

WebENAR DataFest 2026

Information Session for Students and Mentors

Wednesday, November 19, 2026

Speakers:

Anarina Murillo, PhD, Brown University School of Public Health

Ziyi Li, PhD, The University of Texas MD Anderson Cancer Center

Liangliang Zhang, PhD, Case Western Reserve University School of Medicine



What is DataFest?

A competition for students to work together in teams on a real-world challenging research problem involving the analysis of a large data set.

Student teams (4 max) work with one of two datasets: NHANES dataset or genetic data for breast cancer.

Students are paired with an assigned mentor during the project working window.

Learning goals: Students will gain hands-on experience in data analysis, teamwork, writing and presentation skills.

Prizes

- Four teams will be invited to present at the ENAR 2026 Spring Meeting.
- Each final four teams will receive
 - Individual team member certificates
 - Discounted Spring Meeting registration
 - Invitation for a 20-minute oral presentation as part of the ENAR DataFest Invited Session
- The most outstanding final presentation will receive the ENAR DataFest Award, which includes an additional monetary prize



Who can participate?

Students

- Current ENAR Members at start time of project
- Degree candidates (master's or doctoral level)
- Able to present at ENAR 2026 Spring Meeting

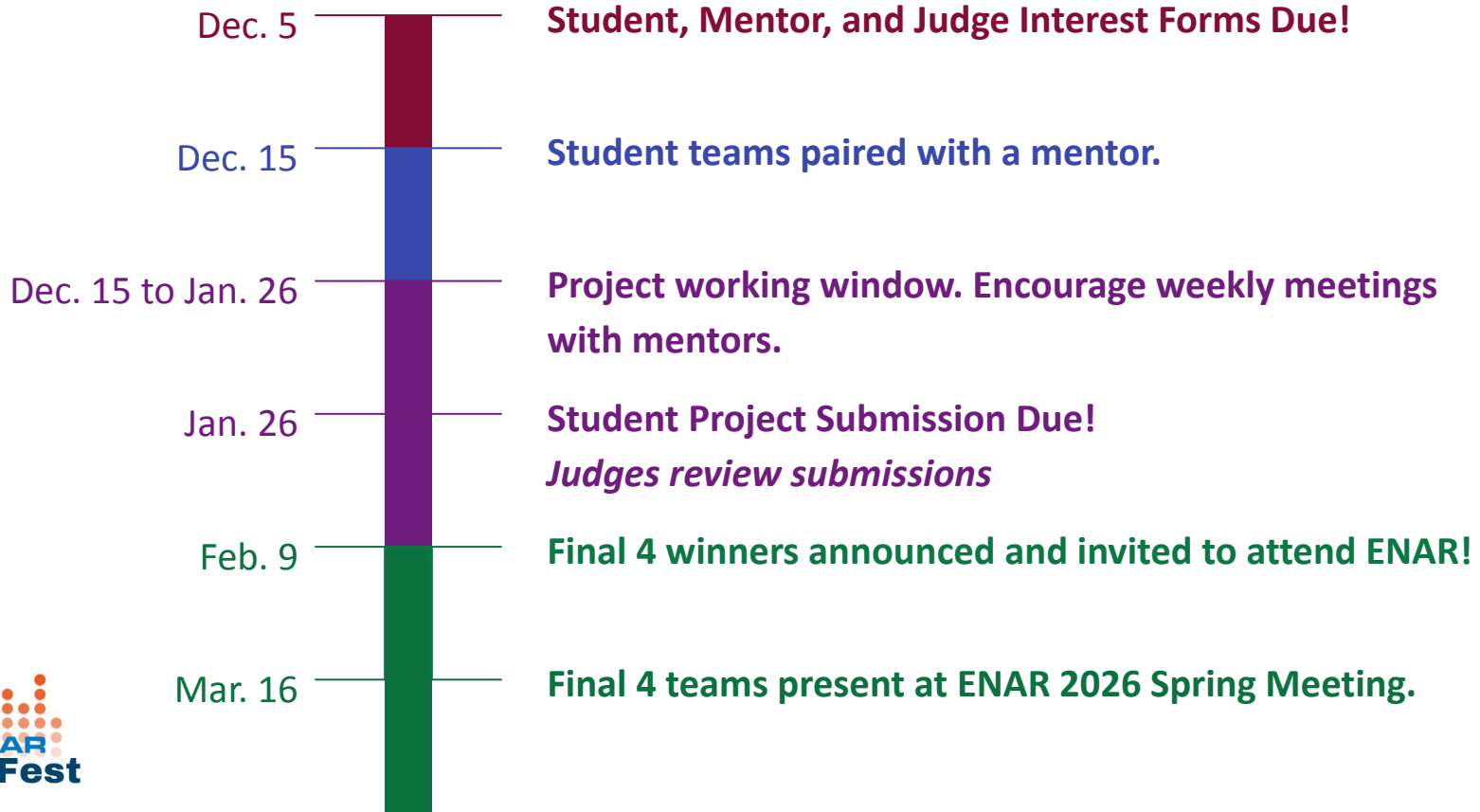
Mentors

- Faculty or working professional available to meet weekly with student team December 15, 2025 - January 26, 2026

Judges

- Faculty or working professional available to review project submissions between January 26 - February 6, 2026 and/or during the ENAR 2026 Spring Meeting

Key Deadlines



Interest Forms Due Dec. 5, 2025

Students

Name *

First Name

Last Name

Email *

example@example.com

Affiliation *

Do you already have a team? *

☐ YES

☐ NO

If not, do you want to be assigned to a team to facilitate participation in DataFest? *

