A comprehensive youth diabetes epidemiological dataset and web portal: Resource Development and Case Studies

Catherine McDonough, Yan Chak Li, Nita Vangeepuram, Bian Liu, Gaurav Pandey

Submitted to: JMIR Public Health and Surveillance
on: October 05, 2023

Table of Contents

A comprehensive youth diabetes epidemiological dataset and web portal: Resource Development and Case Studies

Catherine McDonough[1*] MS; Yan Chak Li[1*] MPhil; Nita Vangeepuram[2, 3] MD, MPH; Bian Liu[2] PhD; Gaurav Pandey[1] PhD

[1]Department of Genetics and Genomic Sciences Icahn School of Medicine at Mount Sinai New York US
[2]Department of Population Health Science and Policy Icahn School of Medicine at Mount Sinai New York US
[3]Department of Pediatrics Icahn School of Medicine at Mount Sinai New York US
[*]these authors contributed equally

Corresponding Author:
Gaurav Pandey PhD
Department of Genetics and Genomic Sciences
Icahn School of Medicine at Mount Sinai
1 Gustave L. Levy Pl
New York
US

Abstract

Background: The prevalence of Type 2 diabetes (DM) and prediabetes (preDM) has been increasing among youth in recent decades in the United States, prompting an urgent need for understanding and identifying their associated risk factors. Such efforts, however, have been hindered by the lack of easily accessible youth preDM/DM data.

Objective: We aimed to first build a high quality, comprehensive epidemiological dataset focused on youth preDM/DM. Subsequently, we aimed to make this data accessible by creating a user-friendly web portal to share it and corresponding codes. Through this, we hope to address this significant gap and facilitate youth preDM/DM research.

Methods: Building on data from the National Health and Nutrition Examination Survey (NHANES) from 1999 to 2018, we cleaned and harmonized hundreds of variables relevant to preDM/DM (fasting plasma glucose level ?100 mg/dL and/or HbA1C ?5.7%) for youth aged 12-19 years (n=15,149). We identified individual factors associated with preDM/DM risk using bivariate statistical analyses and predicted preDM/DM status using our Ensemble Integration (EI) framework for multi-domain machine learning. We then developed a Prediabetes/diabetes in youth ONline Dashboard (POND) to share the data and codes.

Results: We extracted 95 variables potentially relevant to preDM/DM risk organized into 4 domains (socioeconomic status, health status, diet, and other lifestyle behaviors). The bivariate analyses identified 27 significant correlates of preDM/DM (P ?0.0005, Bonferonni adjusted), including race/ethnicity, health insurance, BMI, added sugar intake, and screen time. Seventeen of these factors were also identified based on the EI methodology (Fisher's P of overlap=7.06x10-6). In addition to those, the EI approach identified 11 additional predictive variables, including some known (e.g., meat and fruit intake and family income) and less recognized factors (e.g., number of rooms in homes). The factors identified in both analyses spanned over all 4 of the domains mentioned. These results as well as other exploratory tools can be accessed on POND by users of any background.

Conclusions: Using NHANES data, we built one of the largest public epidemiological datasets for studying youth preDM/DM and identified potential risk factors using complementary analytical approaches. Our results align with the multifactorial nature of preDM/DM with correlates across several domains. Also, our data-sharing platform, POND, facilitates a wide range of applications to inform future youth preDM/DM studies.

(JMIR Preprints 05/10/2023:53330)
DOI: https://doi.org/10.2196/preprints.53330

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
Please make my preprint PDF available to anyone at any time (recommended).

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.

✓ **No, I do not wish to publish my submitted manuscript as a preprint.**

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

Authors

Catherine McDonough[1], Yan Chak Li[1], Nita Vangeepuram[2,3], Bian Liu[3*], Gaurav Pandey[1*]
*Corresponding authors: bian.liu@mountsinai.org; gaurav.pandey@mssm.edu

Affiliations

[1]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[2]Department of Pediatrics, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[3]Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai, New York, NY, USA

A comprehensive youth diabetes epidemiological dataset and web portal: Resource Development and Case Studies

Abstract

Background: The prevalence of Type 2 diabetes (DM) and prediabetes (preDM) has been increasing among youth in recent decades in the United States, prompting an urgent need for understanding and identifying their associated risk factors. Such efforts, however, have been hindered by the lack of easily accessible youth preDM/DM data.

Objective: We aimed to first build a high quality, comprehensive epidemiological dataset focused on youth preDM/DM. Subsequently, we aimed to make this data accessible by creating a user-friendly web portal to share it and corresponding codes. Through this, we hope to address this significant gap and facilitate youth preDM/DM research.

Methods: Building on data from the National Health and Nutrition Examination Survey (NHANES) from 1999 to 2018, we cleaned and harmonized hundreds of variables relevant to preDM/DM (fasting plasma glucose level ≥100 mg/dL and/or HbA1C ≥5.7%) for youth aged 12-19 years (n=15,149). We identified individual factors associated with preDM/DM risk using bivariate statistical analyses and predicted preDM/DM status using our Ensemble Integration (EI) framework for multi-domain machine learning. We then developed a user-friendly web portal named Prediabetes/diabetes in youth ONline Dashboard (POND) to share the data and codes.

Results: We extracted 95 variables potentially relevant to preDM/DM risk organized into 4 domains (sociodemographic, health status, diet, and other lifestyle behaviors). The bivariate analyses identified 27 significant correlates of preDM/DM (P ≤0.0005, Bonferroni adjusted), including race/ethnicity, health insurance, BMI, added sugar intake, and screen time. Sixteen of these factors were also identified based on the EI methodology (Fisher's P of overlap=7.06×10^{-6}). In addition to those, the EI approach identified 11 additional predictive variables, including some known (e.g., meat and fruit intake and family income) and less recognized factors (e.g., number of rooms in homes). The factors identified in both analyses spanned over all 4 of the domains mentioned. These data and results, as well as other exploratory tools, can be accessed on POND (https://rstudio-connect.hpc.mssm.edu/POND/).

Conclusions: Using NHANES data, we built one of the largest public epidemiological datasets for studying youth preDM/DM and identified potential risk factors using complementary analytical approaches. Our results align with the multifactorial nature of preDM/DM with correlates across several domains. Also, our data-sharing platform, POND, facilitates a wide range of applications to inform future youth preDM/DM studies.

Keywords:
Youth prediabetes and diabetes; public dataset; NHANES; web data portal; epidemiology; biostatistics; machine learning

Introduction

Type 2 diabetes mellitus (DM) is a complex disease influenced by several biological and epidemiological factors [1,2], such as obesity [3], family history [4], diet [1,5], physical activity level [1,6–8], and socioeconomic status [9–11]. Prediabetes, characterized by elevated blood glucose levels below the diabetes threshold, is a precursor condition to DM [12]. There has been an alarming increasing trend in the prevalence of youth with prediabetes and DM (preDM/DM) both in the United States [13–19] and worldwide [20,21], and the numbers of newly diagnosed youth living with preDM/DM are also expected to increase [14,20,22]. The latest estimate based on nationally representative data showed that the prevalence of preDM among youth increased from 11.6% in 1999-2002 to 28.2% in 2015-2018 in the United States [13]. This growth is particularly concerning because preDM/DM disproportionately affects racial and ethnic minority groups and those with low socioeconomic status [9–11,22–24], leading to significant health disparities. Having preDM/DM at a younger age also confers a higher health and economic burden resulting from living with the condition for more years and a higher risk of developing other cardiometabolic diseases [25–30]. This serious challenge calls for increased translational research into factors associated with preDM/DM among youth and how they can collectively affect disease risk and inform prevention strategies.

