A DATA ENGINEERING PIPELINE TO RECRUIT CONTRIBUTORS TO FRIEND

Jim Peterman, PhD

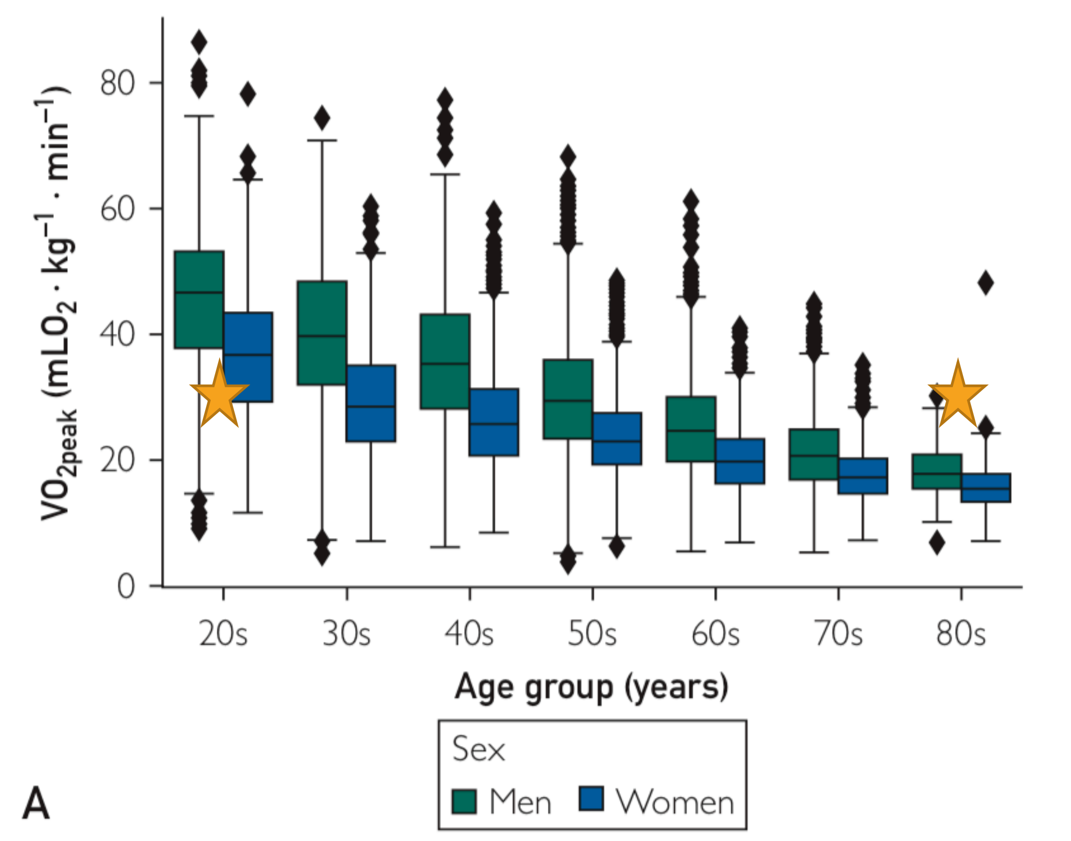
GitHub: JimPeterman

FRIEND

Fitness Registry: Importance of Exercise
National Database

BACKGROUND

- The American Heart Association recommends clinicians regularly assess fitness (“VO_2max”) because it is associated with health outcomes such as mortality and cardiovascular disease.
- Factors such as age and sex influence fitness, so interpreting an individual’s fitness requires reference standards.



BACKGROUND

- The Fitness Registry and Importance of Exercise International Database ("FRIEND") was established to create reference standards for fitness.
- The accuracy of reference standards requires representative population data.
- FRIEND currently has fitness data from >120k tests from around the world, yet more data is needed.

FRIEND

Fitness Registry: Importance of Exercise
National Database

BACKGROUND

Objective

- Build a data storage and processing pipeline that creates data summaries and predictive linear regression models that are then deployed through a web app.

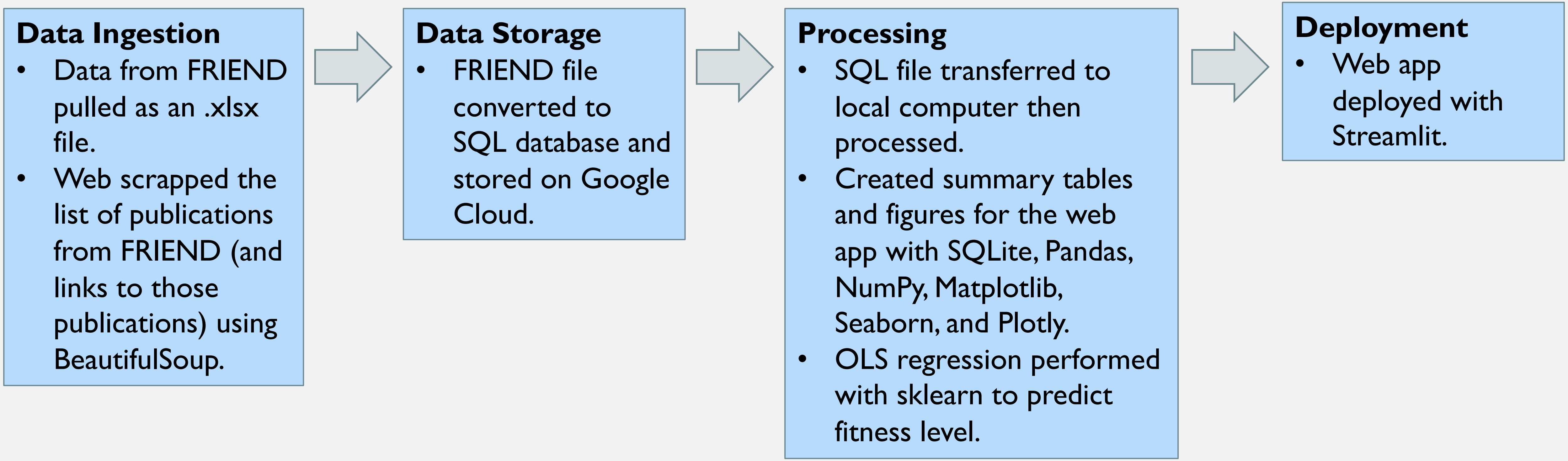
Goal

- Create a web app that highlights various aspects of FRIEND in real time to increase interest in the project and increase data contributions from testing laboratories.

METHODS: DATA PIPELINE

Testing/Robustness

- Filtered unrealistic values.
- User inputs through Streamlit ensure correct data types for regression model.
- Some user interface options determined by the database (eg, list of countries).



Data Ingestion

- Data from FRIEND pulled as an .xlsx file.
- Web scrapped the list of publications from FRIEND (and links to those publications) using BeautifulSoup.

