In [10]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [12]:
```
df = pd.read_csv('covid_clinical_trials.csv')  # change name if needed
```

In [13]:
```
df.head()
```

Out[13]:

	Rank	NCT Number	Title	Acronym	Status	Study Results	
0	1	NCT04785898	Diagnostic Performance of the ID Now™ COVID-19...	COVID-IDNow	Active, not recruiting	No Results Available	
1	2	NCT04595136	Study to Evaluate the Efficacy of COVID19-0001...	COVID-19	Not yet recruiting	No Results Available	SARS-Co
2	3	NCT04395482	Lung CT Scan Analysis of SARS-CoV2 Induced Lun...	TAC-COVID19	Recruiting	No Results Available	
3	4	NCT04416061	The Role of a Private Hospital in Hong Kong Am...	COVID-19	Active, not recruiting	No Results Available	
4	5	NCT04395924	Maternal-foetal Transmission of SARS-Cov-2	TMF-COVID-19	Recruiting	No Results Available	Maternal F Transmissior

5 rows × 27 columns

In [14]:
```
df.shape
```

Out[14]:
```
(5783, 27)
```

In [15]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5783 entries, 0 to 5782
Data columns (total 27 columns):
 #   Column                   Non-Null Count  Dtype  
---  ------                   --------------  -----  
 0   Rank                     5783 non-null   int64  
 1   NCT Number               5783 non-null   object 
 2   Title                    5783 non-null   object 
 3   Acronym                  2480 non-null   object 
 4   Status                   5783 non-null   object 
 5   Study Results            5783 non-null   object 
 6   Conditions               5783 non-null   object 
 7   Interventions            4897 non-null   object 
 8   Outcome Measures         5748 non-null   object 
 9   Sponsor/Collaborators    5783 non-null   object 
 10  Gender                   5773 non-null   object 
 11  Age                      5783 non-null   object 
 12  Phases                   3322 non-null   object 
 13  Enrollment               5749 non-null   float64
 14  Funded Bys               5783 non-null   object 
 15  Study Type               5783 non-null   object 
 16  Study Designs            5748 non-null   object 
 17  Other IDs                5782 non-null   object 
 18  Start Date               5749 non-null   object 
 19  Primary Completion Date  5747 non-null   object 
 20  Completion Date          5747 non-null   object 
 21  First Posted             5783 non-null   object 
 22  Results First Posted     36 non-null     object 
 23  Last Update Posted       5783 non-null   object 
 24  Locations                5198 non-null   object 
 25  Study Documents          182 non-null    object 
 26  URL                      5783 non-null   object 
dtypes: float64(1), int64(1), object(25)
memory usage: 1.2+ MB
```

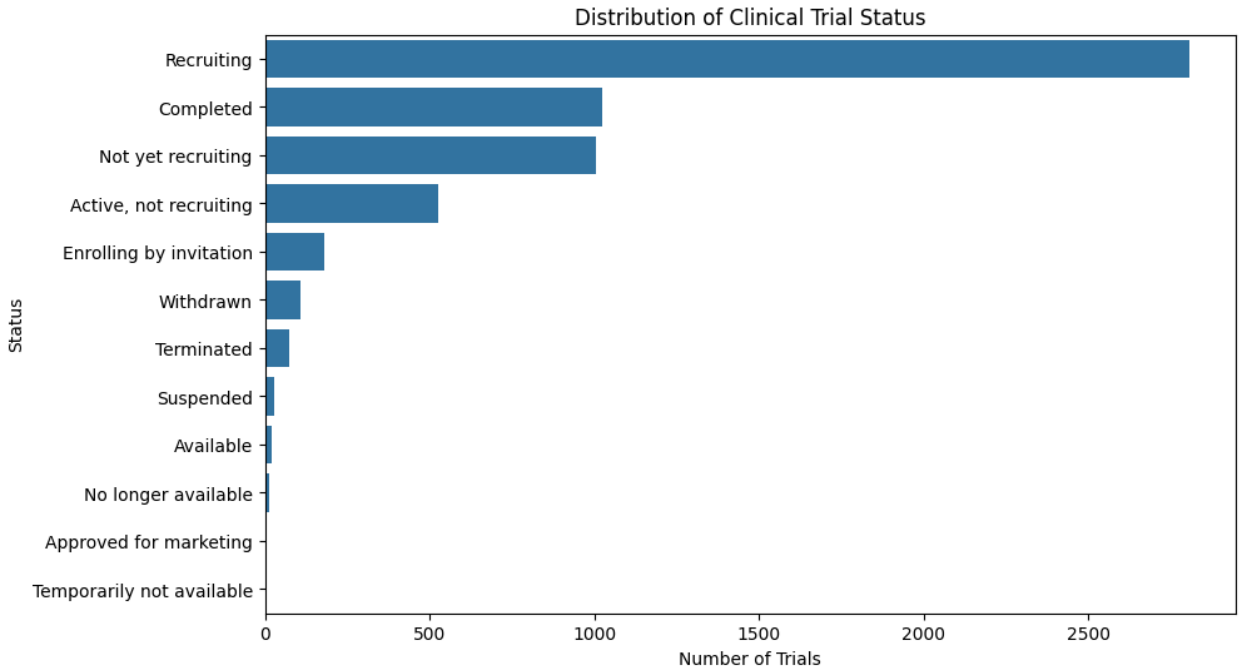
In [16]:
```
status_counts = df['Status'].value_counts()
status_counts
```

Out[16]:

Status	count
Recruiting	2805
Completed	1025
Not yet recruiting	1004
Active, not recruiting	526
Enrolling by invitation	181
Withdrawn	107
Terminated	74
Suspended	27
Available	19
No longer available	12
Approved for marketing	2
Temporarily not available	1

dtype: int64

In [17]:
```
plt.figure(figsize=(10,6))
sns.barplot(
    x=status_counts.values,
    y=status_counts.index
)
plt.title('Distribution of Clinical Trial Status')
plt.xlabel('Number of Trials')
plt.ylabel('Status')
plt.show()
```



In [18]:
```
# Extract country from Locations column
df['Country'] = df['Locations'].astype(str).apply(lambda x: x.split(',')[-1].s

