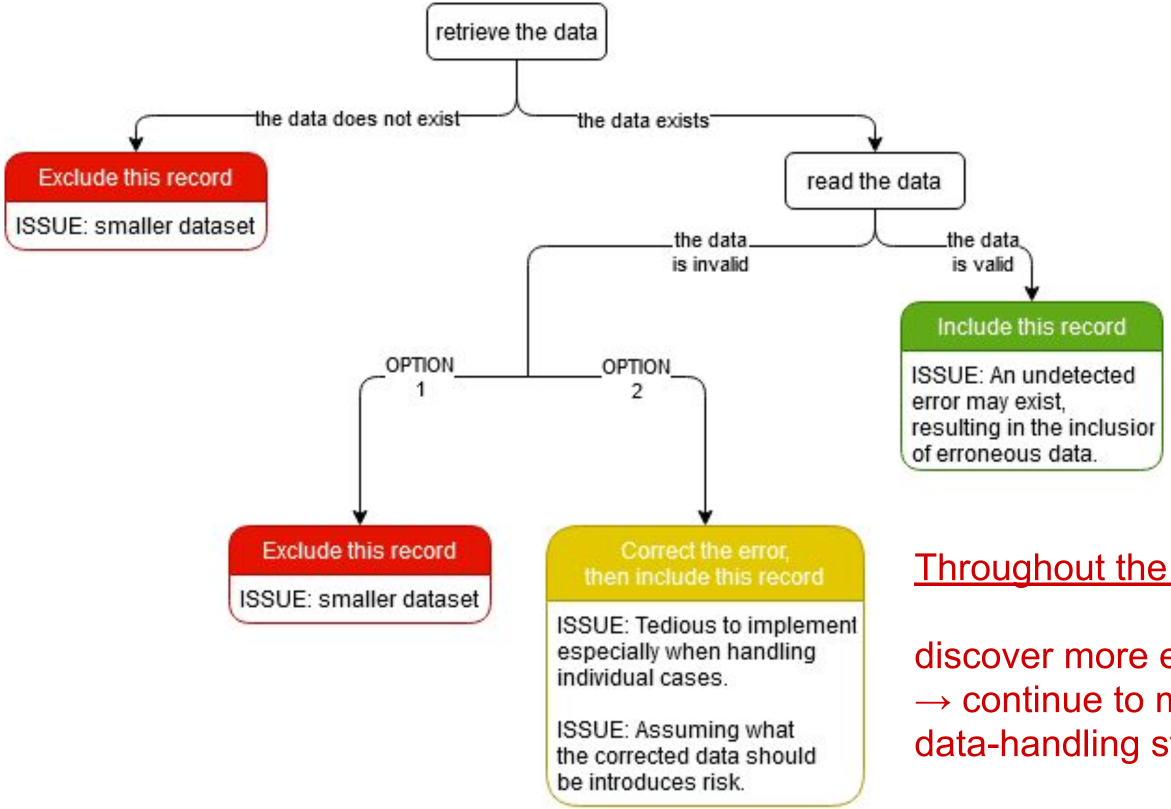
Handling manually-entered data - example of data inclusion



"Oceania / Australia /
New South Wales / Sydney"
Sydney gets categorized into both "Australia" and "Wales".
→ **incorrect data was included due to the categorization methods used**

This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

Handling manually-entered data



<u>Throughout the study</u>

discover more error cases
→ continue to modify the data-handling strategies

This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

Potential solutions for improving data-entry practices

Improving Manual Data Entry

- The use of dropdown menus instead of allowing the user to input information
- Adding Data Validation will ensure that all information being added is formatted correctly

Automated Data Capture instead of Manual Data Entry.

Benefits of automation:
- Significantly reduces errors
- Improved efficiency and cleaner data

This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

RESULTS

Analyzing the dataset

This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

Results - Analyzing the dataset

What did our collection of 75,571 patient records look like?