Data Storage

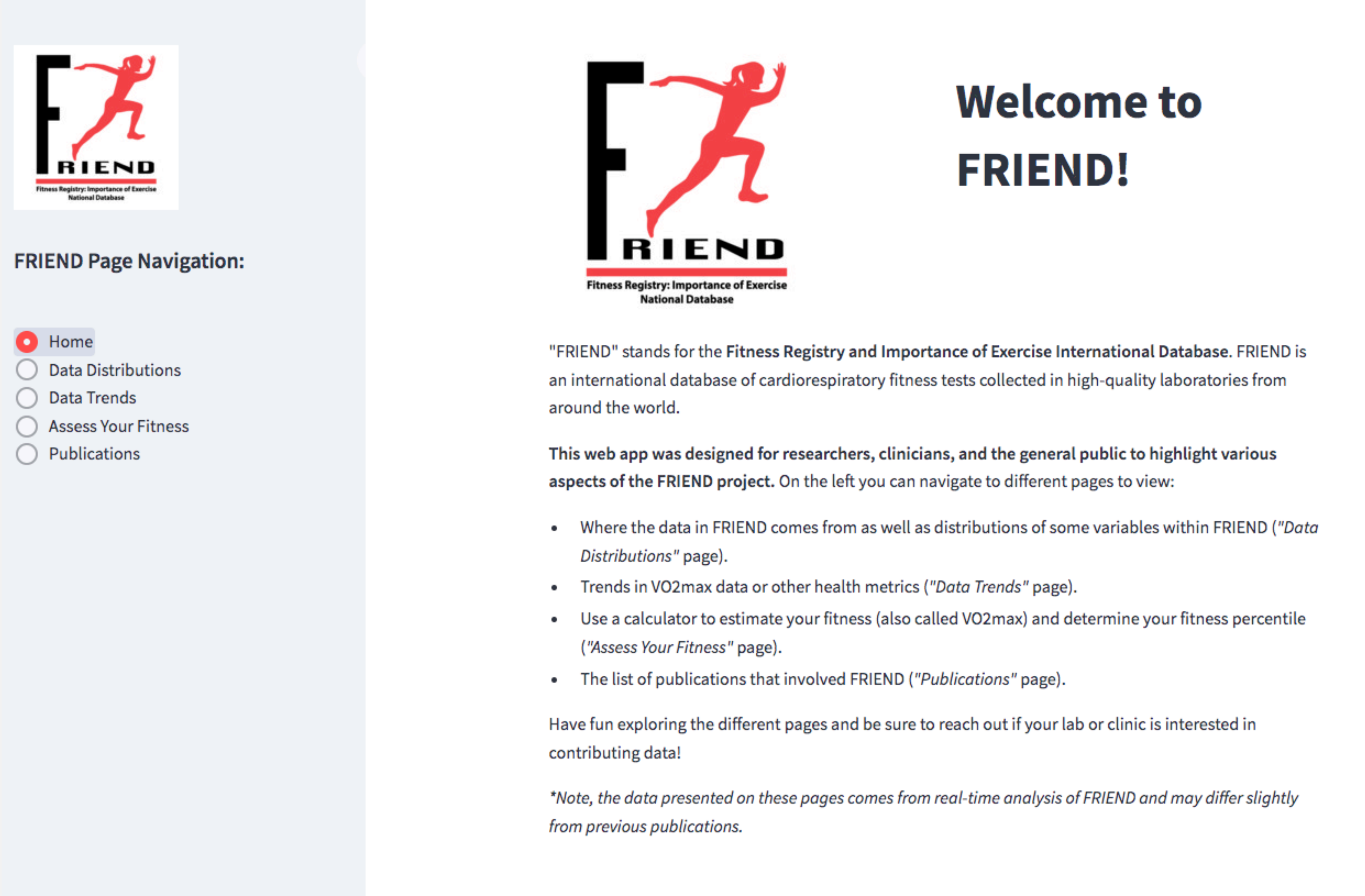
- FRIEND file converted to SQL database and stored on Google Cloud.

Processing

- SQL file transferred to local computer then processed.
- Created summary tables and figures for the web app with SQLite, Pandas, NumPy, Matplotlib, Seaborn, and Plotly.
- OLS regression performed with sklearn to predict fitness level.

Deployment

- Web app deployed with Streamlit.



Welcome to FRIEND!

"FRIEND" stands for the **Fitness Registry and Importance of Exercise International Database.** FRIEND is an international database of cardiorespiratory fitness tests collected in high-quality laboratories from around the world.

This web app was designed for researchers, clinicians, and the general public to highlight various aspects of the FRIEND project. On the left you can navigate to different pages to view:

- Where the data in FRIEND comes from as well as distributions of some variables within FRIEND (*"Data Distributions"* page).
- Trends in VO2max data or other health metrics (*"Data Trends"* page).
- Use a calculator to estimate your fitness (also called VO2max) and determine your fitness percentile (*"Assess Your Fitness"* page).
- The list of publications that involved FRIEND (*"Publications"* page).

Have fun exploring the different pages and be sure to reach out if your lab or clinic is interested in contributing data!

**Note, the data presented on these pages comes from real-time analysis of FRIEND and may differ slightly from previous publications.*

Where the test data comes from.

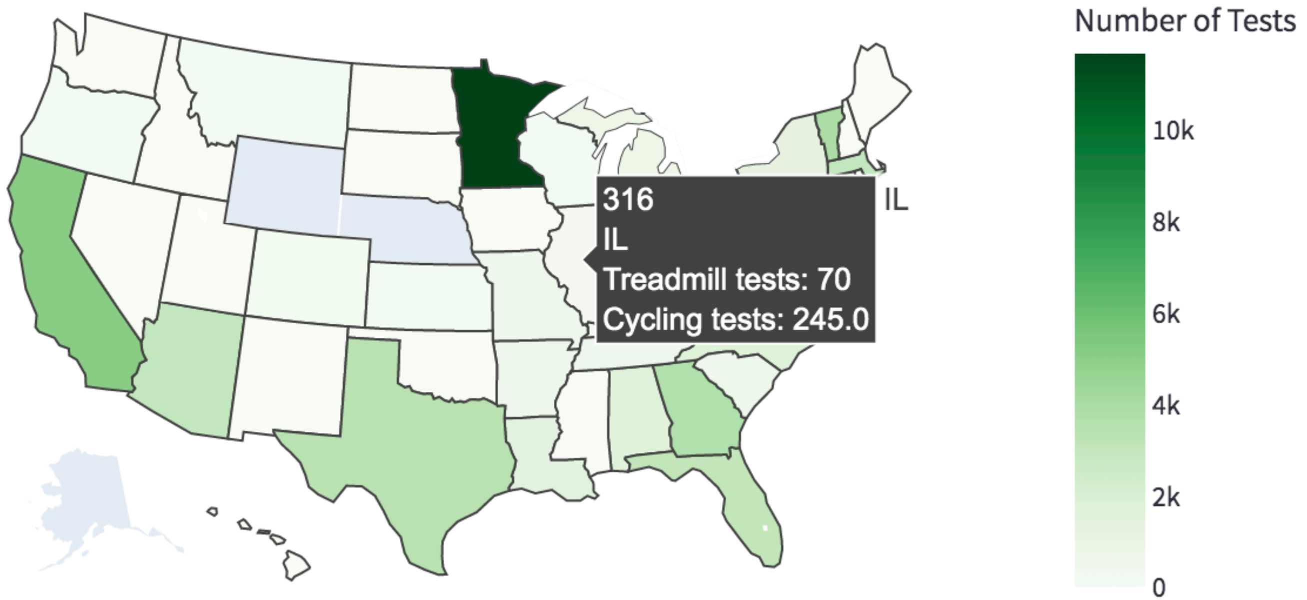
Below you can see where data for FRIEND comes from. Highlighted are the number of tests from different states in the US or from different countries around the world.