# Top 10 countries
top_countries = df['Country'].value_counts().head(10)
top_countries
```

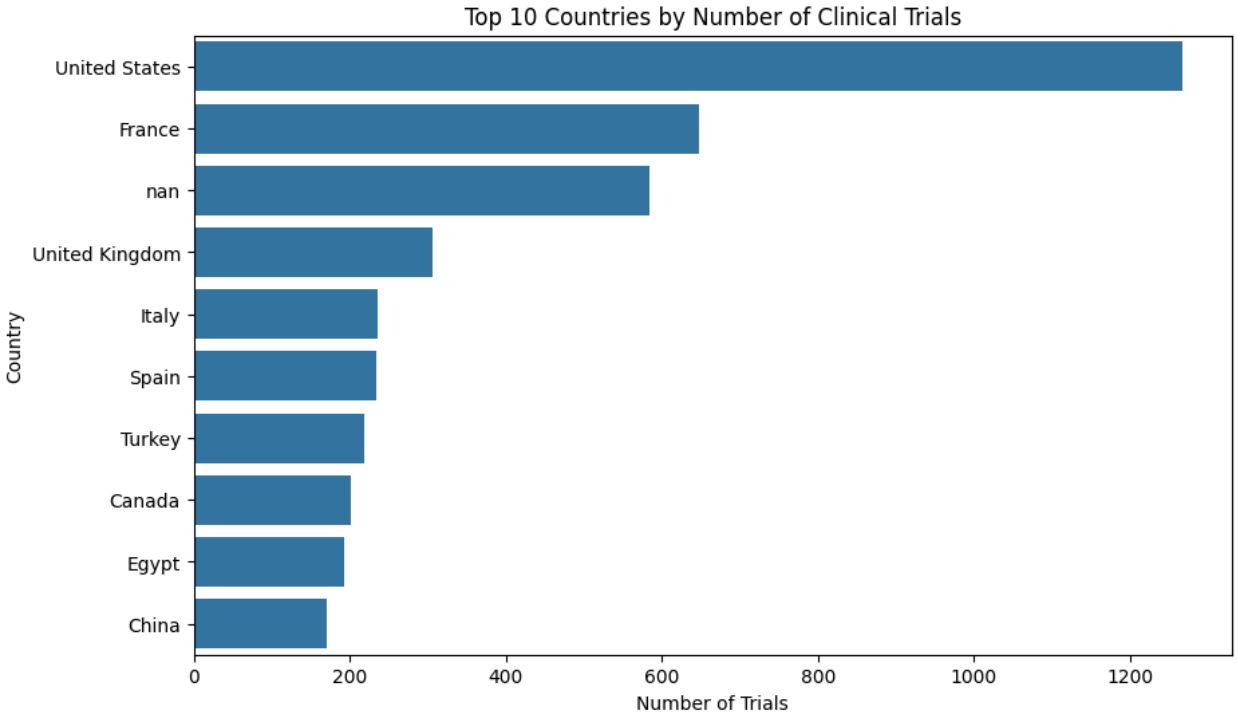
Out[18]:

Country	count
United States	1267
France	647
nan	585
United Kingdom	306
Italy	235
Spain	234
Turkey	219
Canada	202
Egypt	192
China	171

dtype: int64

In [19]:
```
plt.figure(figsize=(10,6))
```