☐ YES

3

Mentors or Judges

Name *

First Name

Last Name

Email *

example@example.com

Affiliation *

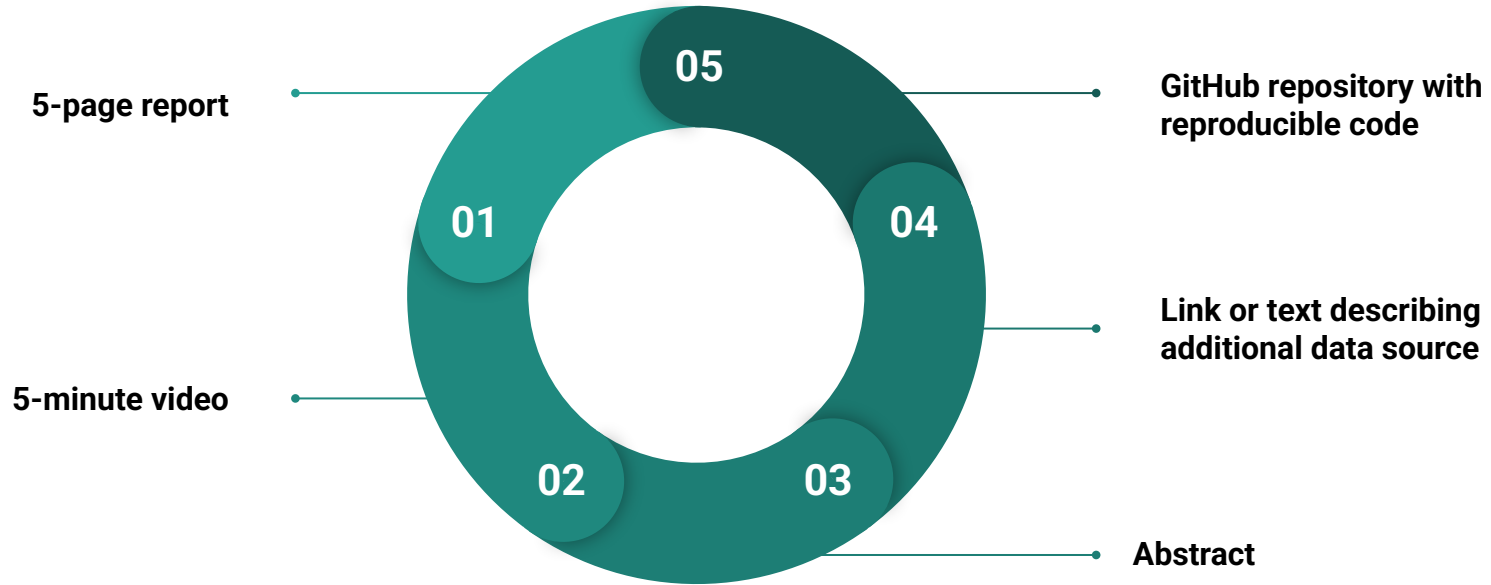
I prefer to serve as a: *

☐ Judge

☐ Mentor

☐ Either

DataFest Project Submission due Jan. 26, 2026



DataFest Project Option 1: NHANES Data in Public Health



The National Health and Nutrition Examination Survey (NHANES) data are widely used in public health and biomedical research, contributing to thousands of peer-reviewed studies that inform national health policy and clinical guidelines.

A wide range of high-impact public health topics can be examined, including trends in obesity and diabetes, nutritional deficiencies, cardiovascular risk factors, environmental toxicants, infectious disease markers, oral health, mental health, aging, and social determinants of health.

Teams should develop their own hypothesis, perform rigorous data cleaning and analysis, and generate evidence-based insights using appropriate statistical or computational methods.

Overview of NHANES Data



The NHANES program began in the early 1960s and has been conducted as a series of surveys focusing on different population groups or health topics.

In 1999, the survey became a continuous program that has a changing focus on a variety of health and nutrition measurements to meet emerging needs. The survey examines a nationally representative sample of about 5,000 persons each year. These persons are located in counties across the country, 15 of which are visited each year.

Continuous NHANES

NHANES 2025-2026	NHANES 08/2021-08/2023	NHANES 2017-March 2020 Pre-Pandemic Data	
NHANES 2019-2020	NHANES 2017-2018	NHANES 2015-2016	NHANES 2013-2014
NHANES 2011-2012	NHANES 2009-2010	NHANES 2007-2008	NHANES 2005-2006
NHANES 2003-2004	NHANES 2001-2002	NHANES 1999-2000	

Data Structure of NHANES Data



You are required to use variables taken from at least three different NHANES data sets. This must include the Demographics data set, in addition to two other data sets taken from at least one of four other available data groups (Examination, Dietary, Laboratory and Questionnaire).

