

A black background with blue and white text

Description automatically generated

**MyNHANES  
A Comprehensive Data Analysis Platform for NHANES**

Pre-project

## **Executive Summary**

This proposal presents a strategic vision for the development and deployment of MyNHANES, a comprehensive data analysis platform designed to streamline the manipulation, normalization, and exploration of NHANES data. The project aims to leverage advanced data processing technologies and machine learning to significantly enhance the efficiency and effectiveness of public health and cancer research. This initiative aligns with the R21 Exploratory/Developmental Research Grant (PAR-23-255), which encourages innovative approaches for secondary data analysis and integration to elucidate cancer risk and related outcomes.

We have already developed a Minimum Viable Product (MVP) version of MyNHANES, which includes functions for reading NHANES data and consolidating it into a single table. This MVP has demonstrated feasibility, with satisfactory response times, and has been applied successfully in a study on ExE interactions with lipids as the phenotype. The objective of this project is to expand the functionalities of MyNHANES.

### Primary Objectives

1. **Develop a Comprehensive Data Analysis Platform**:
   * Centralize and streamline data ingestion, transformation, and normalization processes to support real-time data analysis and ensure data quality and consistency.
   * Provide tools for customizable data queries and advanced filtering capabilities to enhance research flexibility and efficiency.
   * Enable seamless integration of NHANES data with other relevant datasets to support broader and more comprehensive analyses.
2. **Innovate Data Normalization and Transformation Processes**:
   * Design and implement a robust framework for applying data normalization and transformation rules stored in a version-controlled environment (e.g., GitHub).
   * Establish a curation group responsible for creating, modifying, and maintaining transformation rules in a public repository, ensuring transparency and community collaboration.
   * Allow users to disable unnecessary rules or create custom rules for their specific research needs, enhancing flexibility and user control.
3. **Enhance Accessibility and Usability of NHANES Data**:
   * Develop a user-friendly web-based interface for interactive data exploration and analysis.
   * Provide tools for data export in various formats (e.g., CSV, pandas DataFrame).
   * Implement comprehensive documentation and user guides to facilitate platform adoption.

### Alignment with PAR-23-255

The proposed MyNHANES platform directly supports the goals of PAR-23-255 by facilitating the secondary analysis and integration of existing datasets to elucidate cancer risk and related outcomes. By providing a robust and scalable platform for NHANES data, MyNHANES will enable researchers to:

* Conduct innovative analyses of cancer-related factors using comprehensive and normalized datasets.
* Explore new or advanced methods of data analysis to identify trends, patterns, and associations related to cancer risk and outcomes.
* Integrate NHANES data with other genomic, environmental, and lifestyle datasets to support multi-dimensional cancer research.

MyNHANES aims to revolutionize the way researchers interact with NHANES data by providing a powerful, user-friendly platform that leverages advanced AI techniques. This project not only addresses current challenges in data handling but also opens new avenues for public health and cancer research. By integrating with cutting-edge AI models and ensuring seamless integration of new NHANES cycles, MyNHANES will significantly enhance the accessibility, usability, and impact of NHANES data, aligning with the goals of the R21 Exploratory/Developmental Research Grant program.

## Context and Justification

The NHANES (National Health and Nutrition Examination Survey) program is a critical resource for public health research, providing comprehensive data on the health and nutritional status of the American population. Conducted biennially, NHANES collects a wide array of health-related data, which is invaluable for researchers studying disease prevalence, risk factors, and health outcomes.

However, the biennial cycles, with modifications in procedures and variables, pose significant challenges for researchers who need to consolidate and normalize this data for their specific research criteria. This process is often time-consuming, resource-intensive, and prone to errors, necessitating a more efficient and streamlined solution.

Given the critical role of NHANES data in public health research, particularly in studies related to cancer, there is a pressing need to develop a platform that simplifies data manipulation and ensures data integrity. MyNHANES aims to address these challenges by providing a user-friendly, scalable, and robust solution for data ingestion, transformation, and normalization.

## Proposed Technology and Architecture

MyNHANES will feature a centralized client-server architecture, designed to leverage advanced data processing technologies. The platform will support real-time data analysis, ensuring that researchers have access to high-quality, normalized data for replicable studies.

Key components include:

* A robust data ingestion and transformation framework.
* A version-controlled repository for transformation rules, maintained in a public repository like GitHub.
* Integration with cloud platforms such as AnVIL for enhanced data accessibility and usability.

### Advanced Analytics and Machine Learning Integration

The platform will integrate advanced statistical and machine learning methods to automate data insights and predictions. AI-driven data exploration tools will identify hidden patterns and correlations within the NHANES data, suggesting new research directions and generating hypotheses.