In particular, the most critically needed research in this direction is exploring the collective impact of various risk factors across multiple health-related domains. While clinical factors, such as obesity, have been mechanistically linked to insulin resistance [31], it is important to consider the broader perspective. There is an increasing recognition that social determinants of health (SDoH) play a significant role in amplifying the risk of preDM/DM and their related disparities. For example, factors such as limited access to healthcare, food and housing insecurity, and the neighborhood-built environment have been identified as influential contributors [9–11,32]. However, to gain a comprehensive understanding, it is essential to delve into other less studied variables, such as screen time, acculturation, or frequency of eating out, and examine how they interact to increase the risk of preDM/DM among youth [2].

One of the major challenges that has limited translational research into youth preDM/DM risk factors is that there are no publicly available, easily accessible data comprehensively profiling interrelated epidemiological factors for young individuals [2]. Specifically, most available public diabetes data portals focus on providing aggregated descriptive trends, such as preDM/DM prevalence for the entire population or subgroups stratified by race and ethnicity [33–36], which does not allow in-depth examination of the relationships between multiple risk factors and preDM/DM risk using individual level data. While there do exist a few individual-level public diabetes datasets [37–41], they include mainly clinical measurements, while other important risk factors such as those related to diet, physical activity, and SDoH are limited. In addition, these datasets are not available for youth

populations, as they either focus exclusively on adult populations and not on youth specifically [37,39–41]. Furthermore, these datasets are not accompanied by any user-friendly online portals that can help explore or analyze these data to reveal interesting knowledge about youth preDM/DM. This shows that there is a lack of a comprehensive dataset that includes multiple epidemiological variables to study youth preDM/DM, and easily usable functionalities to explore and analyze data.

To directly address this data gap, we turned to the National Health and Nutrition Examination Survey (NHANES), which offers a promising path for examining preDM/DM among the US youth population by providing a rich source of individual- and household-level epidemiological factors. As a result, NHANES has been a prominent data source for studying youth preDM/DM trends and associated factors [18,42–45]. However, the utilization of NHANES data requires extensive data processing that is laborious and time-intensive [46]. This represents a major challenge for the wide-spread use of these high-quality and extensive data for studying youth preDM/DM.

In this work, we directly addressed the above challenges by processing NHANES data from 1999 to 2018 into a large-scale youth diabetes-focused dataset that covers a variety of relevant variable domains, namely sociodemographic factors, health status indicators, diet and other lifestyle behaviors. We also provided public access to this high-quality comprehensive youth preDM/DM dataset, as well as functionalities to explore and analyze it, through the user-friendly Prediabetes/diabetes in youth ONline Dashboard (POND) [47]. We demonstrated the dataset's utility and potential through two case studies that employed statistical analyses and machine learning (ML) approaches, respectively, to identify important epidemiological factors that are associated with youth preDM/DM.

Through this work, we aim to advance youth diabetes research by providing the most comprehensive epidemiological dataset available through a public web portal, and illustrating the value of these resources through our example case studies based on statistical analyses and machine learning. Our overarching goal is to enable researchers to investigate the multifactorial variables associated with youth preDM/DM, which may drive translational advances in prevention and management strategies

Methods

Figure 1 shows the overall study design and workflow. Below, we detail the components of the workflow.

Data source and study population

We built the youth preDM/DM dataset based on publicly available NHANES data [48] spanning the years 1999 to 2018. Developed by the Centers for Disease Control and Prevention (CDC), NHANES is a serial cross-sectional survey that gathers comprehensive health-related information from nationally representative samples of the non-institutionalized population in the United States. The survey employs a multi-stage probability sampling method and collects data through questionnaires, physical examinations, and biomarker analysis. Each year, approximately 5,000 individuals are included in the survey, and the data are publicly released in 2-year cycles. CDCobtained written informed consent from a parent or guardian for participants <18 years old at the time of enrollment. The NHANES survey procedures and protocol were approved by the National Center for Health Statistics Ethics Review Board [49].

Figure 2 details the process used to define our study population. Briefly, of the total 101,316

participants in 1999-2018 NHANES, we excluded individuals who (i) were not within the 12–19-year age range, (ii) did not have either of the biomarkers used to define preDM/DM status, and (iii) answered “Yes” to “Have you ever been told by a doctor or health professional that you have diabetes?”. The youth preDM/DM outcome of this work was derived as follows: youth were considered at risk of preDM/DM if their Fasting Plasma Glucose (FPG) was at or greater than 100 mg/dL, or their glycated hemoglobin (HbA1C) was at or greater than 5.7%, according to the current American Diabetes Association (ADA) pediatric clinical guidelines [2].

Validation of the study population

We estimated preDM/DM prevalence across the ten survey cycles (1999-2018) by incorporating the NHANES design elements in the analysis, and compared the general trend with those reported in the literature [18,19]. We also specifically applied the analytical methods reported in a recent study [13] based on NHANES data to our study population to replicate the trends in preDM among youth in the US from 1999 to 2018 reported in that analysis. Specifically, that study selected a youth population from 12-19 years old with positive sampling weight from the fasting subsample (i.e., non-zero and non-missing WTSAF2YR, personal communication) without a self-reported physician diagnosed DM. In addition, that study focused only on preDM, which was defined as an HbA1c level between 5.7% and 6.4% or a fasting plasma glucose level between 100 mg/dL to 125 mg/dL [13].

Development of youth preDM/DM dataset

Based on the most recent ADA standard of care recommendations including factors related to preDM/DM risk and management [2], we selected 27 potentially relevant NHANES questionnaires and grouped them into four domains: sociodemographic, health status, diet, and other lifestyle behaviors. For example, under the health status domain, body mass index (BMI) was included as a potential risk factor for youth preDM/DM [2]. Similarly, lifestyle and behavioral variables included factors, such as diet and physical activity, that have been shown to be critical for preDM/DM prevention in both observational studies and randomized clinical trials [50–52]. Our sociodemographic domain included demographic, socioeconomic, and SDoH variables (e.g., age, gender, poverty status, and food security). Except for commonly available clinical measurements, such as blood pressure and total cholesterol, we did not include laboratory data (e.g., triglycerides, transferrin, CRP, IL-6, WBC, etc.), since these measurements were not collected for all NHANES participants, and were not commonly accessible for the general population.

From the selected questionnaires, we identified a list of 95 variables based on the above methodology. The complete list of variables are provided in Table S1 in Section S1 of Multimedia Appendix 1 and on our POND web portal [47]. All the code developed, processed data and detailed description of variables are also available on the web portal [47]. The process of extracting these variables involved extensive examination of the questions that were asked, consultation of the literature, and discussions to reach consensus within the study team. The details of this process are provided in Section S2 of Multimedia Appendix 1**.** We used SAS (version 9.4) and R (version 4.2.2; R Core Team, 2022) in R Studio (version 4.2.2; R Core Team, 2022) for data processing and dataset development.

Building the *pre*DM/DM in youth *ON*line *D*ashboard (POND)

To facilitate other researchers’ use of our youth preDM/DM dataset and make our methodology

transparent and reproducible, we developed POND to share our processed dataset and enable users to understand and explore the data on their own. The web portal was developed using R markdown and the flexdashboard package [53], and was published as a Shiny application [54]. Table S2 and Section S3 in Multimedia Appendix 1 provides details of all the R packages used to develop POND, and the related code is available on the portal's download page.