Attributes that we looked at

Location
Where the data was submitted from

→ Continent
→ Country
→ State

Gender

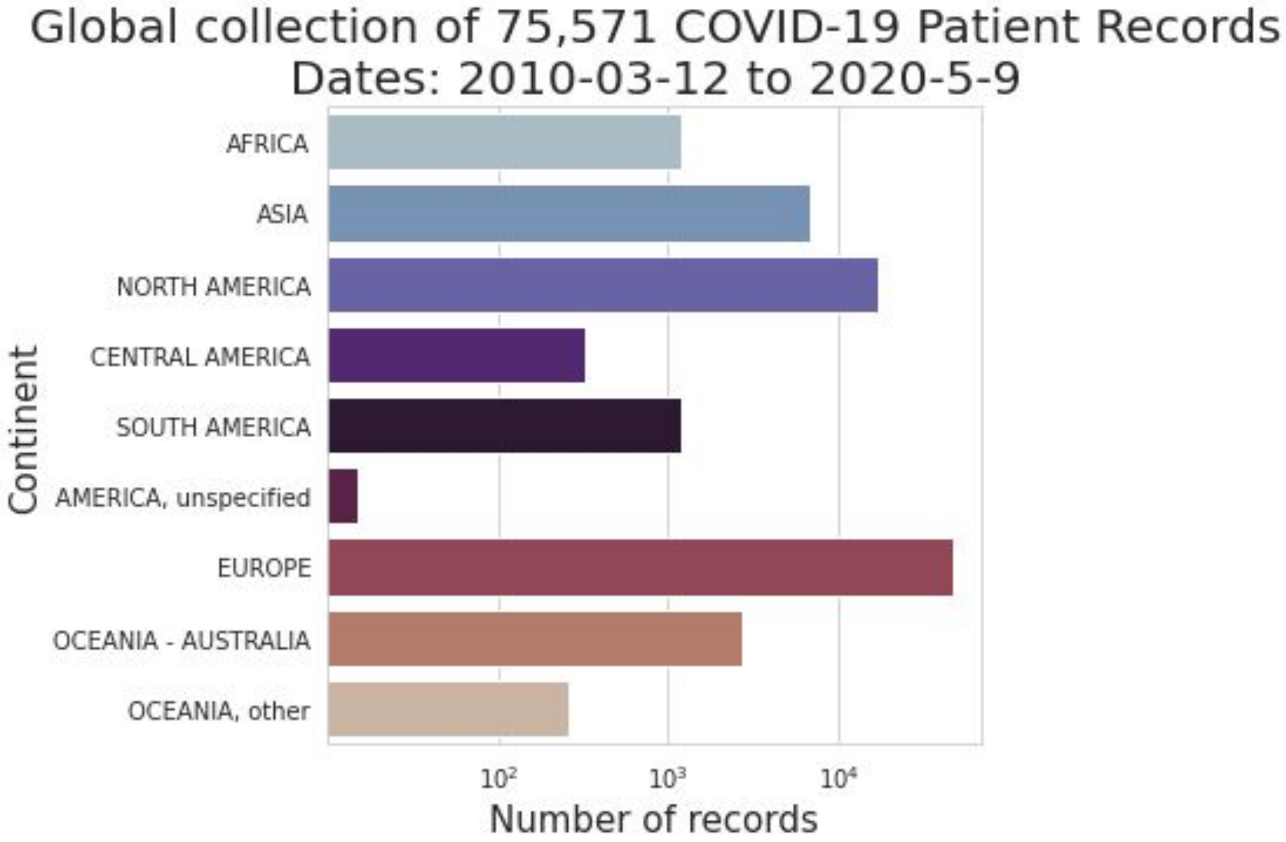
→ The gender data that was provided

Age

→ The age in years of the patient

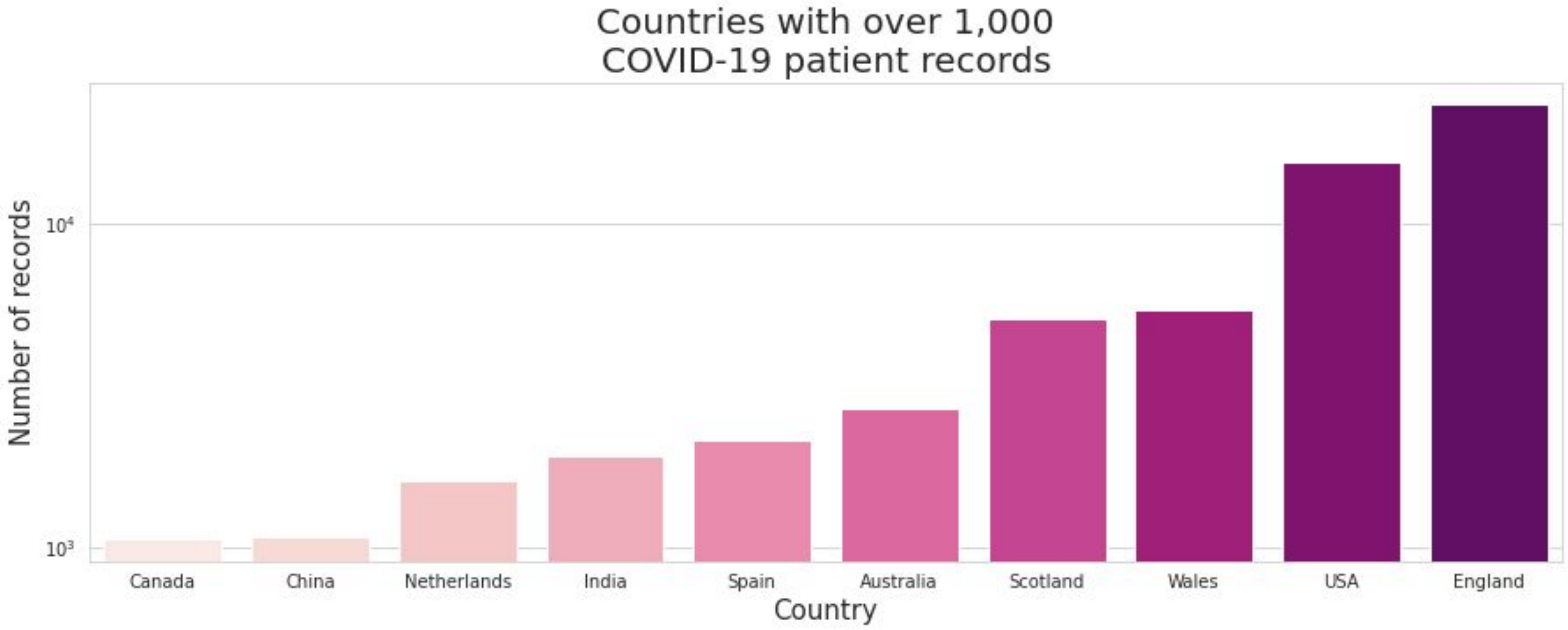