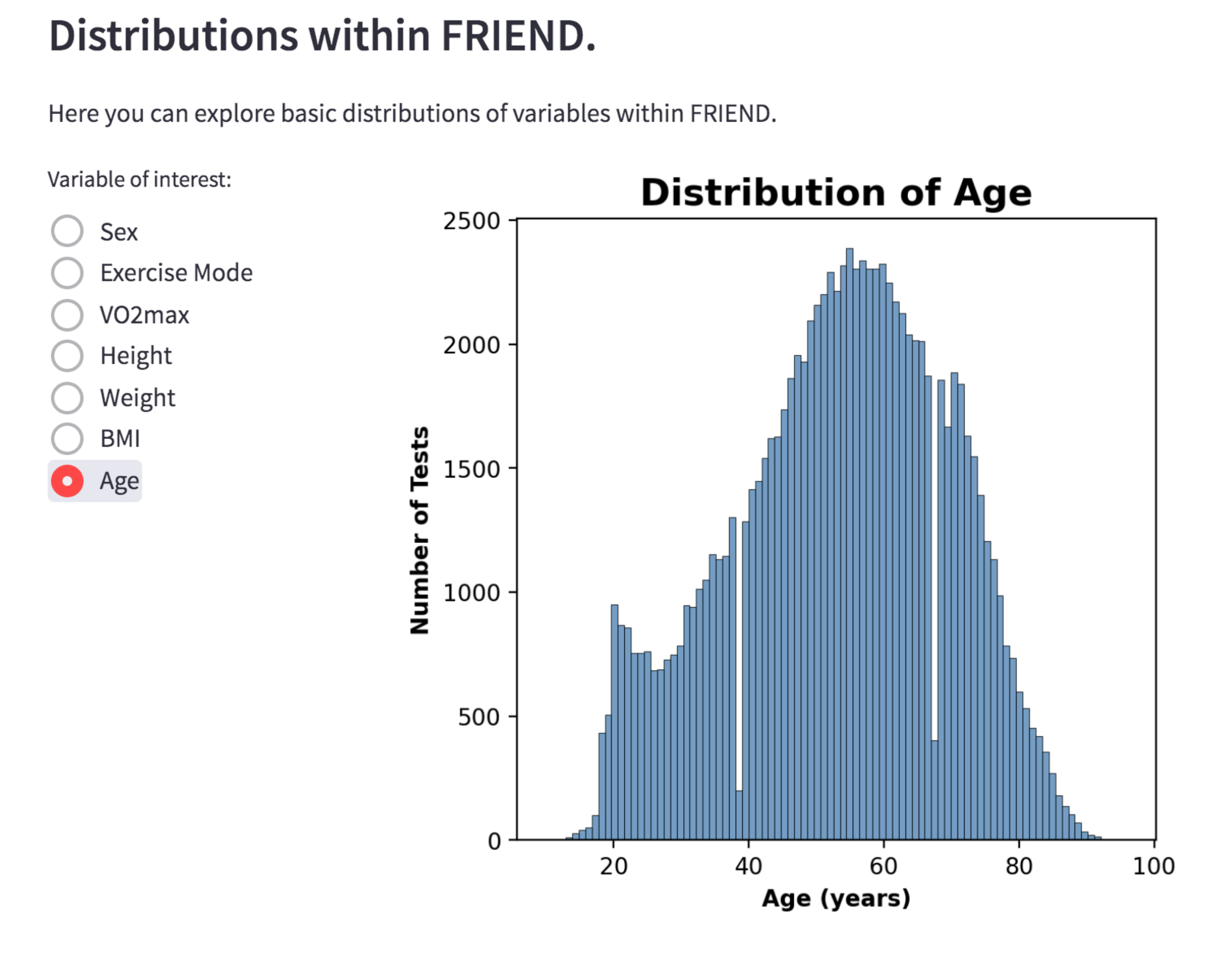
Location of interest:

● United States

○ Global

Current Distribution of Tests from the US in FRIEND
(Hover for breakdown by test mode)