Case studies in using the dataset to better understand youth preDM/DM

To examine the validity and utility of our dataset for advancing translational research on youth preDM/DM, we conducted two complementary data analyses. We first conducted bivariate analyses to assess the statistical associations between each of the 95 variables and youth preDM/DM status. In the second analysis, we used machine learning methods to examine the ability to predict preDM/DM status of youth based on the 95 variables. The methodological details of these analyses are provided below.

Bivariate analyses to identify variables associated with preDM/DM status

We examined associations between individual variables and youth preDM/DM status using Chi-square and Wilcoxon rank sum tests for categorical and continuous variables, respectively. Cell sizes were checked for sufficient size (≥ 5) prior to Chi-square tests. Independence and equal variance were assessed for continuous variables. Distribution normality was assured through adequate sample size in accordance with Central Limit Theorem [55]. We applied Bonferroni correction for multiple hypothesis testing (n=95 tests) at an alpha level of 0.05 to determine the statistical significance of each association at the adjusted alpha level of 0.0005 (i.e., approximately 0.05/95). We used Cramer's V and Wilcoxon R-values [56] as the effect size measures for categorical and continuous variables, respectively. To better compare with results from the machine learning approach, the main bivariate analyses did not account for NHANES survey design; thus, the results were only applicable to the study population included in the analytical sample and were not generalizable to the entire U.S. youth population. For completeness, we provide the survey-weighted analyses using NHANES examination weights (WTMEC2YR) in Section S4 of Multimedia Appendix 1.

Prediction of preDM/DM status using machine learning algorithms

Several machine learning algorithms have been employed to predict adult preDM/DM status using NHANES data [57–59], and we have previously utilized these algorithms to predict preDM/DM status specifically among youth in a subsample of our current study population [42]. We expanded these existing analyses by taking into account the multi-domain nature of our dataset with the goal of building an effective and interpretable predictive model of youth preDM/DM. To that end, we leveraged our recently developed machine learning framework, Ensemble Integration (EI) [60,61], with all four domains and their variables in our dataset. EI incorporates both consensus and complementarity in our dataset by first inferring local predictive models from the individual domains, i.e., sociodemographic, health status, diet, and other lifestyle behaviors, that are expected to capture information and interactions specific to the domains. These local models and information are then integrated into a global preDM/DM, comprehensive preDM/DM prediction model using heterogeneous ensemble algorithms [62] (Figure S2 and Section S5 in Multimedia Appendix 1). These algorithms, such as stacking, allow the integration of an unrestricted number and variety of

local models into the global predictive model, thus offering improved performance and robustness.EI also enables the identification of the most predictive variables in the final model, thus offering deeper insights into the outcome being predicted.

We used both the above capabilities of EI to build and interpret a predictive model of youth preDM/DM status based on our dataset. We also compared the predictive performance of the model with three alternative approaches: (i) a modified form of the ADA screening guideline [63], which is based on BMI, total cholesterol level, hypertension, and race/ethnicity, to assess the utility of data-driven screening for youth preDM/DM, (ii) EI applied to individual variable domains, namely sociodemographic, health status, diet and other lifestyle behaviors, to assess the value of multi-domain data for youth preDM/DM prediction, and (iii) eXtreme Gradient Boosting (XGBoost) [64] applied to our combined multi-domain dataset as a representative alternate machine learning algorithm. This alternative was chosen as XGBoost is considered the most effective classification algorithm for tabular data [65], since it can potentially capture feature interactions across different domains [66,67]. The prediction performance of EI and all the alternative approaches were assessed in terms of the commonly used Area Under the receiver operating characteristic Curve (AUC) [68] and Balanced Accuracy (BA, average of specificity and sensitivity) [69] measures. The performance of the machine learning-based prediction approaches, namely multi- and single-domain EI and XGBoost, were evaluated in a five-fold cross-validation setting repeated ten times [70]. These performance scores were statistically compared using the Wilcoxon rank sum test, and the resultant p-values were corrected for multiple hypothesis testing using the Benjamini-Hochberg procedure to yield false discovery rates (FDRs) [71]. More details of machine learning model building, the alternative approaches and the evaluation methodology, including cross-validation, model selection and comparison, are available in Section S5 in Multimedia Appendix 1**.**

Finally, we used EI's interpretation capabilities [60,61] to identify the variables in our dataset that were the most predictive of youth preDM/DM status and compare them to the variables identified from the bivariate analyses described above.

Results

Study population derived from NHANES

Our study population consisted of 15,149 youth aged 12 to 19 years who participated in the 1999-2018 NHANES cycles and met our selection criteria (Figure 2). Approximately 13.3% of US youth were at risk of preDM/DM according to the clinically standard criteria for defining preDM/DM per ADA guidelines (FPG≥100 mg/dL and/or HbA1c≥5.7%) (Table 1).

Table 1. Unweighted study population characteristics. Unweighted statistics of some key variables describing the study population in the youth preDM/DM dataset overall and by preDM/DM status. More detailed statistics for all the variables in our dataset can be found in the Data Exploration section of POND.

Variables	Overall	With preDM/DM	without preDM/DM
	(N=15,149)	(N=2,010) (Unweighted % = 13.3)	(N=13,139)
	n (%) or median (interquartile range)		
Sociodemographic			

Age (years)	15 (13, 17)	15 (13, 17)	16 (14, 17)
Female Sex	7430 (49.0)	691 (34.4)	6739 (51.3)
Race/ethnicity			
Hispanic	5565 (36.7)	711 (35.4)	4854 (36.9)
White, Non-Hispanic	4033 (26.6)	431 (21.4)	3602 (27.4)
Black, Non-Hispanic	4292 (28.3)	676 (33.6)	3616 (27.5)
Other	1259 (8.3)	192 (9.6)	1067 (8.1)
Insurance			
Private	6392 (43.0)	744 (37.7)	5648 (43.8)
Medicare, government, or single Service	2026 (13.6)	268 (13.6)	1758 (13.6)
Medicaid/CHIP[a]	3637 (24.4)	564 (28.6)	3073 (23.8)
No insurance	2821 (19.0)	395 (20.0)	2426 (18.8)
Authorized for food stamps	7833 (69.4)	1037 (61.1)	6796 (70.8)
Health status			
Body Mass Index (BMI) percentile			
Underweight (BMI %ile < 5th)	462 (3.1)	40 (2.0)	422 (3.2)
Normal weight (5th ≤ BMI %ile < 85th)	8516 (56.8)	933 (46.8)	7583 (58.4)
Overweight (85th ≤ BMI %ile < 95th)	2788 (18.6)	356 (17.9)	2432 (18.7)
Obese (95th ≤ BMI %ile)	3214 (21.5)	663 (33.3)	2551 (19.6)
Hypertensive[b]	2552 (17.4)	502 (26.1)	2050 (16.1)
High total cholesterol (≥ 170 mg/dL)	4951 (33.2)	707 (35.6)	4244 (32.8)
Fasting plasma glucose (mg/dL)	93 (88, 98)	102 (100, 106)	91 (86, 95)
Hemoglobin A1c (%)	5.2 (5.0, 5.4)	5.5 (5.2, 5.7)	5.2 (5.0, 5.3)
Diet			
Meals eaten out per week	2 (1, 3)	2 (1, 3)	2 (1, 3)
Total grain (oz eq.) intake 24 hours prior	6.55 (4.24, 9.66)	6.43 (4.19, 9.58)	6.57 (4.25, 9.67)
Total fruits (cup eq.) intake 24 hours prior	0.38 (0.00, 1.44)	0.26 (0.00, 1.37)	0.40 (0.00, 1.45)
Total vegetable (cup eq.) intake 24 hours prior	0.88 (0.39, 1.58)	0.84 (0.37, 1.54)	0.89 (0.39, 1.59)
Total protein (oz eq.) intake 24 hours prior	5.29 (2.71, 9.15)	4.73 (2.46, 8.37)	5.38 (2.76, 9.34)
Added sugar (tsp eq.) intake 24 hours prior	20.42 (11.49, 32.49)	20.09 (11.15, 31.89)	20.48 (11.57, 32.59)
Other lifestyle behavior			
Physical activity minutes per week	209 (45, 488)	210 (49, 476)	209 (45, 491)
Screen time hours per day	5 (3, 8)	5 (3, 8)	5 (2, 7)
Exposed to secondhand smoke at home	3297 (21.9)	469 (23.6)	2828 (21.7)