```
sns.barplot(
    x=top_countries.values,
    y=top_countries.index
)
plt.title('Top 10 Countries by Number of Clinical Trials')
plt.xlabel('Number of Trials')
plt.ylabel('Country')
plt.show()
```



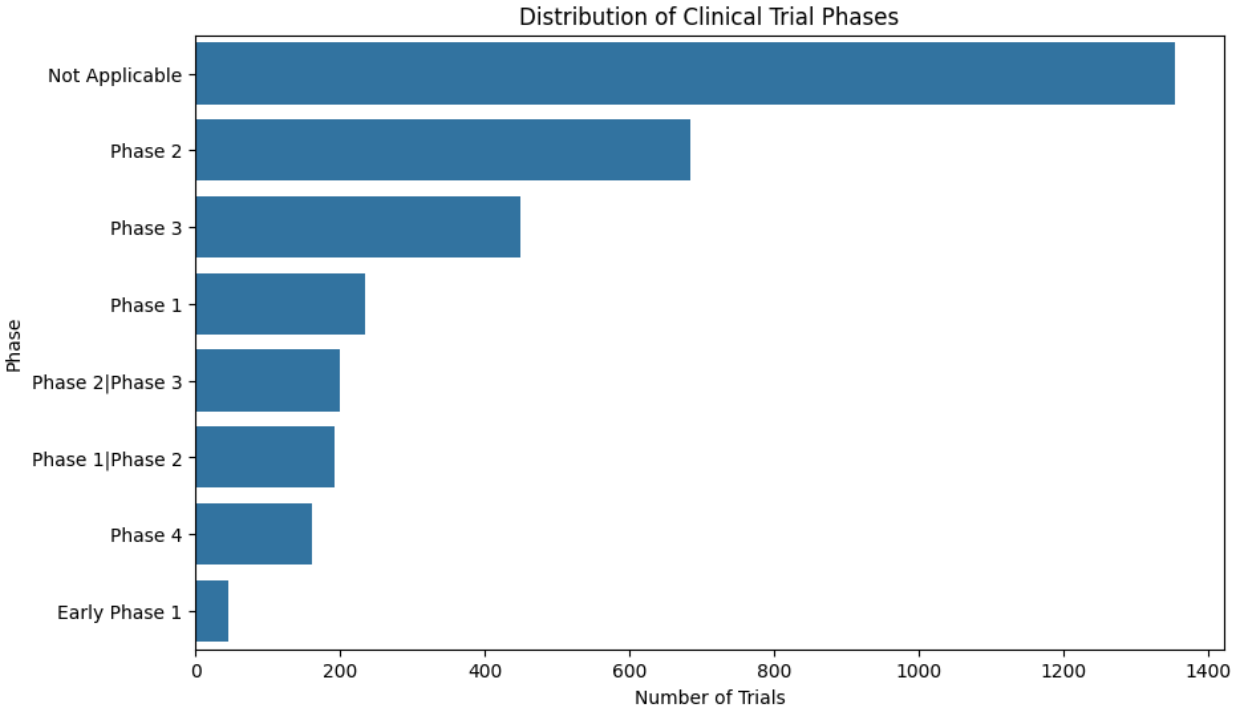
In [20]:
```
phase_counts = df['Phases'].value_counts()
phase_counts
```

Out[20]:

Phases	count
Not Applicable	1354
Phase 2	685
Phase 3	450
Phase 1	234
Phase 2|Phase 3	200
Phase 1|Phase 2	192
Phase 4	161
Early Phase 1	46

dtype: int64

In [21]:
```
plt.figure(figsize=(10,6))
sns.barplot(