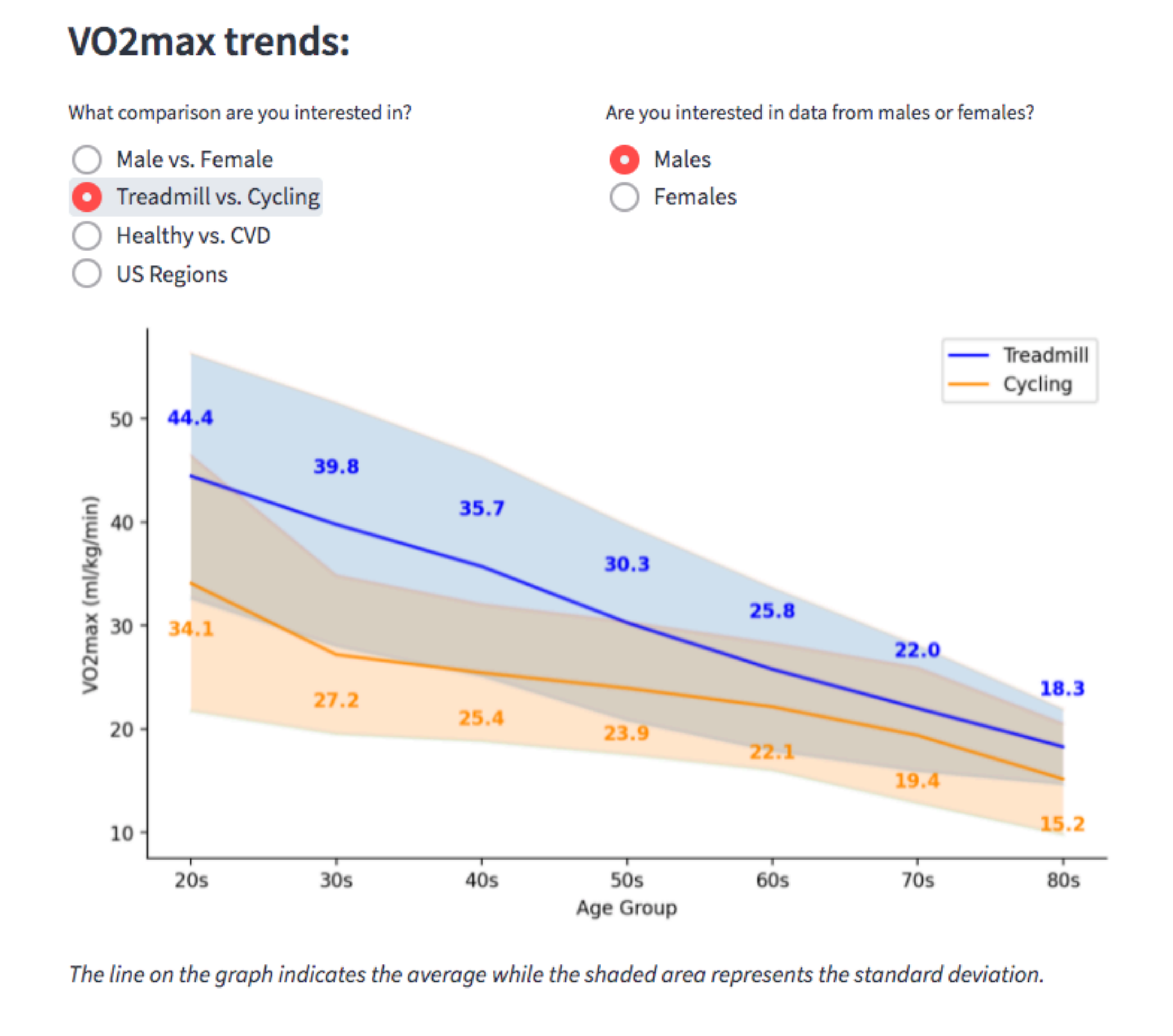


Distributions within FRIEND.

Here you can explore basic distributions of variables within FRIEND.

Variable of interest:

○ Sex
○ Exercise Mode
○ VO2max
○ Height
○ Weight
○ BMI
● Age



VO2max trends:

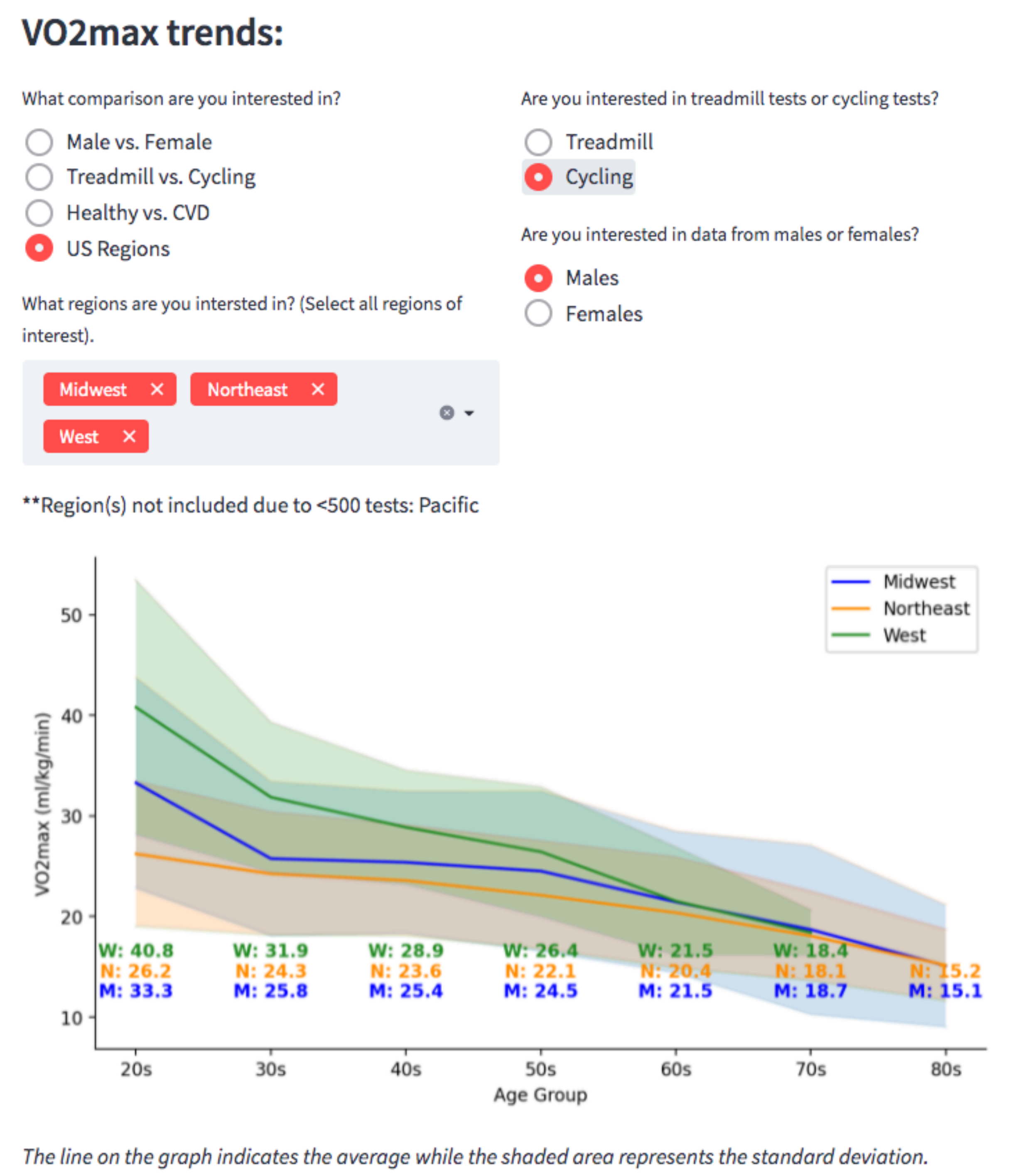
What comparison are you interested in?

- ○ Male vs. Female
- ● Treadmill vs. Cycling
- ○ Healthy vs. CVD
- ○ US Regions

Are you interested in data from males or females?

- ● Males
- ○ Females

The line on the graph indicates the average while the shaded area represents the standard deviation.



VO2max trends:

What comparison are you interested in?

○ Male vs. Female
○ Treadmill vs. Cycling
○ Healthy vs. CVD
● US Regions

What regions are you intersted in? (Select all regions of interest).

Midwest ✕ Northeast ✕ West ✕

Are you interested in treadmill tests or cycling tests?