[a]CHIP = Child Health Insurance Program.

[b]Hypertensive was defined by blood pressure ≥ 90th percentile or ≥ 120/80 mm Hg for children ≥ 13 years [2].

Validation of the study population

We estimated that the survey-weighted prevalence of preDM/DM in our study population rose substantially from 4.1% (95% confidence interval (CI) 2.8-5.4) in 1999 to 22.0% (95% CI 18.5-25.6) in 2018 (Figure S3 in Multimedia Appendix 1). This increasing trend of preDM/DM prevalence was consistent with that reported in other NHANES-based studies, which had preDM/DM prevalence ranging from 17.7% to 18.0% [18,19]. We also applied the study population and preDM definition criteria reported in a recent study [13] to NHANES data, and derived a similarly sized study

population (n=6,656 vs n=6,598 in the current vs previous analysis [13]) and youth preDM prevalence, which ranged from 11.1% (95% CI 8.9-13.3) to 37.3% (95% CI 31.0-43.6) in our analysis, compared to 11.6% (95% CI 9.5-14.1) to 28.2% (95% CI 23.3-33.6) in Liu et al. [13] (Table S6 in Multimedia Appendix 1)**.**

Youth preDM/DM-focused dataset

We extracted 95 epidemiological variables from NHANES, and organized them into four preDM/DM-related domains, namely sociodemographic, health status, diet, and other lifestyle behaviors (Table S1 in Multimedia Appendix 1). Table 1 shows the unweighted statistics of some key study population characteristics. Among youth with preDM/DM, the proportion of youth who were Non-Hispanic Black, Non-Hispanic White, Hispanic and other race/ethnicity (including Non-Hispanic persons that reported races other than Black or White and Non-Hispanic Asian) were 33.6%, 27.4%, 35.4%, and 9.6%, respectively. Approximately, half of the population were males, and they represented 65.6% of those with preDM/DM. Approximately 32.4% of the youth had a family income below poverty level, and 69.4% were from households receiving food stamps. The proportion of youth covered by private insurance was higher among those with than without preDM/DM (43.8% vs 37.7%). Overall, 21.5% of the youth were obese as defined by having a BMI at or above the 95th percentile based on age and gender, and the proportion was 33.3% among youth with preDM/DM. Youth with preDM/DM tended to have less fruit and vegetable intake and ate lower amounts of protein and total grains than those without. Youth with and without preDM/DM showed similar amounts of physical activity with 209 and 210 minutes per week, respectively (Table 1).

PreDM/DM in youth ONline Dashboard (POND)

To facilitate other researchers' use of our youth preDM/DM dataset and make our methodology transparent and reproducible, we developed POND, which is available on [47]. Users can navigate POND through its built-in functionalities. For example, users are able to explore the details of the 95 individual variables and their distributions by preDM/DM status, as well as examine the risk factors of youth preDM/DM identified from the case studies described below (Figure 3). POND also allows users to easily download the data to conduct their own analyses and explore other youth preDM/DM-related research questions. In addition, we make available all the code used to develop the dataset, our case studies, and POND itself.

Case studies using our dataset to better understand youth preDM/DM

We examined the validity and utility of our processed multi-domain dataset for translational studies on youth preDM/DM by the following two complementary types of data analyses.

Identifying individual variables associated with preDM/DM status

In our bivariate analyses, we found 27 variables to be significantly ($P \leq 0.0005$, Bonferonni adjusted) associated with preDM/DM status (Figure 4, Table S7 in Multimedia Appendix 1). These variables spanned all four domains, and included gender, race/ethnicity, use of food stamps, health insurance status, BMI, total protein intake and screen time. Similar results were found when repeating these bivariate association tests after accounting for NHANES survey design elements (Table S7 in Multimedia Appendix 1).

Predicting youth preDM/DM status with machine learning

We used a machine learning framework, Ensemble Integration (EI) [60,61], to leverage the multi-

domain nature of our dataset, and predict youth preDM/DM status. We also compared EI's performance with alternative prediction approaches, most prominently the widely used XGBoost algorithm [64].

The best-performing multi-domain EI methodology, stacking [68] using Logistic Regression, predicted youth preDM/DM status (AUC=0.67, BA=0.62) more accurately than all the alternative approaches (Figure 5), namely XGBoost (AUC=0.64, BA=0.60, Wilcoxon rank-sum FDR=$1.7x10^{-4}$ and $1.8x10^{-4}$, respectively), the ADA pediatric screening guidelines (AUC=0.57, BA=0.57; Wilcoxon rank-sum FDR=$1.7x10^{-4}$ and $1.8x10^{-4}$, respectively), and four single-domain EI (AUC=0.63-0.54, BA=0.60-0.53; FDR<$1.7x10^{-4}$ and $1.8x10^{-4}$, respectively).

The multi-domain EI also identified 27 variables (the same as the number of significant variables from bivariate analyses) that contributed the most to predicting youth preDM/DM status. Among these variables, 16 overlapped with those identified from the bivariate statistical analyses (Figure 6; Fisher's P of overlap=$7.06x10^{-6}$). These variables identified by both approaches included some established preDM/DM risk factors like BMI and high total cholesterol, as well as some less-recognized ones like screen time and taking prescription drugs [2].

Discussion

Principal Results

Leveraging the rich information in NHANES spanning nearly 20 years, we built the most comprehensive epidemiological dataset for studying youth preDM/DM. We accomplished this by selecting and harmonizing variables relevant to youth preDM/DM from sociodemographic, health status, diet and other lifestyle behaviors domains. This youth preDM/DM dataset, as well as several functionalities to explore and analyze it, are publicly available in our user-friendly web portal, POND. We also conducted case studies using the dataset with both traditional statistical methods and machine learning approaches to demonstrate the potential of using this dataset to identify factors relevant to youth preDM/DM. The combination of the comprehensive public dataset and POND provide avenues for more informed investigations of youth preDM/DM.