Additionally, MyNHANES will connect with state-of-the-art Large Language Models (LLMs) like OpenAI's GPT or Meta's LLaMA, enabling natural language querying and real-time assistance for researchers.

### User Interface and Experience

MyNHANES will feature a user-friendly web-based interface for interactive data exploration and analysis. The interface will provide tools for customizable data queries, advanced filtering capabilities, and data export in various formats (e.g., CSV, pandas DataFrame). Comprehensive documentation and user guides will facilitate platform adoption.

### Scalability and Sustainability

The platform will be designed to handle increasing data volumes and adapt to future research needs and technologies. Data will be stored in SQLite, with an option to migrate to PostgreSQL for concurrent multi-user access. The system can be scaled on a server, directing a namespace and port to the service, ensuring robust and sustainable infrastructure.

## Project Methodology and Strategic Implementation

### Team

Programmer

* Responsibilities: Development of the data flow, software implementation, AI integration, creation of the graphical interface, and maintenance of the transformation rules repository.
* Time: Full-time in the first year, part-time in the second year for support and adjustments.

Project Manager

* Responsibilities: Project coordination, schedule management, communication among team members, and ensuring objectives and deadlines are met.
* Time: Part-time during both years.

Researcher or Ph.D. Student

* Responsibilities: Assisting in data validation, application of use cases, development of new transformation rules, and conducting analyses and tests.
* Time: Part-time during both years, with increased hours in the second year focusing on analyses and tests.

### Budget

Total: $275,000 over two years.

Budget Breakdown

Year 1: $xx,000

* Programmer: $xx,000 (Full-time)
  + Salary and benefits for developing the data flow, implementing the software, and creating the transformation rules repository.
* Project Manager: $xx,000 (Part-time)
  + Project coordination and schedule management.
* Researcher/Ph.D. Student: $xx,000 (Part-time)
  + Assisting with initial data validation and development of new transformation rules.

Year 2: $xx,000

* Programmer: $xx,000 (Part-time)
  + Support, adjustments, and maintenance of the software and transformation rules repository.
* Project Manager: $xx,000 (Part-time)
  + Continued project coordination and schedule management.
* Researcher/Ph.D. Student: $xx,000 (Part-time)
  + Focus on analyses and tests, application of use cases, and additional development of transformation rules.

### Timeline

**Year 1: Development of Data Flow and Repository**

* **Q1**: Project kickoff, setup of the development environment, start development of the data flow.
* **Q2**: Continue data flow development, start implementation of the transformation rules repository.
* **Q3**: Finalize data flow development, continue implementation of the transformation rules repository.
* **Q4**: Initial system testing, adjustments, and improvements based on feedback.

**Year 2: Feeding Rules and Applying Analyses and Tests**

* **Q1**: Initial feeding of transformation rules into the repository, start analyses and testing on the data.
* **Q2**: Continue analyses and testing, adjust transformation rules as necessary.
* **Q3**: Final validation of rules and analyses, prepare for data publication on platforms like AnVIL.
* **Q4**: Data publication, project finalization, documentation, and preparation for future expansions.

## Expected Benefits

**Enhanced Data Management**: Centralized and streamlined data ingestion and processing, supporting real-time data analysis and ensuring replicable results through precise version control.

**Reference Platform for NHANES Data:** Establish MyNHANES as a reference platform where normalization and refinement rules have been applied based on a repository of rules supported by previous studies and curated by a dedicated team. This ensures that researchers have access to high-quality, pre-processed data for their studies.

**Advanced Analytics**: Implementation of machine learning models to automate data insights and predictions, facilitating more profound and faster research discoveries.

**Scalable and Sustainable Infrastructure**: Robust infrastructure capable of handling increasing data volumes and adaptable to future research needs and technologies.

**Support for Cancer Research**: Enabling studies on cancer risk factors, prevalence, and outcomes by providing standardized and easily accessible NHANES data.

## Next Steps

**Finalize Project Plan**: Detail the project plan and milestones, ensuring alignment with grant requirements.

**Submit the grant application** through the appropriate channels with the assistance of the responsible department at UPenn.

**Team Recruitment**: Assemble a team comprising a programmer, project manager, and researcher or Ph.D. student.

**Development and Testing**: Begin development of the platform, followed by rigorous testing and validation.

**Community Engagement**: Engage with the research community to gather feedback and ensure the platform meets user needs.

**Launch and Support**: Launch the platform, provide ongoing support, and gather continuous feedback for future improvements.