This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

What continents were the records submitted from?



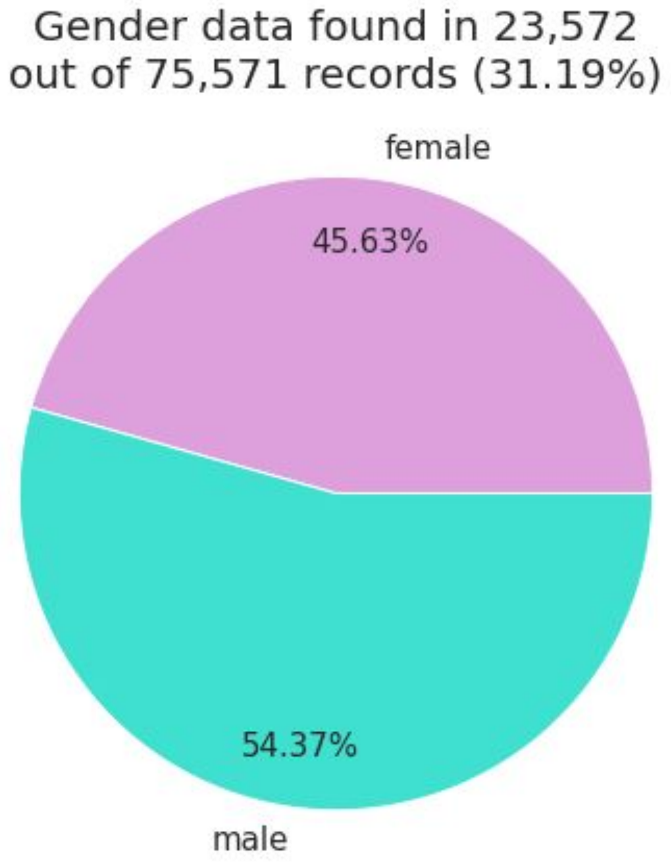
This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

Which countries have submitted over 1,000 records?