The future translational impact of preDM/DM research, facilitated by comprehensive datasets like the one developed in this study, holds significant promise for advancing our understanding of the disease and its risk factors among youth. By enabling researchers to investigate multifactorial variables associated with preDM/DM, this dataset contribute to several areas of research and has a broader impact on the scientific community. Firstly, the dataset's comprehensive nature allows researchers to explore the collective impact of various risk factors across multiple health domains. By incorporating sociodemographic factors, health status indicators, diet, and lifestyle behaviors, researchers can gain a holistic understanding of the interplay between these factors and preDM/DM risk among youth. This knowledge can be used to generate hypotheses for further studies and inform the development of targeted interventions and prevention strategies that address the specific needs of at-risk populations. Furthermore, the dataset provides an opportunity to delve into less-studied variables and their interactions in relation to preDM/DM risk. Variables such as screen time, acculturation, or frequency of eating out, which are often overlooked in traditional research, can be examined to uncover their potential influence on preDM/DM risk among youth. This expands the scope of translational research and enhances our understanding of the multifaceted nature of the disease.

One of the major contributions of our work was POND, our publicly available web portal, which

provided access to all materials related to our dataset and analyses, thus enabling transparency and reproducibility. Although several such portals are available in other biomedical areas, such as genomics [69–71], there is a general lack of such tools in epidemiology and public health. We hope that, in addition to facilitating studies into preDM/DM, POND illustrates the utility of such portals for population and epidemiological studies as well.

The results of the case studies and validation exercises we conducted were also consistent with existing literature. The case studies identified known preDM/DM risk factors, such as gender [15,17,19], race/ethnicity [2,9,10,24], health measures (BMI, hypertension and cholesterol) [2,63], income [9,11], insurance status [9,10] and healthcare availability [9,10], thus affirming the validity of the dataset. In addition, our analyses revealed some less studied variables, such as screen time, home ownership status, self-reported health status, soy and nut consumption, and frequency of school meal intake, that may influence youth preDM/DM risk. Further study of these variables may reveal new knowledge about preDM/DM among youth. More generally, such novel findings further demonstrate the utility of our dataset and data-driven methods for further translational discoveries about this complex disorder.

Limitations

Although our work has several strengths and high potential utility for youth preDM/DM studies, it is not without limitations. First, as our dataset was derived from NHANES, we adopt limitations to the survey in our dataset. Since NHANES is a cross-sectional survey, the preDM/DM status and its related variables only provide consecutive snapshots of youth in the U.S. over time across the available survey cycles. Thus, and the associations identified are better suited for hypothesis generation purposes, and require in-depth investigation using prospective longitudinal and randomized trial designs. Additionally, we modified the ADA guideline for determining preDM/DM status according to variable availability. Due to the high missingness of 45% in family history (DIQ170) and the complete missingness of maternal history (DIQ175S) from 1999-2010 in the raw NHANES data, we were unable to include family history of diabetes in the dataset. Similarly, NHANES does not provide data regarding every condition associated with insulin resistance. Therefore, we used hypertension and high cholesterol as proxies for insulin resistance. On the other hand, as our main purpose is to use POND as a conduit between this comprehensive youth preDM/DM database and interested researchers, our method can be adopted to longitudinal data sets should they become available in the future. Second, for the prediction of preDM/DM status, EI's performance was found to be significantly better than the alternative approaches, including a modified form of the suggested guideline [45]. However, this performance assessment was only based on cross-validation, which is no substitute for validation on external datasets that is necessary for rigorous assessment. Finally, while our preliminary case study analyses identified a wide range of variables associated with youth prediabetes and diabetes, other known risk factors, such as current asthma status [72–74], added sugar consumption [75–77], sugary fruit and juice intake [75–78], and physical activity per week [6–8,50], were not identified. This limitation can be addressed by employing other data analysis methods beyond our bivariate testing and machine learning approaches, highlighting more potential use cases of our dataset.

Conclusions

Overall, the future impact of translational preDM/DM research facilitated by comprehensive datasets and web servers like ours extends beyond individual studies. It creates opportunities for interdisciplinary collaboration and reproducibility, strengthens evidence-based decision-making, and supports the development of targeted interventions for the prevention and management of preDM/DM among youth. By providing rich resources, our work can enable researchers to build

upon existing knowledge and push the boundaries of translational preDM/DM research, ultimately leading to improved health outcomes for at-risk populations.

Acknowledgements

This study was enabled in part by computational resources provided by Scientific Computing and Data at the Icahn School of Medicine at Mount Sinai. The Ensemble Integration used in this work was implemented by Jamie J.R. Bennett. This work was funded by National Institutes of Health grant #s R21DK131555 and R01HG011407.

Data Availability

The dataset and code used in this study are available online on Zenodo [79] and our web portal POND [47].

Author Contributions

NV, BL and GP conceptualized the project. CM, YCL, NV, BL and GP designed the methodology. CM and BL implemented the data curation and bivariate analyses. YCL implemented the machine learning case study and POND. CM and YCL conducted formal analysis and visualization. CM, YCL, NV, BL and GP wrote the manuscript. NV, BL and GP supervised the project.

Conflicts of Interest

None declared.

Abbreviations

ADA: American Diabetes Association
AUC: Area Under the receiver operating characteristic Curve
BA: Balanced Accuracy
BMI: body mass index
CI: confidence interval
DM: type 2 diabetes mellitus
EI: Ensemble Integration
FDR: false discovery rate
NHANES: National Health and Nutrition Examination Survey
POND: prediabetes/diabetes in youth online dashboard
preDM/DM: prediabetes/diabetes mellitus
SDoH: social determinants of health
WTSAF2YR: Fasting Subsample 2 Year Mobile Examination Centers Weight
XGBoost: eXtreme Gradient Boosting

References

1. Temneanu OR, Trandafir LM, Purcarea MR. Type 2 diabetes mellitus in children and adolescents: a relatively new clinical problem within pediatric practice. J Med Life