1. The Demographics data group contains only 1 data sets, and should be part of all projects
2. The Dietary data group contains 14 data sets.
3. The Examination data group contains 11 data sets.
4. The Laboratory data group contains 36 data sets.
5. The Questionnaire data group also contains 37 data sets.



Data, Documentation, Codebooks

-  Demographics Data
-  Dietary Data
-  Examination Data
-  Laboratory Data
-  Questionnaire Data
-  Limited Access Data

Contents in Detail

-  Questionnaire Instruments
-  Laboratory Methods
-  Procedure Manuals
-  Brochures and Consent Documents

Using the Data

-  Brief Overview and Analytic Guidance
-  Estimates of Selected Health Conditions
-  Release Notes
-  Survey Methods and Analytic Guidelines
-  Response Rates and Population Totals
-  NHANES Web Tutorial

Contents at a Glance

-  What's New
-  Survey Content Brochure [PDF - 1.15 MB]
-  Frequently Asked Questions (FAQs)
-  General Information about NHANES Documentation Files

Documentation of NHANES Data



- For example, the Demographic Variables and Sample Weights for NHANES 2017-March 2020 [are described here](#).
- Each NHANES data set is associated with a Doc File (which stands for Data Documentation, Codebook and Frequencies). For instance, [here's the one](#) for Demographics in 2017-March 2020. This file can be viewed online (it's an HTML file) and it will tell you what variables are included in that data set.
- Each NHANES data set is available as a SAS transport file. For example, it's the P_DEMO file for Demographics in 2017-March 2020, as [you can see here](#).
- How to read XPT (SAS Transport) files in R? Try read_xpt function in the “haven” package.
- You can use the “[nhanesA](#)” package, but it might not include all information and recent years.
- For more information, please refer to [the supporting materials](#).