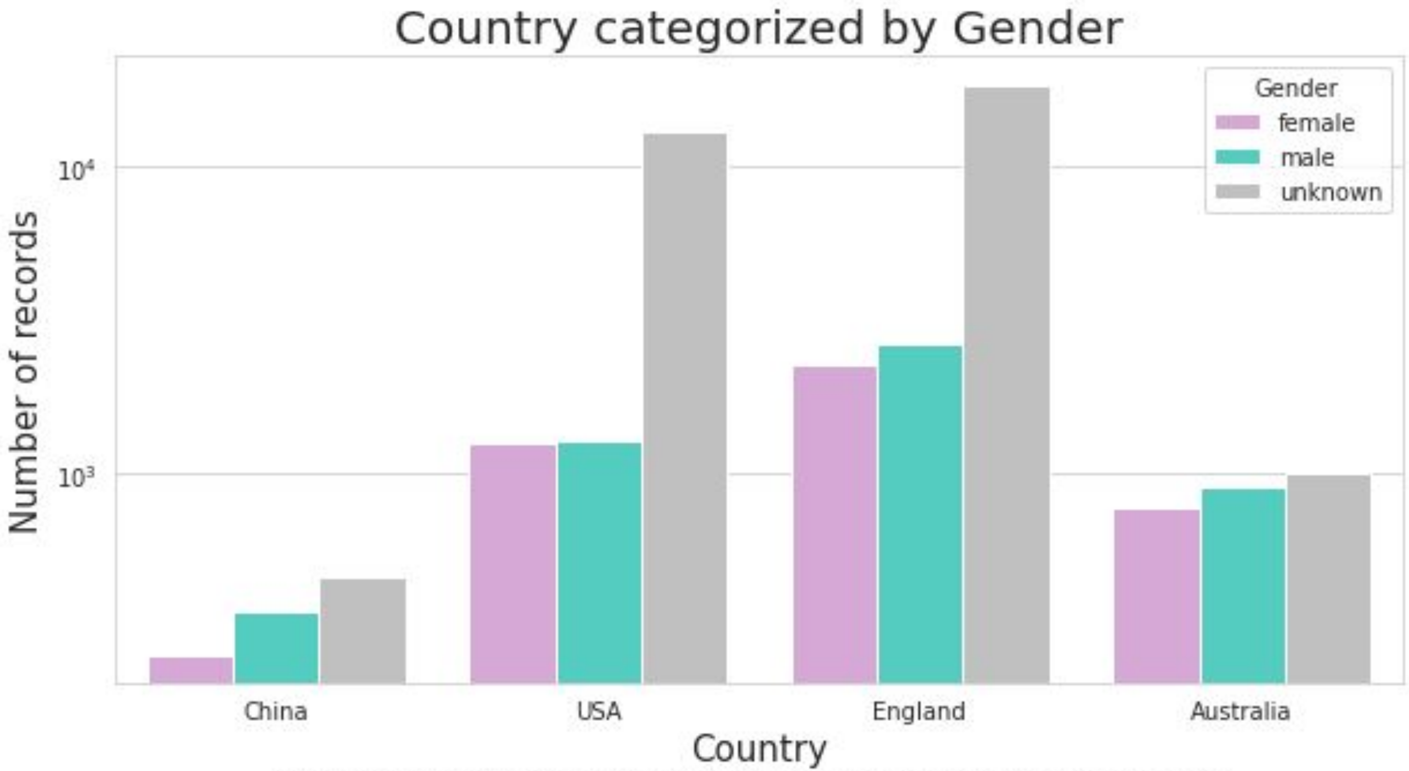
This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

What genders were found in the dataset?