2016;9(3):235–239. PMID:27974926
2. ElSayed NA, Aleppo G, Aroda VR, Bannuru RR, Brown FM, Bruemmer D, Collins BS, Hilliard ME, Isaacs D, Johnson EL, Kahan S, Khunti K, Leon J, Lyons SK, Perry ML, Prahalad P, Pratley RE, Seley JJ, Stanton RC, Gabbay RA. 2. Classification and Diagnosis of Diabetes: Standards of Care in Diabetes—2023. Diabetes Care 2023 Jan;46(Suppl 1):S19–S40. PMID:36507649
3. Weiss R, Dufour S, Taksali SE, Tamborlane WV, Petersen KF, Bonadonna RC, Boselli L, Barbetta G, Allen K, Rife F, Savoye M, Dziura J, Sherwin R, Shulman GI, Caprio S. Prediabetes in obese youth: a syndrome of impaired glucose tolerance, severe insulin resistance, and altered myocellular and abdominal fat partitioning. The Lancet 2003 Sep;362(9388):951–957. doi: 10.1016/S0140-6736(03)14364-4
4. Zhang Y, Luk AOY, Chow E, Ko GTC, Chan MHM, Ng M, Kong APS, Ma RCW, Ozaki R, So WY, Chow CC, Chan JCN. High risk of conversion to diabetes in first-degree relatives of individuals with young-onset type 2 diabetes: a 12-year follow-up analysis. Diabet Med J Br Diabet Assoc 2017 Dec;34(12):1701–1709. PMID:28945282
5. Zhuang P, Liu X, Li Y, Wan X, Wu Y, Wu F, Zhang Y, Jiao J. Effect of Diet Quality and Genetic Predisposition on Hemoglobin A1c and Type 2 Diabetes Risk: Gene-Diet Interaction Analysis of 357,419 Individuals. Diabetes Care 2021 Nov 1;44(11):2470–2479. doi: 10.2337/dc21-1051
6. Pivovarov JA, Taplin CE, Riddell MC. Current perspectives on physical activity and exercise for youth with diabetes: Perspectives on exercise. Pediatr Diabetes 2015 Jun;16(4):242–255. doi: 10.1111/pedi.12272
7. Colberg SR, Sigal RJ, Yardley JE, Riddell MC, Dunstan DW, Dempsey PC, Horton ES, Castorino K, Tate DF. Physical Activity/Exercise and Diabetes: A Position Statement of the American Diabetes Association. Diabetes Care 2016 Nov;39(11):2065–2079. PMID:27926890
8. Rietz M, Lehr A, Mino E, Lang A, Szczerba E, Schiemann T, Herder C, Saatmann N, Geidl W, Barbaresko J, Neuenschwander M, Schlesinger S. Physical Activity and Risk of Major Diabetes-Related Complications in Individuals With Diabetes: A Systematic Review and Meta-Analysis of Observational Studies. Diabetes Care 2022 Dec 1;45(12):3101–3111. doi: 10.2337/dc22-0886
9. Hill-Briggs F, Adler NE, Berkowitz SA, Chin MH, Gary-Webb TL, Navas-Acien A, Thornton PL, Haire-Joshu D. Social Determinants of Health and Diabetes: A Scientific Review. Diabetes Care 2021 Jan;44(1):258–279. PMID:33139407
10. Butler AM. Social Determinants of Health and Racial/Ethnic Disparities in Type 2 Diabetes in Youth. Curr Diab Rep 2017 Aug;17(8):60. doi: 10.1007/s11892-017-0885-0
11. Walker RJ, Smalls BL, Campbell JA, Strom Williams JL, Egede LE. Impact of Social Determinants of Health on Outcomes for Type 2 Diabetes: A Systematic Review. Endocrine 2014 Sep;47(1):29–48. PMID:24532079
12. Bansal N. Prediabetes diagnosis and treatment: A review. World J Diabetes 2015;6(2):296. doi: 10.4239/wjd.v6.i2.296
13. Liu J, Li Y, Zhang D, Yi SS, Liu J. Trends in Prediabetes Among Youths in the US From 1999 Through 2018. JAMA Pediatr 2022 Jun 1;176(6):608–611. PMID:35344013
14. Tönnies T, Brinks R, Isom S, Dabelea D, Divers J, Mayer-Davis EJ, Lawrence JM, Pihoker C, Dolan L, Liese AD, Saydah SH, Jr. RBD, Hoyer A, Imperatore G. Projections of type 1 and type 2 diabetes burden in the US population aged <20 years through 2060: The SEARCH for Diabetes in Youth Study. 2022 Dec. doi: 10.2337/figshare.21514014

15. Lawrence JM, Divers J, Isom S, Saydah S, Imperatore G, Pihoker C, Marcovina SM, Mayer-Davis EJ, Hamman RF, Dolan L, Dabelea D, Pettitt DJ, Liese AD, SEARCH for Diabetes in Youth Study Group. Trends in Prevalence of Type 1 and Type 2 Diabetes in Children and Adolescents in the US, 2001-2017. JAMA 2021 Aug 24;326(8):717. doi: 10.1001/jama.2021.11165
16. Jensen ET, Dabelea D. Type 2 Diabetes in Youth: New Lessons from the SEARCH Study. Curr Diab Rep 2018 Jun;18(6):36. doi: 10.1007/s11892-018-0997-1
17. Dabelea D, Mayer-Davis EJ, Saydah S, Imperatore G, Linder B, Divers J, Bell R, Badaru A, Talton JW, Crume T, Liese AD, Merchant AT, Lawrence JM, Reynolds K, Dolan L, Liu LL, Hamman RF. Prevalence of Type 1 and Type 2 Diabetes Among Children and Adolescents From 2001 to 2009. JAMA 2014 May 7;311(17):1778. doi: 10.1001/jama.2014.3201
18. Andes LJ, Cheng YJ, Rolka DB, Gregg EW, Imperatore G. Prevalence of Prediabetes Among Adolescents and Young Adults in the United States, 2005-2016. JAMA Pediatr 2020 Feb 1;174(2):e194498. PMID:31790544
19. Menke A, Casagrande S, Cowie CC. Prevalence of Diabetes in Adolescents Aged 12 to 19 Years in the United States, 2005-2014. JAMA 2016 Jul 19;316(3):344–345. PMID:27434447
20. Khan MAB, Hashim MJ, King JK, Govender RD, Mustafa H, Al Kaabi J. Epidemiology of Type 2 Diabetes – Global Burden of Disease and Forecasted Trends. J Epidemiol Glob Health 2020 Mar;10(1):107–111. PMID:32175717
21. Lin X, Xu Y, Pan X, Xu J, Ding Y, Sun X, Song X, Ren Y, Shan P-F. Global, regional, and national burden and trend of diabetes in 195 countries and territories: an analysis from 1990 to 2025. Sci Rep 2020 Sep 8;10:14790. PMID:32901098
22. Imperatore G, Boyle JP, Thompson TJ, Case D, Dabelea D, Hamman RF, Lawrence JM, Liese AD, Liu LL, Mayer-Davis EJ, Rodriguez BL, Standiford D, SEARCH for Diabetes in Youth Study Group. Projections of type 1 and type 2 diabetes burden in the U.S. population aged <20 years through 2050: dynamic modeling of incidence, mortality, and population growth. Diabetes Care 2012 Dec;35(12):2515–2520. PMID:23173134
23. Herman WH, Ma Y, Uwaifo G, Haffner S, Kahn SE, Horton ES, Lachin JM, Montez MG, Brenneman T, Barrett-Connor E, for the Diabetes Prevention Program Research Group. Differences in A1C by Race and Ethnicity Among Patients With Impaired Glucose Tolerance in the Diabetes Prevention Program. Diabetes Care 2007 Oct 1;30(10):2453–2457. doi: 10.2337/dc06-2003
24. Kahkoska AR, Shay CM, Crandell J, Dabelea D, Imperatore G, Lawrence JM, Liese AD, Pihoker C, Reboussin BA, Agarwal S, Tooze JA, Wagenknecht LE, Zhong VW, Mayer-Davis EJ. Association of Race and Ethnicity With Glycemic Control and Hemoglobin A1c Levels in Youth With Type 1 Diabetes. JAMA Netw Open 2018 Sep 7;1(5):e181851. PMID:30370425
25. Lascar N, Brown J, Pattison H, Barnett AH, Bailey CJ, Bellary S. Type 2 diabetes in adolescents and young adults. Lancet Diabetes Endocrinol 2018 Jan;6(1):69–80. PMID:28847479
26. Lee AM, Fermin CR, Filipp SL, Gurka MJ, DeBoer MD. Examining trends in prediabetes and its relationship with the metabolic syndrome in US adolescents, 1999-2014. Acta Diabetol 2017 Apr;54(4):373–381. PMID:28070750
27. Weiss R, Taksali SE, Tamborlane WV, Burgert TS, Savoye M, Caprio S. Predictors of changes in glucose tolerance status in obese youth. Diabetes Care 2005 Apr;28(4):902–909. PMID:15793193