One published example using NHANES Data



The US NHANES contain information about demographics, blood pressure levels, hypertension status, antihypertensive medication usage, and co-morbidities for a sample of 59,799 noninstitutionalized US adults collected between 1999 and 2020.

Which preventive strategies (e.g., smoking reduction, diet interventions, physical activity) would yield the greatest reduction in CVD burden at the national level?

The data can be downloaded from GitHub or accessed via the cardioStatsUSA R package. For both the file and information about the R package, see <https://github.com/jhs-hwg/cardioStatsUSA>.

- **Documentation:** https://jhs-hwg.github.io/cardioStatsUSA/reference/nhanes_data.html.
- **Interactive exploration with a shiny app:** <https://bcjaeger.shinyapps.io/nhanesShinyBP/>
- **Supporting paper:** <https://doi.org/10.1161/hypertensionaha.123.20900>

DataFest Project Option 2: Genomics Data in Cancer Research

Patient survival prediction remains a key challenge in cancer research, with implications for both risk stratification and treatment planning. Although TCGA data is publicly available, integrating multi-omics data for outcome prediction is still an open problem.

- This project invites teams to analyze TCGA breast cancer data (pre-downloaded from cBioPortal) to develop predictive models.
- The cBioPortal platform offers access to many additional datasets across various cancer types for deeper or broader analyses.
- While the main focus can be survival prediction, teams are encouraged to explore other questions in cancer biology and treatment.

Participants are urged to think creatively about using TCGA data to advance cancer modeling and prediction efforts.

Breast Cancer (MSK, 2025)

Targeted sequencing of 3879 breast clinical cases and their matched normals via MSK-IMPACT platform.