    x=phase_counts.values,
    y=phase_counts.index
)
plt.title('Distribution of Clinical Trial Phases')
plt.xlabel('Number of Trials')
plt.ylabel('Phase')
plt.show()
```



In [22]:
```
gender_counts = df['Gender'].value_counts()
gender_counts
```

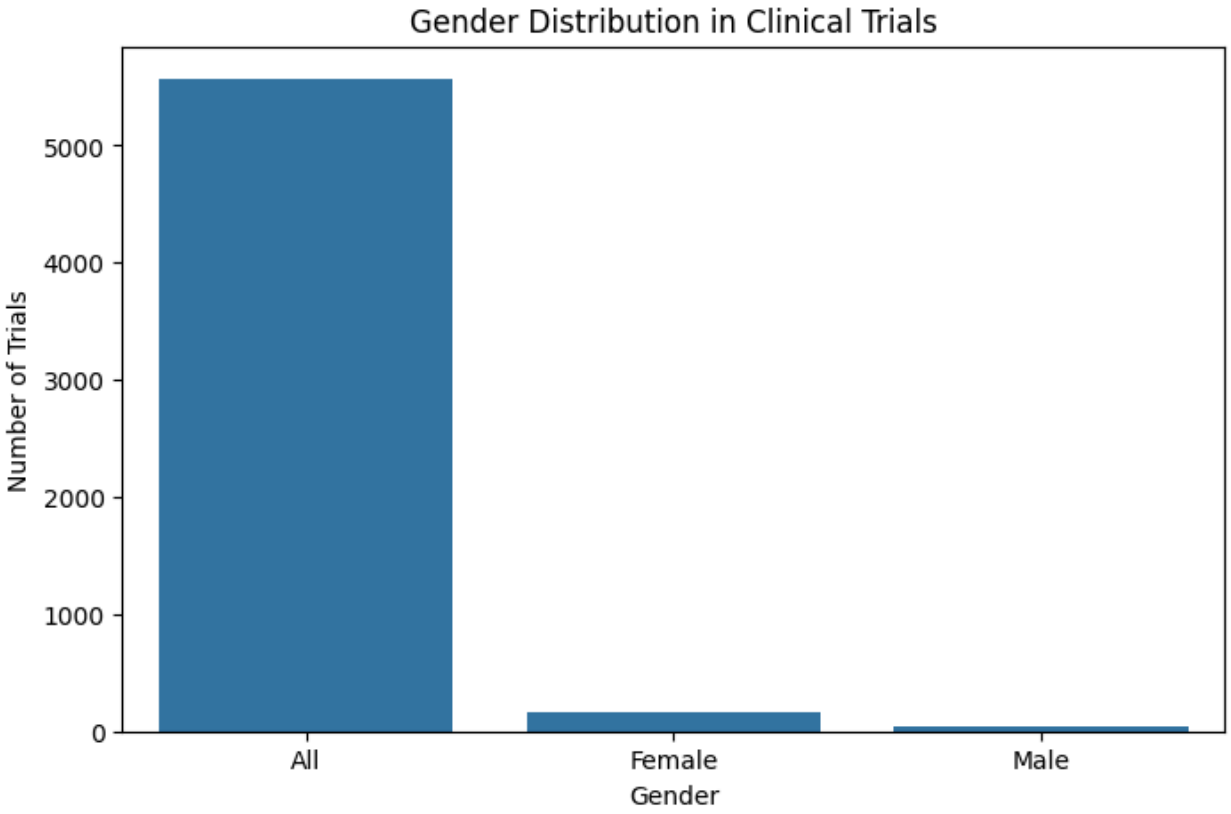
Out[22]:

Gender	count
All	5567
Female	162
Male	44

dtype: int64

In [23]:
```
plt.figure(figsize=(8,5))
sns.barplot(
    x=gender_counts.index,
    y=gender_counts.values
)
plt.title('Gender Distribution in Clinical Trials')
```