28. Nadeau KJ, Anderson BJ, Berg EG, Chiang JL, Chou H, Copeland KC, Hannon TS, Huang TT-K, Lynch JL, Powell J, Sellers E, Tamborlane WV, Zeitler P. Youth-Onset Type 2 Diabetes Consensus Report: Current Status, Challenges, and Priorities. Diabetes Care 2016 Sep 1;39(9):1635–1642. doi: 10.2337/dc16-1066
29. Dart AB, Martens PJ, Rigatto C, Brownell MD, Dean HJ, Sellers EA. Earlier Onset of Complications in Youth With Type 2 Diabetes. Diabetes Care 2014 Feb 1;37(2):436–443. doi: 10.2337/dc13-0954
30. American Diabetes Association. Economic Costs of Diabetes in the U.S. in 2017. Diabetes Care 2018 May 1;41(5):917–928. doi: 10.2337/dci18-0007
31. Al-Goblan AS, Al-Alfi MA, Khan MZ. Mechanism linking diabetes mellitus and obesity. Diabetes Metab Syndr Obes Targets Ther 2014 Dec 4;7:587–591. PMID:25506234
32. Chan JCN, Lim L-L, Wareham NJ, Shaw JE, Orchard TJ, Zhang P, Lau ESH, Eliasson B, Kong APS, Ezzati M, Aguilar-Salinas CA, McGill M, Levitt NS, Ning G, So W-Y, Adams J, Bracco P, Forouhi NG, Gregory GA, Guo J, Hua X, Klatman EL, Magliano DJ, Ng B-P, Ogilvie D, Panter J, Pavkov M, Shao H, Unwin N, White M, Wou C, Ma RCW, Schmidt MI, Ramachandran A, Seino Y, Bennett PH, Oldenburg B, Gagliardino JJ, Luk AOY, Clarke PM, Ogle GD, Davies MJ, Holman RR, Gregg EW. The Lancet Commission on diabetes: using data to transform diabetes care and patient lives. The Lancet 2020 Dec;396(10267):2019–2082. doi: 10.1016/S0140-6736(20)32374-6
33. International Diabetes Federation. IDF Diabetes Atlas, 10th Edition. Available from: https://diabetesatlas.org/
34. U.S. Chronic Disease Indicators: Diabetes | Chronic Disease and Health Promotion Data & Indicators. Available from: https://chronicdata.cdc.gov/Chronic-Disease-Indicators/U-S-Chronic-Disease-Indicators-Diabetes/f8ti-h92k [accessed May 17, 2023]
35. NCD Risk Factor Collaboration. Available from: https://ncdrisc.org/index.html [accessed May 17, 2023]
36. Zhou B, Lu Y, Hajifathalian K, Bentham J, Cesare MD, Danaei G, Bixby H, Cowan MJ, Ali MK, Taddei C, Lo WC, Reis-Santos B, Stevens GA, Riley LM, Miranda JJ, Bjerregaard P, Rivera JA, Fouad HM, Ma G, Mbanya JC, McGarvey ST, Mohan V, Onat A, Pilav A, Ramachandran A, Romdhane HB, Paciorek CJ, Bennett JE, Ezzati M, Abdeen ZA, Kadir KA, Abu-Rmeileh NM, Acosta-Cazares B, Adams R, Aekplakorn W, Aguilar-Salinas CA, Agyemang C, Ahmadvand A, Al-Othman AR, Alkerwi A, Amouyel P, Amuzu A, Andersen LB, Anderssen SA, Anjana RM, Aounallah-Skhiri H, Aris T, Arlappa N, Arveiler D, Assah FK, Avdicová M, Azizi F, Balakrishna N, Bandosz P, Barbagallo CM, Barceló A, Batieha AM, Baur LA, Romdhane HB, Benet M, Bernabe-Ortiz A, Bharadwaj S, Bhargava SK, Bi Y, Bjerregaard P, Bjertness E, Bjertness MB, Björkelund C, Blokstra A, Bo S, Boehm BO, Boissonnet CP, Bovet P, Brajkovich I, Breckenkamp J, Brenner H, Brewster LM, Brian GR, Bruno G, Bugge A, León AC de, Can G, Cândido AP, Capuano V, Carlsson AC, Carvalho MJ, Casanueva FF, Casas JP, Caserta CA, Castetbon K, Chamukuttan S, Chaturvedi N, Chen CJ, Chen F, Chen S, Cheng CY, Chetrit A, Chiou ST, Cho Y, Chudek J, Cifkova R, Claessens F, Concin H, Cooper C, Cooper R, Costanzo S, Cottel D, Cowell C, Crujeiras AB, D'Arrigo G, Dallongeville J, Dankner R, Dauchet L, Gaetano G de, Henauw SD, Deepa M, Dehghan A, Deschamps V, Dhana K, Castelnuovo AD, Djalalinia S, Doua K, Drygas W, Du Y, Dzerve V, Egbagbe EE, Eggertsen R, Ati JE, Elosua R, Erasmus RT, Erem C, Ergor G, Eriksen L, Peña JE la, Fall CH, Farzadfar F, Felix-Redondo FJ, Ferguson TS, Fernández-Bergés D, Ferrari M, Ferreccio C,