Click

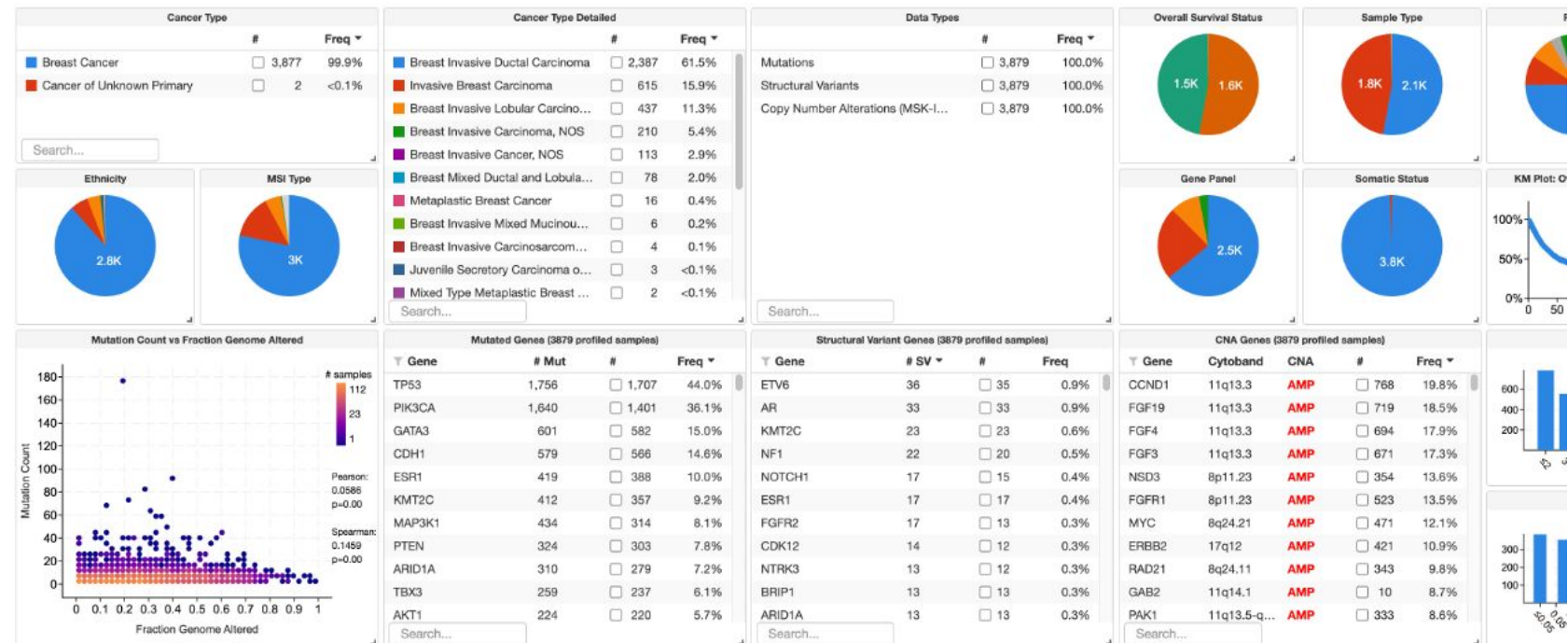
Summary

Clinical Data

CN Segments

Plots **Beta**

Selected: 3,116 patients | 3,879 samples



Datasets

The table below lists the number of available samples per cancer study and data type. It also provides links to download the data for each study. For alternative ways of downloading, see the [Download Documentation](#).

