DATA AND METHODS

NHANES Data Structure

In the current protocol, NHANES data are collected over the course of two years for each survey cycle. There is a suite of data products available on health outcomes and risk factors covering the years 1999-2000 up to the most current release for 2013-2014. Prior to the 1999-2000 cycle, the design and collection were slightly different. This study uses data from 2013-2014, so following is a brief description of how the data are organized and disseminated for public use using the protocol established for the 1999-2000 cycles and later.

The NHANES data sets are grouped into Demographic, Dietary, Examination, Laboratory and Questionnaire components. The Demographic component is a master file of all participants in a given survey cycle that contains data on age, gender, race, ethnicity, socioeconomic indicators, and other characteristics. In contrast, the Dietary, Examination, Laboratory and Questionnaire components contain a collection of data sets with measurements and responses from participants in the Demographic master file. All data sets contain a unique identification number for each participant that allows researchers to link data sets from different components by person.

Each component outside of the Demographic master file is a collection of data sets that focus on a family of related outcomes or measurements. For context, the 2013-2014 Laboratory component has over 50 available data sets. In this component, for example, there is a data set that covers HDL cholesterol levels, while another one contains data on blood lead concentration.

Description of Data in Analysis

This project pulls data from the 2013-2014 Demographic master file and data from the Examination, Questionnaire, and Laboratory components from 2013-2014. The demographic data, design weights, and sampling weight for each participant are contained in DEMO_H, the master file of participants in the Demographic component.

Body Mass Index (BMI) is extracted from the Body Measures data set BMX_H within the Examination component. Disease status on thyroid disorder is obtained from the Medical Outcomes data set MCQ_H in the Questionnaire component. The exposures to PFOS and phthalates are in the data sets

SSPFAS_H and PHTHTE_H, respectively, within the Laboratory component. The table below lists the data sets used in this analysis, their respective components, and the location of their documentation.

2013-2014 NHANES Data Included in Analysis

Data Set	Data Component	Documentation
Demographic Variables and Sample Weights	Demographic	DEMO_H
Body Measurements	Examination	BMX_H
Medical Outcomes	Questionnaire	MCQ_H
Perfluoroalkyl and Polyfluoroalkyl Substances	Laboratory	SSPFAS_H
Phthalates and Plasticizers Metabolites	Laboratory	PHTHTE_H

The Demographic data master file (DEMO_H) contains 47 variables on all 10,715 participants from the 2013-2014 NHANES survey cycle, where variables include gender, age, race, citizenship, foreign-born status, language preference and other characteristics. The Body Measures data set (BMX_H) in the Examination component contains collected data on body measurements such as weight, height, and BMI for 9,743 of the 10,715 participants.

The Medical Outcomes data set (MCQ_H) in the Questionnaire component contain data on 9,970 participants with 95 variables covering major organ systems and disease types.

The PFO data set (SSPFAS_H) from the Laboratory component contains measurements on 2,165 participants and 10 variables, covering concentrations and detection status on 4 major PFO species extracted from the serum sample. Also from the Laboratory component, the Phthalate data set (PHTHTE_H) contains 2,777 participants and 29 variables, covering concentrations and detection status on 13 major phthalate species extracted from the urine samples.

The missing data on the thyroid disorder question in the Medical Outcomes data and the missing data in the Body Measures set from the Examination component are participant non-response. In contrast, both of the data sets from the Laboratory components are a random subsample of the entire participant sample from the Demographic master file and should not be handled as missing data. Appropriate sampling weights are needed here to accommodate this subsampling design feature.

It is also important to note that minority groups and the elderly are oversampled by NHANES design in order to obtain reliable estimates for these subpopulations. Additionally, the NHANES data has a complex, multi-stage, stratified sampling design. As a result, the racial and socioeconomic makeup of the sample population does not necessarily reflect the nations's population as a whole. In other words, the participant sample is not a simple random sample from the nation's population. Specifically, the 2013-2014 survey sample is 25.8% Hispanic, 33.3% White, 25.0% Black, 11.7% Asian and 4.3% others, compared to 16.3% Hispanic, 63.7% White, 12.2% Black, 4.7% Asian and 3.0% others in overall racial composition of the nation based on census data (7).

NHANES Weighting Scheme: Design and Sampling Weights

Sampling weights are included in each of the data sets so that the sample can be adjusted to reflect the composition of the nation. If an analysis uses data from the Questionnaire component only, then the sampling weight in the Demographic master file is used. Data from the other components should use the sampling weight provided in that data set. In addition to sampling weights, the Demographic master file of participants provides design weights in the form of variables “SDMVSTRA” and “SDMVPSU” to account for the multi-stage sampling design.

Not every individual in the Demographic master file has a corresponding measurement from a Laboratory component due to cost and time constraints and additional burden on survey participants. All of the data sets in the Laboratory component are random subsamples of the Demographic master file of participants. As a result, sampling weights are also provided in the Laboratory data sets (variable TSB2YR), to account for the bias that results from this subsampling scheme in a laboratory examination (8). Using the NHANES data requires careful selection of both sampling and design weights. Applying appropriate sampling weights produces unbiased estimates from a sample that is representative of the entire nation, and using the correct design weights ensures that standard errors are accurate in statistical inference. The selection of sampling weights depends on which NHANES components are used in the analysis, as described previously.

Mothods

The first step in this study is to identify the demographic covariates associated with thyroid disorder and candidate predictor variables from the laboratory results, using a series of t-tests. Subsequently a logistic model is constructed using the binary outcome of the disease status as the response. The baseline model will include all the identified demographic covariates as controls, followed by addition of each candidate predictor variable to examine whether any of laboratory results would correlate with the thyroid disease. The model would be further refined if multiple variables could be added together to improve the model fit.