This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

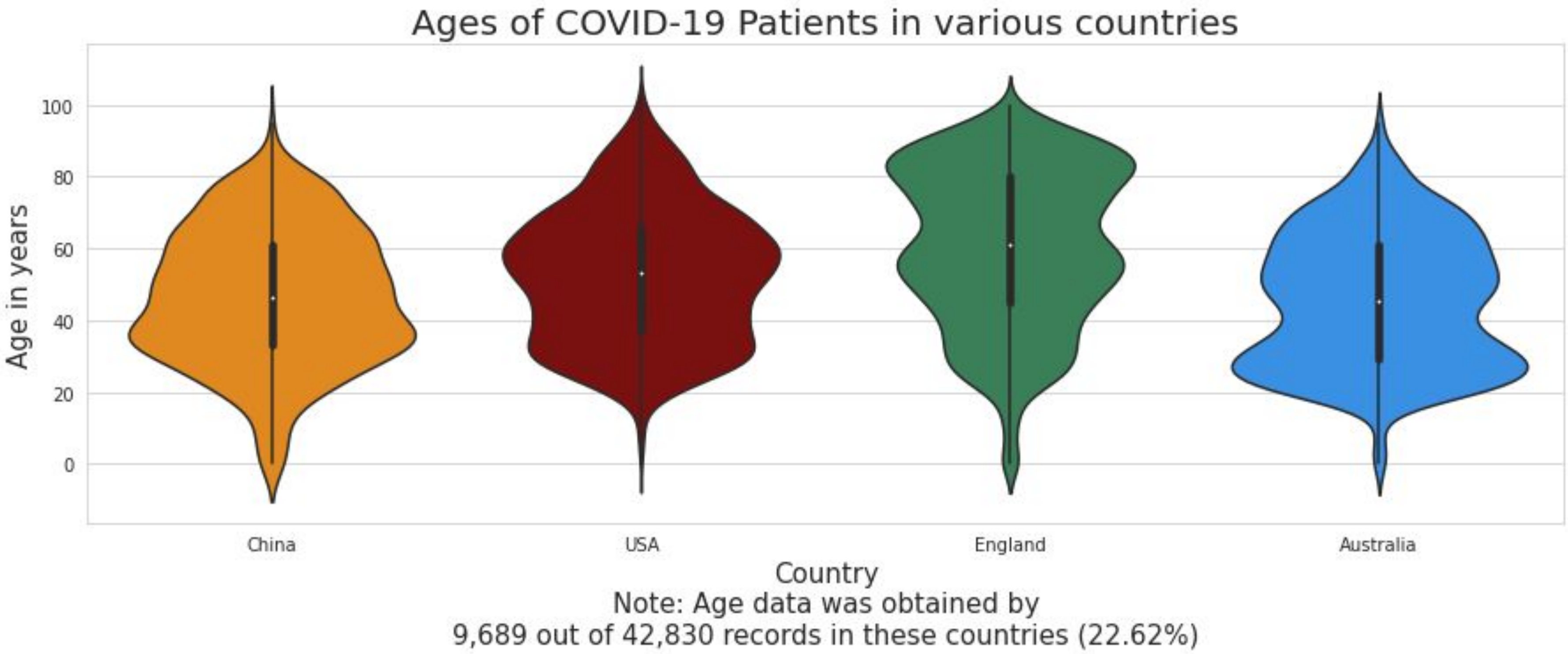
What was the gender data in various countries?



Out of the 42,615 patient records from these countries,
55.0% held either female of male gender data.