Columns ▾

Name ▾	Reference	All	Mutations	CNA	RNA-Seq
Acinar Cell Carcinoma of the Pancreas (JHU, J Pathol 2014)	Jia et al. J Pathol 2014	23	23	0	0
Acral Melanoma (TCGA, Genome Res 2017)	Liang et al. Genome Res 2017	38	38	38	36
Acute Leukemias of Ambiguous Lineage (TARGET GDC, 2025)		251	123	0	0
Acute Lymphoblastic Leukemia (St Jude, Nat Genet 2015)	Andersson et al. Nat Genet 2015	93	93	0	0
Acute Lymphoblastic Leukemia (St Jude, Nat Genet 2016)	Zhang et al. Nat Genet 2016	73	73	0	0
Acute Myeloid Leukemia (OHSU, Cancer Cell 2022)	Bottomly et al. Cancer Cell 2022	942	903	0	698
Acute Myeloid Leukemia (OHSU, Nature 2018)	Tyner et al. Nature 2018	672	622	0	451
Acute Myeloid Leukemia (TARGET GDC, 2025)		2766	584	92	0
Acute Myeloid Leukemia (TCGA GDC, 2025)		200	72	190	0
Acute Myeloid Leukemia (TCGA, Firehose Legacy)		394	197	191	173
Acute Myeloid Leukemia (TCGA, NEJM 2013)	TCGA, NEJM 2013	394	200	191	173
Acute Myeloid Leukemia (TCGA, PanCancer Atlas)	TCGA, Cell 2018	200	200	191	173
Acute myeloid leukemia or myelodysplastic syndromes (WashU, 2016)	Welch et al. N Engl J Med. 2016	136	136	0	0
Adenoid Cystic Carcinoma (FMI, Am J Surg Pathol. 2014)	Ross et al. Am J Surg Pathol 2014	28	28	28	0
Adenoid Cystic Carcinoma (JHU, Cancer Prev Res 2016)	Rettig et al. Cancer Prev Res 2016	25	25	0	0
Adenoid Cystic Carcinoma (MDA, Clin Cancer Res 2015)	Mitani et al. Clin Cancer Res 2015	102	65	0	0
Adenoid Cystic Carcinoma (MGH, Nat Gen 2016)	Drier et al. Nature Genetics 2016	10	10	0	0
Adenoid Cystic Carcinoma (MSK, Nat Genet 2013)	Ho et al. Nat Genet 2013	60	60	60	0
Adenoid Cystic Carcinoma (Sanger/MDA, JCI 2013)	Stephens et al. JCI 2013	24	24	0	0
Adenoid Cystic Carcinoma of the Breast (MSK, J Pathol. 2015)	Martelletto et al. J Pathol 2015	12	12	12	0
Adenoid Cystic Carcinoma Project (J Clin Invest 2019)	Allen et al. J Clin Invest 2019	1049	1049	928	0
Adrenocortical Carcinoma (TCGA GDC, 2025)		92	90	90	0
Adrenocortical Carcinoma (TCGA, Firehose Legacy)		172	90	90	79
Adrenocortical Carcinoma (TCGA, PanCancer Atlas)	TCGA, Cell 2018	92	91	89	78
Adult Soft Tissue Sarcomas (TCGA, Cell 2017)	TCGA, Cell 2017	206	206	206	206
Ampullary Carcinoma (Baylor College of Medicine, Cell Reports 2016)	Gingras et al. Cell Rep 2016	160	160	0	0
Anaplastic Oligodendroglioma and Anaplastic Oligoastrocytoma (MSK, Neuro Oncol 2017)	Thomas et al. Neuro Oncol 2017	22	22	22	0
Anaplastic Thyroid Cancers (GATC, Cell Reports 2024)	Zang, Peter Y F et al. Cell reports	190	190	190	0
Appendiceal Cancer (MSK, J Clin Oncol 2022)	Michael B et al. J Clin Oncol 2022	273	273	273	0
Atypical Small Cell Lung Cancer (MSK, Cancer Discov 2024)	Rekhtman N et al. Cancer Discov 2024	31	31	31	31
B-Lymphoblastic Leukemia/Lymphoma (TARGET GDC, 2025)		989	739	294	0
UNIGE, Nat Genet 2016)	Bonilla et al. Nat Genet 2016	293	293	0	0
nbia University/MSK, Cell 2018)	Lee, Suk Hyung et al. Cell 2018	130	130	130	0
Cell Reports 2022)	Clinton et al. Cell Reports 2022	1659	1659	1659	0
Clin Cancer Res 2023)	Guercio et al. Clin Cancer Res. 2023	526	526	526	0
Eur Urol 2014)	Kim et al. Eur Urol 2015	109	109	109	0

TCGA Data Description

The dataset includes multiple datasheets that can be linked by patient ID and sample ID (note that one patient may have multiple samples). The datasheets include:

- [Data_clinical_patient.txt](#): Contains clinical information about patients, including diagnosis, treatments received, age at diagnosis, overall survival time, and survival status.
- [Data_clinical_sample.txt](#): Provides details about the tumor samples, such as the associated patient, tumor site, stage, and other sample-specific annotations.
- [Data_methylation_promoters_rrbs.txt](#): Contains gene-level methylation data across samples.
- [Data_mrna_illumina_microarray.txt](#): Includes normalized mRNA expression levels from Illumina microarrays.
- [Data_mutation.txt](#): Contains mutation data for the tumor samples.

Accessing Additional TCGA Data: We encourage teams to explore and integrate additional datasets or analyze data from multiple cancer types available through the cBioPortal website (<https://www.cbioportal.org/datasets>).

FAQ

1. Can teams include students from different institutions ? **yes**
2. Do you need to be a student member to participate? **You must be a member at the start time of the project in January. Membership fee for students is \$25**
3. What is the time expectation for mentors? **We encourage meeting with student teams at least once a week between December 15, 2025 - January 26, 2026**

Thank you for your
attention!
Hope you join us this
year!

More Info:

https://www.enar.org/meetings/spring2026/program/datafest_submission.cfm

More Questions? Email us at enar@enar.org