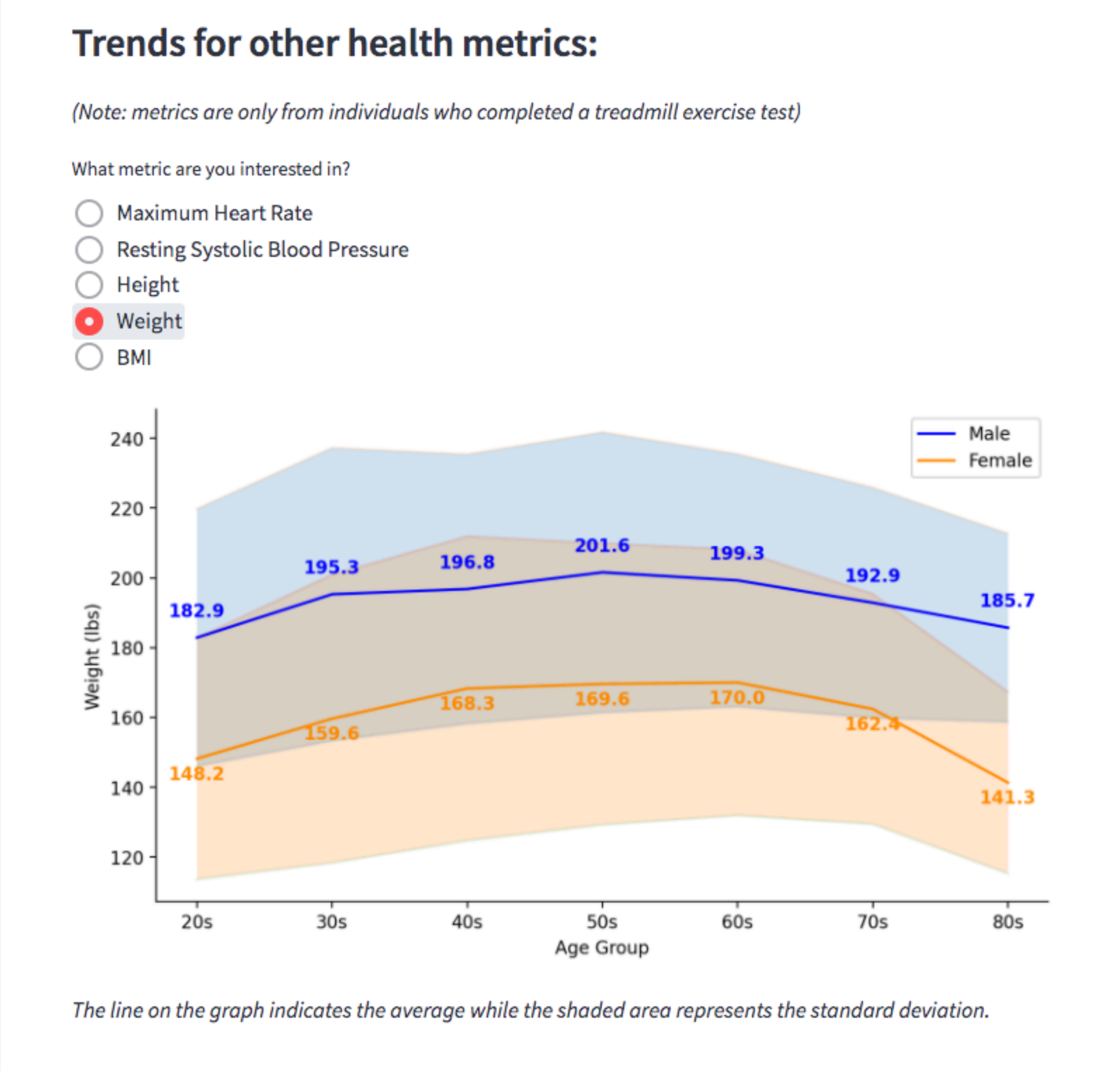
○ Treadmill
● Cycling

Are you interested in data from males or females?

● Males
○ Females

**Region(s) not included due to <500 tests: Pacific

The line on the graph indicates the average while the shaded area represents the standard deviation.



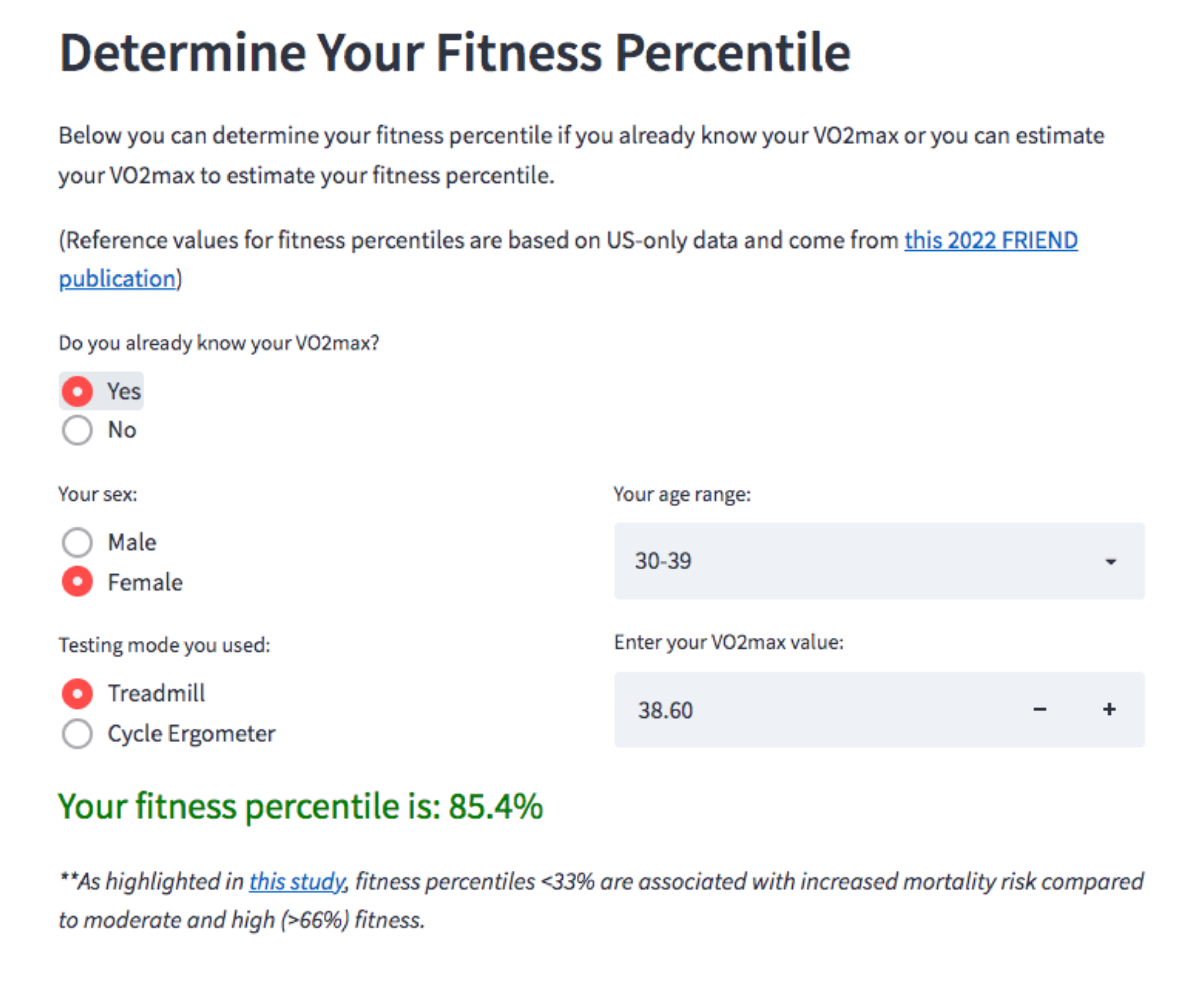
Trends for other health metrics:

(Note: metrics are only from individuals who completed a treadmill exercise test)

What metric are you interested in?