```
plt.xlabel('Gender')
plt.ylabel('Number of Trials')
plt.show()
```



In [24]:
```
# Convert Start Date to datetime
df['Start Date'] = pd.to_datetime(df['Start Date'], errors='coerce')

# Group by month
trials_over_time = df['Start Date'].dt.to_period('M').value_counts().sort_inde
trials_over_time
```

Out[24]:

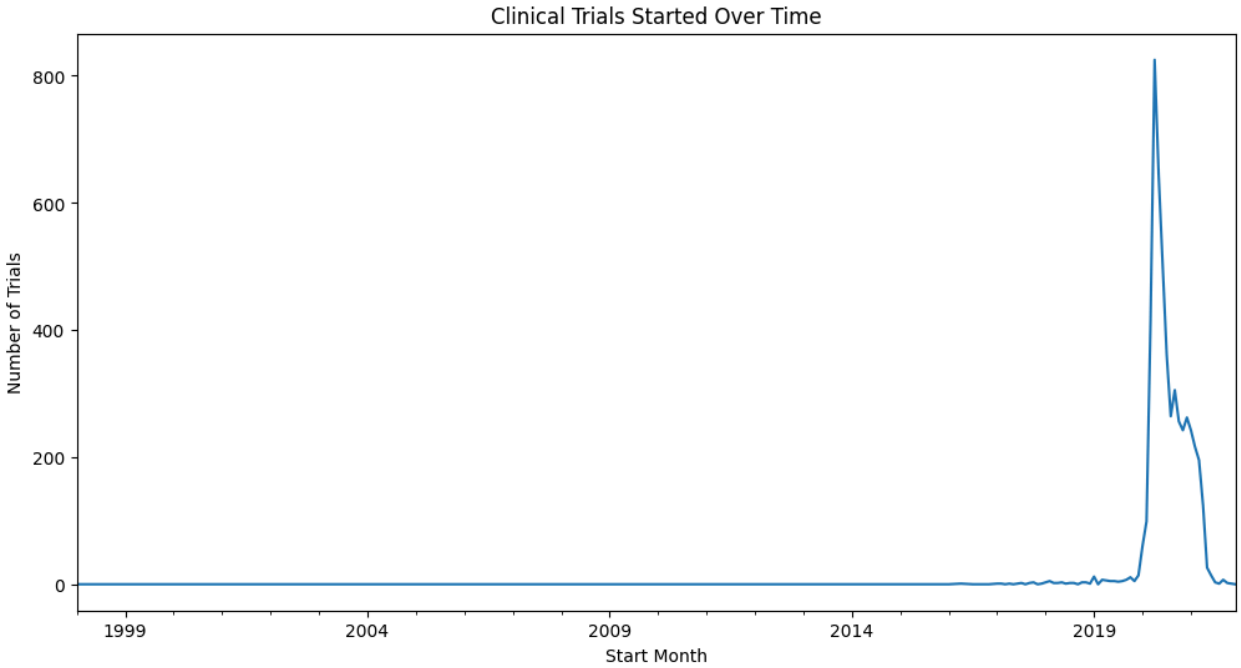
Start Date	count
1998-01	1
2010-03	1
2011-02	1
2011-03	1
2012-01	1
...	...
2021-08	2
2021-09	8
2021-10	3
2021-11	2
2021-12	1

78 rows × 1 columns

dtype: int64

In [25]:

```
plt.figure(figsize=(12,6))
trials_over_time.plot(kind='line')
plt.title('Clinical Trials Started Over Time')
plt.xlabel('Start Month')
plt.ylabel('Number of Trials')
plt.show()
```


Clinical Trials Started Over Time
Number of Trials
800
600
400
200
0
1999
2004
2009
2014
2019
Start Month