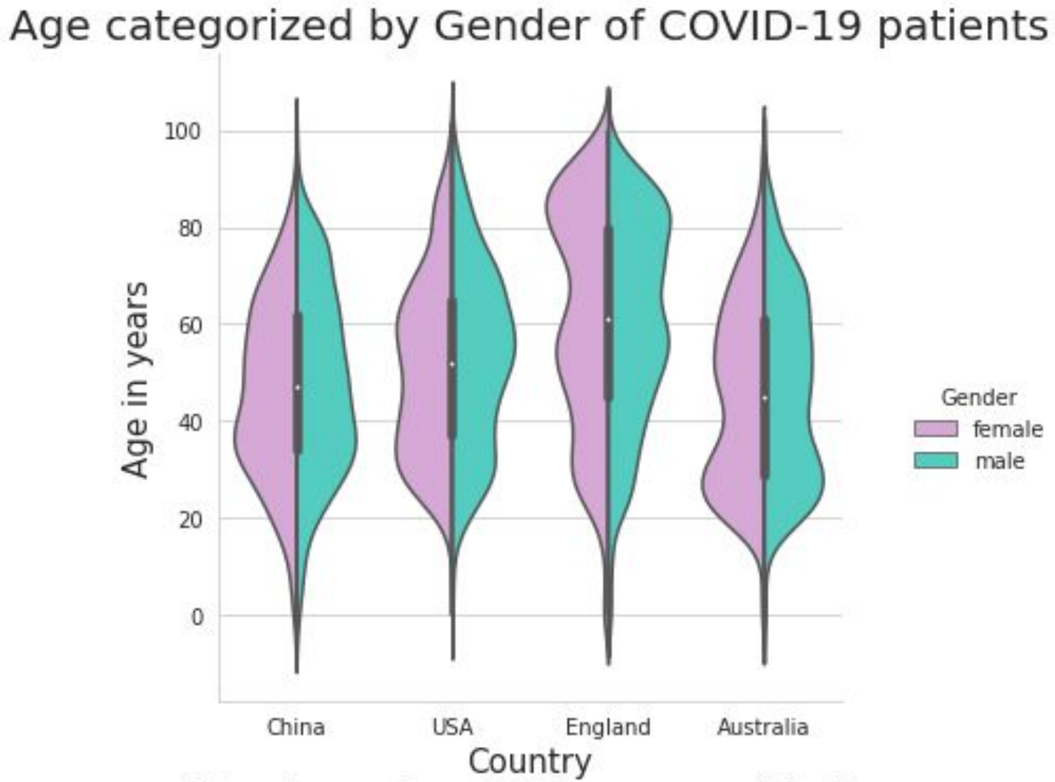
This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

What was the age data in various countries?



This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

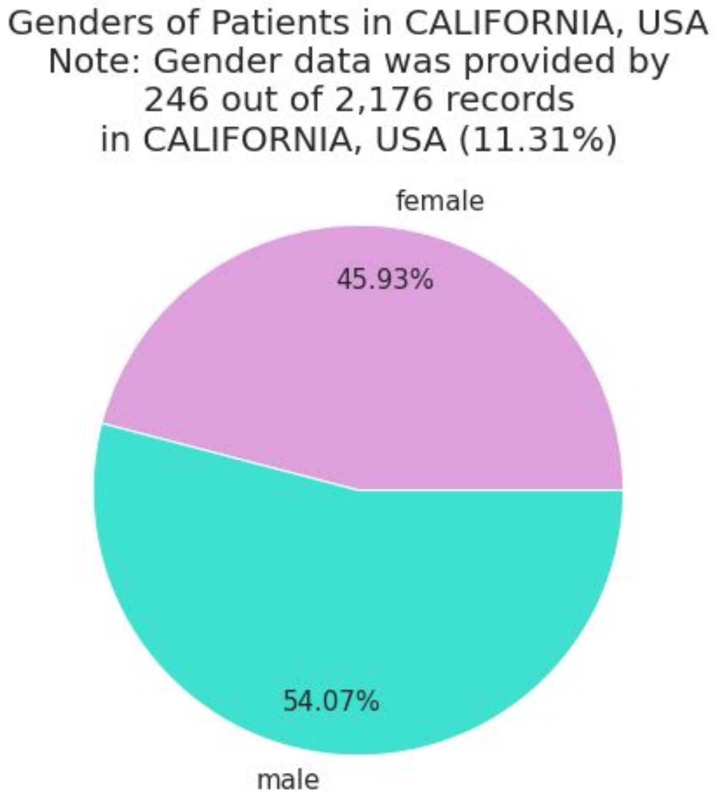
What were the combinations of age and gender?



