Data Visualization Project

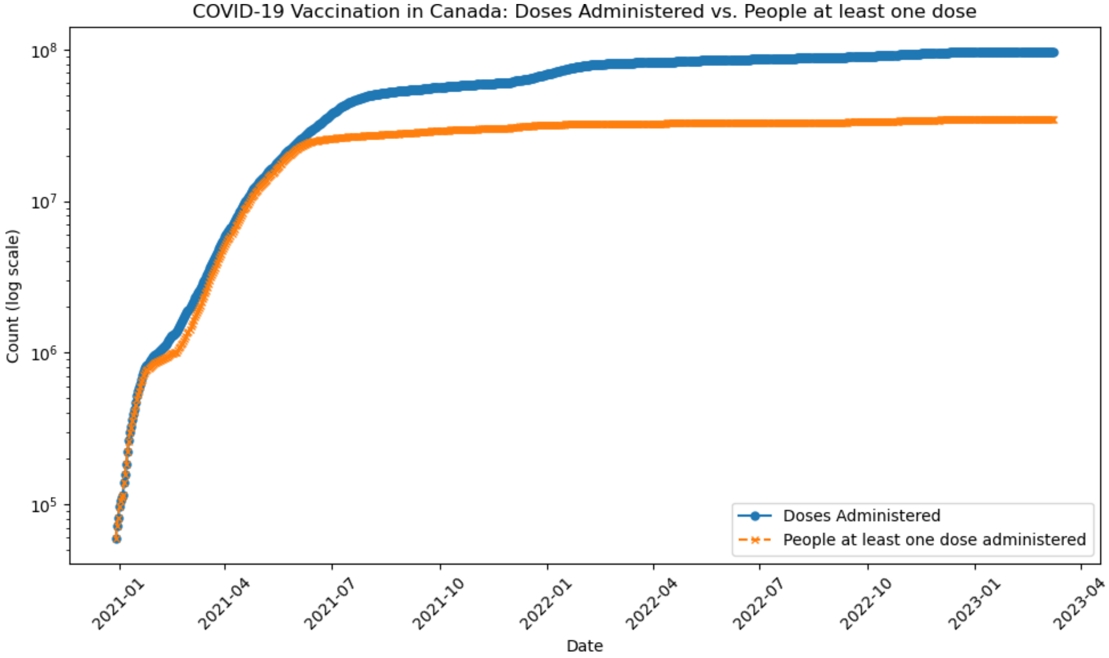
Pengxi Chen
McMaster Universtiy
July 21, 2024

[18]:
```
import fitz   # PyMuPDF
import pandas as pd
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from wordcloud import WordCloud
import matplotlib.pyplot as plt


doc = fitz.open("Donoho (2024)_Just Accepted-11706563057147.pdf")
text = ""
for page in doc:
    text += page.get_text()

lines = text.split('\n')
df = pd.DataFrame(lines, columns=['text'])
df = df.iloc[20:]
df['words'] = df['text'].apply(word_tokenize)
df = df.explode('words')
df['words'] = df['words'].str.lower()
stop_words = set(stopwords.words('english'))
df = df[~df['words'].isin(stop_words)]
df['words'] = df['words'].fillna("")
df = df[~df['words'].str.match(r'^\w$|^[^\w\s]+$|^[a-zA-Z]\.$')]
word_freq = df['words'].value_counts().reset_index()
word_freq.columns = ['word', 'freq']

wordcloud = WordCloud(width=800, height=400, background_color␣
 ↪='#ffffff',collocations=False).generate_from_frequencies(dict(word_freq.
 ↪values))

plt.figure(figsize=(10, 5))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis('off')
plt.show()
```

```
[1]: import pandas as pd
     import matplotlib.pyplot as plt

     # (1
     data_dictionary = pd.read_csv("data_dictionary.csv")
     vaccine_data = pd.read_csv("time_series_covid19_vaccine_global.csv")

     # 2
     vaccine_data_dimensions = vaccine_data.shape
     data_dictionary_preview = data_dictionary
     vaccine_data_dimensions

     # 3
     #Doses_admin: Cumulative number of doses administered. When a vaccine requires␣
      ↪multiple doses, each one is counted independently
     #People_at_least_one_dose: Cumulative number of people who received at least␣
      ↪one vaccine dose. When the person receives a prescribed second dose, it is␣
      ↪not counted twice
```

```
[1]: (142597, 6)
```

```
[2]: # (4)
     vaccine_data["Date"] = pd.to_datetime(vaccine_data["Date"])
     canada_data = vaccine_data[vaccine_data["Country_Region"] == "Canada"]

     #(5)
     plt.figure(figsize=(10, 6))
plt.plot(canada_data['Date'], canada_data['Doses_admin'], label='Doses␣
  ↪Administered', marker='o', linestyle='-', markersize=5)
plt.plot(canada_data['Date'], canada_data['People_at_least_one_dose'],␣
  ↪label='People at least one dose administered', marker='x', linestyle='--',␣
  ↪markersize=5)
plt.yscale('log')
plt.xticks(rotation=45)
plt.xlabel('Date')
plt.ylabel('Count (log scale)')
plt.title('COVID-19 Vaccination in Canada: Doses Administered vs. People at␣
  ↪least one dose')
plt.legend()
plt.tight_layout()
plt.show()
```


COVID-19 Vaccination in Canada: Doses Administered vs. People at least one dose
Count (log scale)
10^8
10^7
10^6
10^5
Doses Administered
People at least one dose administered
Date