○ Maximum Heart Rate
○ Resting Systolic Blood Pressure
○ Height
● Weight
○ BMI

The line on the graph indicates the average while the shaded area represents the standard deviation.



Determine Your Fitness Percentile

Below you can determine your fitness percentile if you already know your VO2max or you can estimate your VO2max to estimate your fitness percentile.

(Reference values for fitness percentiles are based on US-only data and come from this 2022 FRIEND publication)

Do you already know your VO2max?

● Yes
○ No

Your sex:

○ Male
● Female

Your age range:

30-39

Testing mode you used:

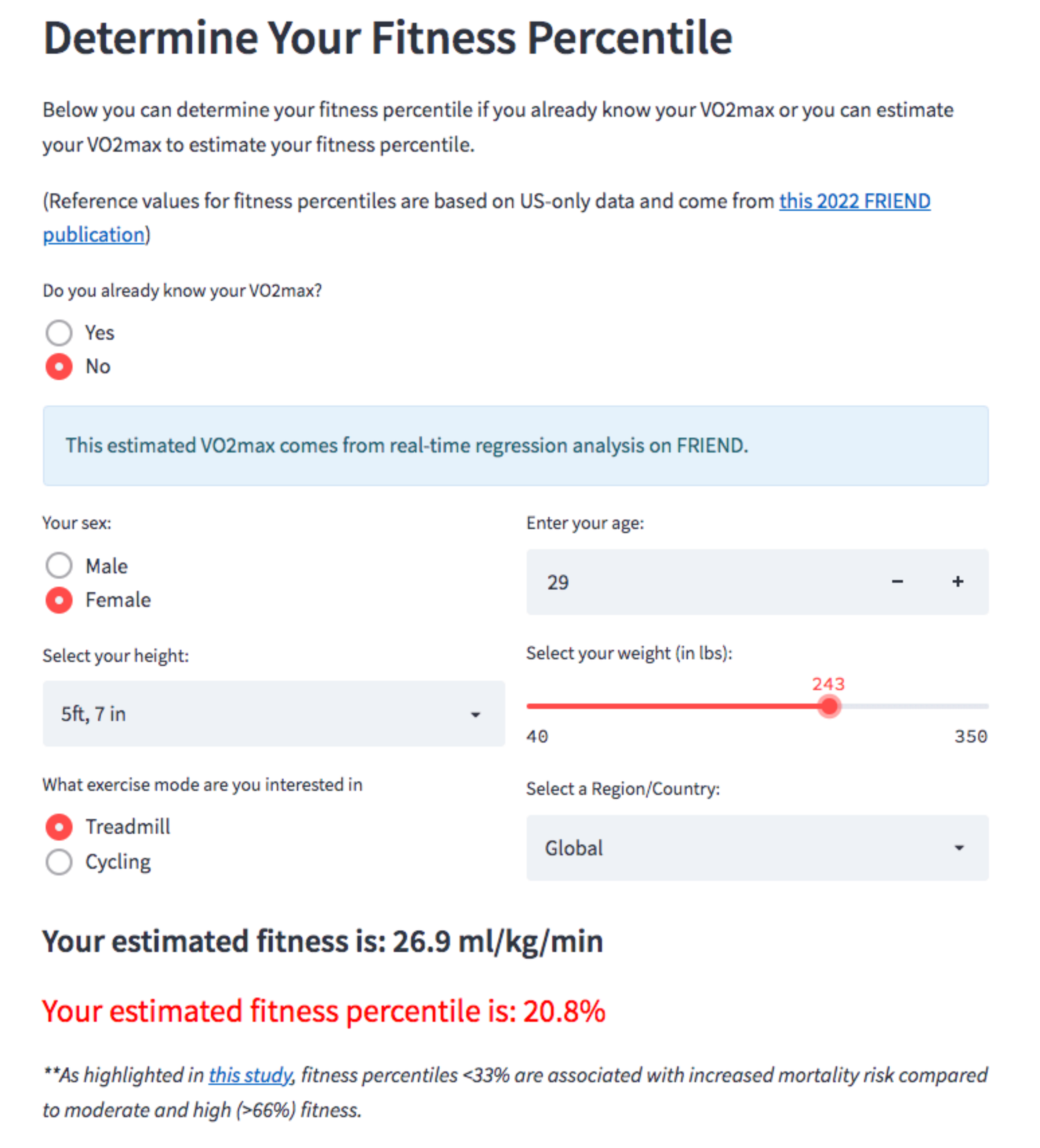
● Treadmill
○ Cycle Ergometer

Enter your VO2max value:

38.60

Your fitness percentile is: 85.4%

***As highlighted in this study, fitness percentiles <33% are associated with increased mortality risk compared to moderate and high (>66%) fitness.*



Determine Your Fitness Percentile

Below you can determine your fitness percentile if you already know your VO2max or you can estimate your VO2max to estimate your fitness percentile.

(Reference values for fitness percentiles are based on US-only data and come from this 2022 FRIEND publication)

Do you already know your VO2max?

○ Yes
● No

This estimated VO2max comes from real-time regression analysis on FRIEND.