Note: Age and gender data was provided by
9,457 out of 42,830 records in these countries (22.08%)

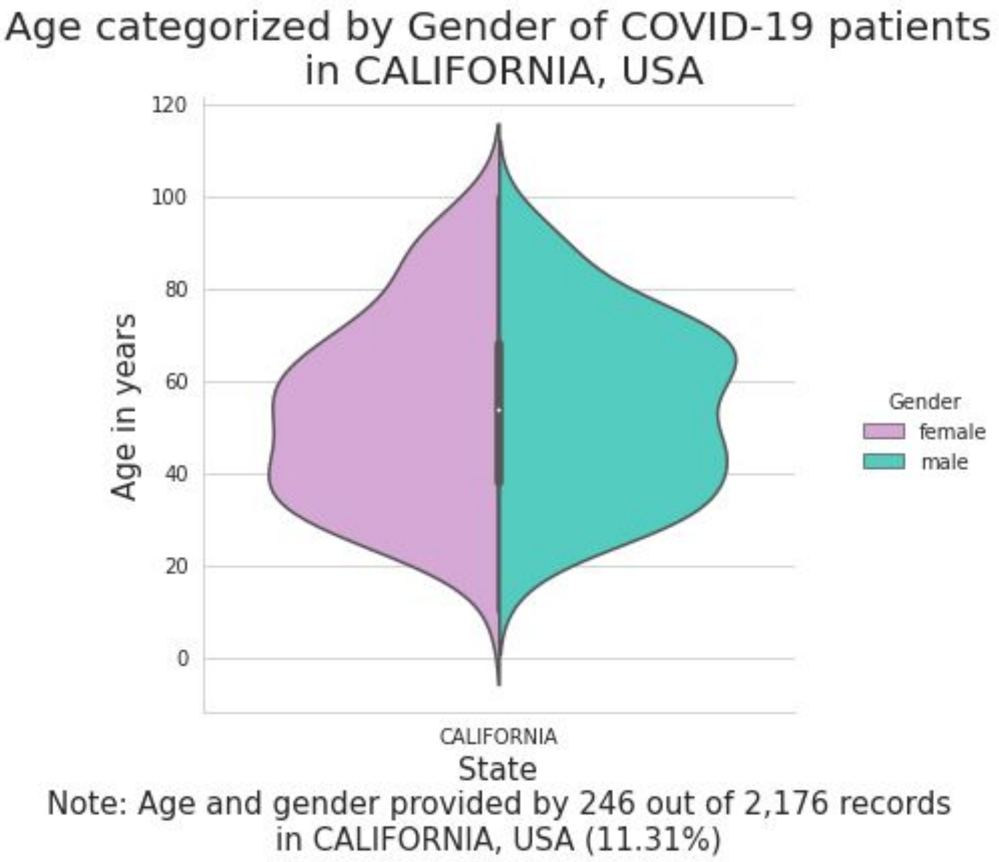
This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

What genders were found in California?



This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

What ages and genders were found in California?



This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

Lessons Learned - Biomedical Research

Reading patient data

Data analysis is affected by the data-entry methods that were used to create the data.

Confounding errors such as the lack of correct data entry can cause misinterpretation.

Analysis and visualization

In order to generalize inferences about a sample of patients, the differences in the patients should be accounted for instead of treating each patient as if they are the same person.

This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

Lessons Learned - Data Science

Reading patient data

Loading and reading a collection of data from a file.

Identifying what information should be included in analysis and visualization.

Analysis and visualization

Choosing the appropriate visualization types to show trends in data.

Using tools to generate the visualizations.

This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

Next Steps

Improve data-handling strategies in our code

Current focus: error-handling for location data

Example: Use pre-defined lists of all states in a country to prevent misrepresentation

Improve **efficiency**, **readability**, and **reusability** of our code so that our methods can be used to assist with the study other datasets

Advocate for **improvement in data-entry practices**

This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.

Acknowledgements

DBMI Summer 2020 Internship Program Leaders

Nancy Herbst

Dr. Jejo Koola

Dr. Tsung-Ting (Tim) Kuo

Dr. Lucila Ohno-Machado

Elizabeth Santillanez

Project Mentors

Dr. Niema Moshiri | Dr. Youwen Ouyang

Research Funding

The National Library of Medicine (NLM) training grant number T15LM011271

The National Science Foundation RAPID grant number NSF-2028040

Google Cloud Platform (GCP) Research Credits Program

Thank you for your time!

Any questions are welcome.

Mhealyssah Bustria
bustr003@cougars.csusm.edu

Anjelina Velazquez
velaz035@cougars.csusm.edu

This research was supported by the NLM training grant T15LM011271,
the NSF RAPID grant NSF-2028040, and the GCP Research Credits Program.