Your sex:

○ Male
● Female

Enter your age:

29 − +

Select your height:

5ft, 7 in

Select your weight (in lbs):

243

40 350

What exercise mode are you interested in

● Treadmill
○ Cycling

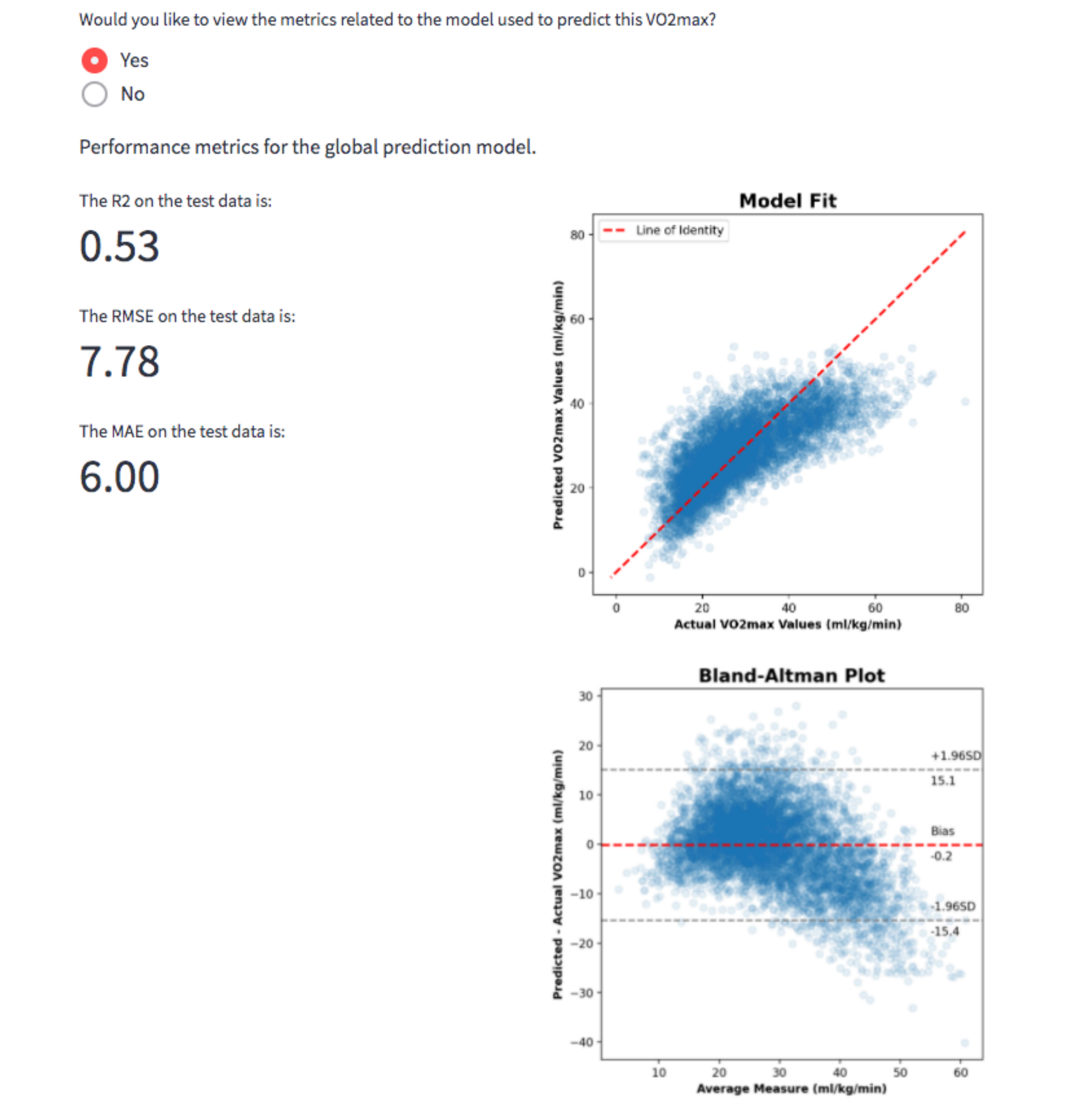
Select a Region/Country:

Global

Your estimated fitness is: 26.9 ml/kg/min

Your estimated fitness percentile is: 20.8%

***As highlighted in this study, fitness percentiles <33% are associated with increased mortality risk compared to moderate and high (>66%) fitness.*



Would you like to view the metrics related to the model used to predict this VO2max?

● Yes

○ No

Performance metrics for the global prediction model.

The R2 on the test data is:

0.53

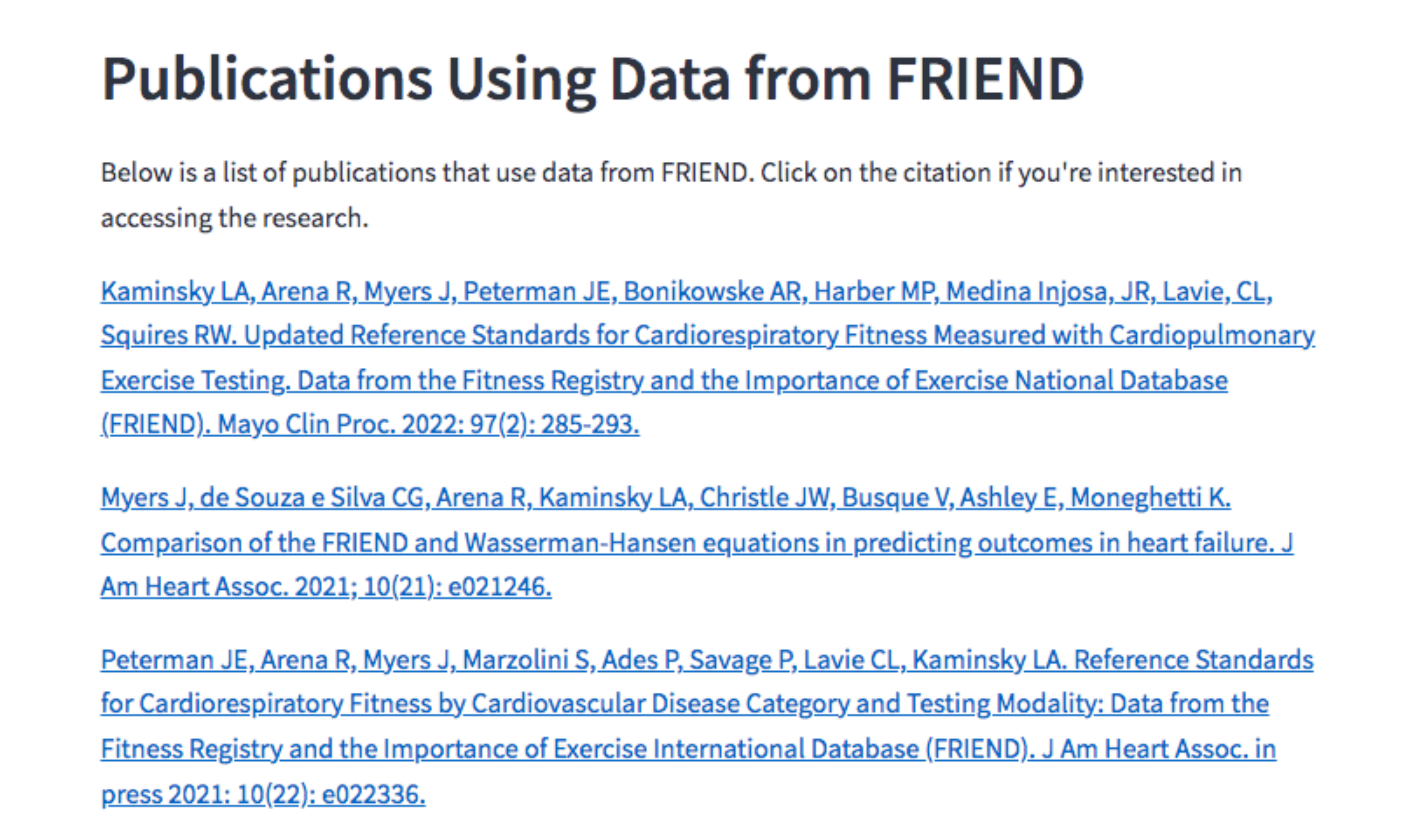
The RMSE on the test data is:

7.78

The MAE on the test data is:

6.00

Model Fit
Line of Identity
Predicted VO2max Values (ml/kg/min)
Actual VO2max Values (ml/kg/min)
0
20
40
60
80
Bland-Altman Plot
Predicted - Actual VO2max (ml/kg/min)
Average Measure (ml/kg/min)
+1.96SD
15.1
Bias
-0.2
-1.96SD
-15.4
30
20
10
0
−10
−20
−30
−40
10
20
30
40
50
60



Publications Using Data from FRIEND

Below is a list of publications that use data from FRIEND. Click on the citation if you're interested in accessing the research.

Kaminsky LA, Arena R, Myers J, Peterman JE, Bonikowske AR, Harber MP, Medina Injosa, JR, Lavie, CL, Squires RW. Updated Reference Standards for Cardiorespiratory Fitness Measured with Cardiopulmonary Exercise Testing. Data from the Fitness Registry and the Importance of Exercise National Database (FRIEND). Mayo Clin Proc. 2022: 97(2): 285-293.

Myers J, de Souza e Silva CG, Arena R, Kaminsky LA, Christle JW, Busque V, Ashley E, Moneghetti K. Comparison of the FRIEND and Wasserman-Hansen equations in predicting outcomes in heart failure. J Am Heart Assoc. 2021; 10(21): e021246.

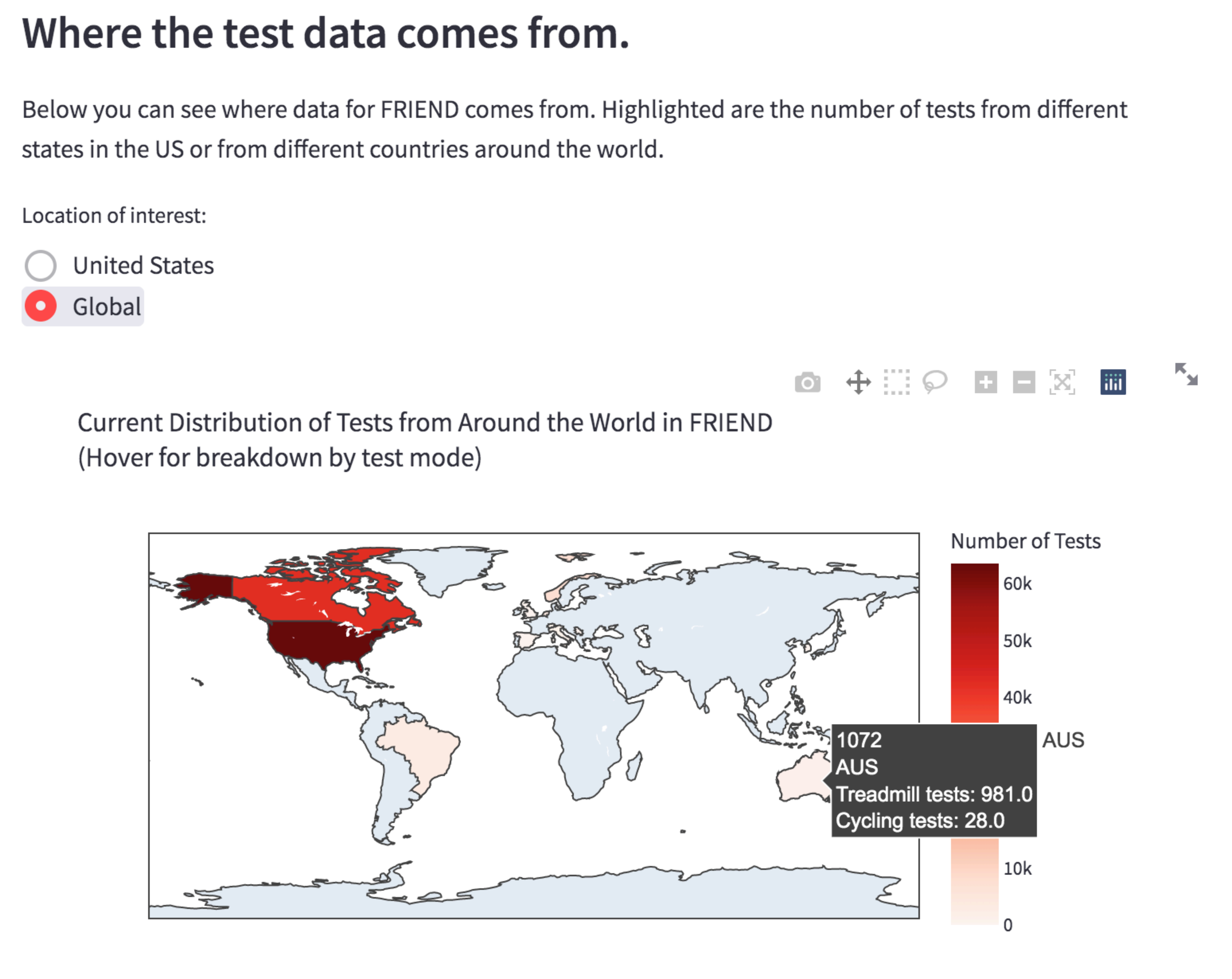
Peterman JE, Arena R, Myers J, Marzolini S, Ades P, Savage P, Lavie CL, Kaminsky LA. Reference Standards for Cardiorespiratory Fitness by Cardiovascular Disease Category and Testing Modality: Data from the Fitness Registry and the Importance of Exercise International Database (FRIEND). J Am Heart Assoc. in press 2021: 10(22): e022336.

CONCLUSION

- The data pipeline allows for easy updates to the web app as data is added to FRIEND.
- Areas of need for FRIEND are highlighted with the web app (eg, the need for more tests from individuals ~20-50 years of age)
- The web app has practical applications: individuals or clinicians can use the web app to interpret assessments of fitness.
- Overall, the web app increases interest in FRIEND for both the general public and researchers, which will hopefully lead to increased in data contributions.

FUTURE STEPS

- Add different types of regression analyses (eg, XG Boost) to improve predictions of fitness
- Add a page for researchers to download the data from FRIEND



Where the test data comes from.

Below you can see where data for FRIEND comes from. Highlighted are the number of tests from different states in the US or from different countries around the world.

Location of interest:

○ United States

● Global

Current Distribution of Tests from Around the World in FRIEND
(Hover for breakdown by test mode)