Feskens EJ, Finn JD, Föger B, Foo LH, Forslund AS, Fouad HM, Francis DK, Mdo CF, Franco OH, Frontera G, Furusawa T, Gaciong Z, Garnett SP, Gaspoz JM, Gasull M, Gates L, Geleijnse JM, Ghasemian A, Ghimire A, Giampaoli S, Gianfagna F, Giovannelli J, Giwercman A, Gross MG, Rivas JG, Gorbea MB, Gottrand F, Grafnetter D, Grodzicki T, Grøntved A, Gruden G, Gu D, Guan OP, Guerrero R, Guessous I, Guimaraes AL, Gutierrez L, Hambleton IR, Hardy R, Kumar RH, Hata J, He J, Heidemann C, Herrala S, Hihtaniemi IT, Ho SY, Ho SC, Hofman A, Hormiga CM, Horta BL, Houti L, Howitt C, Htay TT, Htet AS, Htike MM, Hu Y, Hussieni AS, Huybrechts I, Hwalla N, Iacoviello L, Iannone AG, Ibrahim MM, Ikeda N, Ikram MA, Irazola VE, Islam M, Iwasaki M, Jacobs JM, Jafar T, Jamil KM, Jasienska G, Jiang CQ, Jonas JB, Joshi P, Kafatos A, Kalter-Leibovici O, Kasaeian A, Katz J, Kaur P, Kavousi M, Keinänen-Kiukaanniemi S, Kelishadi R, Kengne AP, Kersting M, Khader YS, Khalili D, Khang YH, Kiechl S, Kim J, Kolsteren P, Korrovits P, Kratzer W, Kromhout D, Kujala UM, Kula K, Kyobutungi C, Laatikainen T, Lachat C, Laid Y, Lam TH, Landrove O, Lanska V, Lappas G, Laxmaiah A, Leclercq C, Lee J, Lehtimäki T, Lekhraj R, León-Muñoz LM, Li Y, Lim WY, Lima-Costa MF, Lin HH, Lin X, Lissner L, Lorbeer R, Lozano JE, Luksiene D, Lundqvist A, Lytsy P, Ma G, Machado-Coelho GL, Machi S, Maggi S, Magliano DJ, Makdisse M, Rao KM, Manios Y, Manzato E, Margozzini P, Marques-Vidal P, Martorell R, Masoodi SR, Mathiesen EB, Matsha TE, Mbanya JC, McFarlane SR, McGarvey ST, McLachlan S, McNulty BA, Mediene-Benchekor S, Meirhaeghe A, Menezes AM, Merat S, Meshram II, Mi J, Miquel JF, Miranda JJ, Mohamed MK, Mohammad K, Mohammadifard N, Mohan V, Yusoff MM, Møller NC, Molnár D, Mondo CK, Morejon A, Moreno LA, Morgan K, Moschonis G, Mossakowska M, Mostafa A, Mota J, Motta J, Mu TT, Muiesan ML, Müller-Nurasyid M, Mursu J, Nagel G, Námešná J, Nang EE, NangThetia VB, Navarrete-Muñoz EM, Ndiaye NC, Nenko I, Nervi F, Nguyen ND, Nguyen QN, Nieto-Martínez RE, Ning G, Ninomiya T, Noale M, Noto D, Nsour MA, Ochoa-Avilés AM, Oh K, Onat A, Ordunez P, Osmond C, Otero JA, Owusu-Dabo E, Pahomova E, Palmieri L, Panda-Jonas S, Panza F, Parsaeian M, Peixoto SV, Pelletier C, Peltonen M, Peters A, Peykari N, Pham ST, Pilav A, Pitakaka F, Piwonska A, Piwonski J, Plans-Rubió P, Porta M, Portegies ML, Poustchi H, Pradeepa R, Price JF, Punab M, Qasrawi RF, Qorbani M, Radisauskas R, Rahman M, Raitakari O, Rao SR, Ramachandran A, Ramke J, Ramos R, Rampal S, Rathmann W, Redon J, Reganit PF, Rigo F, Robinson SM, Robitaille C, Rodríguez-Artalejo F, Mdel CR-P, Rodríguez-Villamizar LA, Rojas-Martinez R, Ronkainen K, Rosengren A, Rubinstein A, Rui O, Ruiz-Betancourt BS, Horimoto RR, Rutkowski M, Sabanayagam C, Sachdev HS, Saidi O, Sakarya S, Salanave B, Salonen JT, Salvetti M, Sánchez-Abanto J, Santos D, Santos R dos, Santos R, Saramies JL, Sardinha LB, Sarrafzadegan N, Saum KU, Scazufca M, Schargrodsky H, Scheidt-Nave C, Sein AA, Sharma SK, Shaw JE, Shibuya K, Shin Y, Shiri R, Siantar R, Sibai AM, Simon M, Simons J, Simons LA, Sjostrom M, Slowikowska-Hilczer J, Slusarczyk P, Smeeth L, Snijder MB, So HK, Sobngwi E, Söderberg S, Solfrizzi V, Sonestedt E, Soumare A, Staessen JA, Stathopoulou MG, Steene-Johannessen J, Stehle P, Stein AD, Stessman J, Stöckl D, Stokwiszewski J, Stronks K, Strufaldi MW, Sun CA, Sundström J, Sung YT, Suriyawongpaisal P, Sy RG, Tai ES, Tamosiunas A, Tang L, Tarawneh M, Tarqui-Mamani CB, Taylor A, Theobald H, Thijs L, Thuesen BH, Tolonen HK, Tolstrup JS, Topbas M, Torrent M, Traissac P, Trinh OT, Tulloch-Reid MK, Tuomainen TP, Turley ML, Tzourio C, Ueda P, Ukoli FA, Ulmer H, Uusitalo HM, Valdivia G, Valvi D, Rossem L van, Valkengoed I van, Vanderschueren D, Vanuzzo D, Vega T, Velasquez-Melendez G, Veronesi G, Verschuren WM, Verstraeten R, Viet L, Vioque J, Virtanen JK, Visvikis-Siest S, Viswanathan B,

Vollenweider P, Voutilainen S, Vrijheid M, Wade AN, Wagner A, Walton J, Mohamud WW, Wang F, Wang MD, Wang Q, Wang YX, Wannamethee SG, Weerasekera D, Whincup PH, Widhalm K, Wiecek A, Wijga AH, Wilks RJ, Willeit J, Wilsgaard T, Wojtyniak B, Wong TY, Woo J, Woodward M, Wu FC, Wu SL, Xu H, Yan W, Yang X, Ye X, Yoshihara A, Younger-Coleman NO, Zambon S, Zargar AH, Zdrojewski T, Zhao W, Zheng Y, Cisneros JZ. Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4·4 million participants. The Lancet Elsevier; 2016 Apr 9;387(10027):1513–1530. PMID:27061677

37. UCI Machine Learning Repository: Diabetes 130-US hospitals for years 1999-2008 Data Set. Available from: https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008 [accessed May 20, 2023]
38. Type 2 Diabetes Knowledge Portal. Available from: https://t2d.hugeamp.org/ [accessed May 17, 2023]
39. Rashid A. Diabetes Dataset. Mendeley Data; 2020 Jul 18;1. doi: 10.17632/wj9rwkp9c2.1
40. Diabetes Dataset 2019. Available from: https://www.kaggle.com/datasets/tigganeha4/diabetes-dataset-2019 [accessed May 20, 2023]
41. Diabetes Health Indicators Dataset. Available from: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset [accessed May 17, 2023]
42. Vangeepuram N, Liu B, Chiu P, Wang L, Pandey G. Predicting youth diabetes risk using NHANES data and machine learning. Sci Rep Nature Publishing Group; 2021 May 27;11(1):11212. doi: 10.1038/s41598-021-90406-0
43. Nagarajan S, Khokhar A, Holmes DS, Chandwani S. Family Consumer Behaviors, Adolescent Prediabetes and Diabetes in the National Health and Nutrition Examination Survey (2007-2010). J Am Coll Nutr 2017;36(7):520–527. PMID:28853988
44. Wallace AS, Wang D, Shin J-I, Selvin E. Screening and Diagnosis of Prediabetes and Diabetes in US Children and Adolescents. Pediatrics 2020 Sep;146(3):e20200265. PMID:32778539
45. Chu P, Patel A, Helgeson V, Goldschmidt AB, Ray MK, Vajravelu ME. Perception and Awareness of Diabetes Risk and Reported Risk-Reducing Behaviors in Adolescents. JAMA Netw Open 2023 May 3;6(5):e2311466. doi: 10.1001/jamanetworkopen.2023.11466
46. Patel CJ, Pho N, McDuffie M, Easton-Marks J, Kothari C, Kohane IS, Avillach P. A database of human exposomes and phenomes from the US National Health and Nutrition Examination Survey. Sci Data Nature Publishing Group; 2016 Oct 25;3(1):160096. doi: 10.1038/sdata.2016.96
47. PreDM/DM in youth ONline Dashboard (POND). Available from: https://rstudio-connect.hpc.mssm.edu/POND/ [accessed Feb 2, 2024]
48. Zipf G, Chiappa M, Porter KS, Ostchega Y, Lewis BG, Dostal J. National health and nutrition examination survey: plan and operations, 1999-2010. Vital Health Stat Ser 1 Programs Collect Proced 2013 Aug;(56):1–37. PMID:25078429
49. NHANES - NCHS Research Ethics Review Board Approval. 2022. Available from: https://www.cdc.gov/nchs/nhanes/irba98.htm [accessed Jan 19, 2024]
50. Sampath Kumar A, Maiya AG, Shastry BA, Vaishali K, Ravishankar N, Hazari A, Gundmi S, Jadhav R. Exercise and insulin resistance in type 2 diabetes mellitus: A systematic review and meta-analysis. Ann Phys Rehabil Med 2019 Mar;62(2):98–103. PMID:30553010
51. Karstoft K, Winding K, Knudsen SH, Nielsen JS, Thomsen C, Pedersen BK, Solomon TPJ. The Effects of Free-Living Interval-Walking Training on Glycemic Control, Body Composition, and

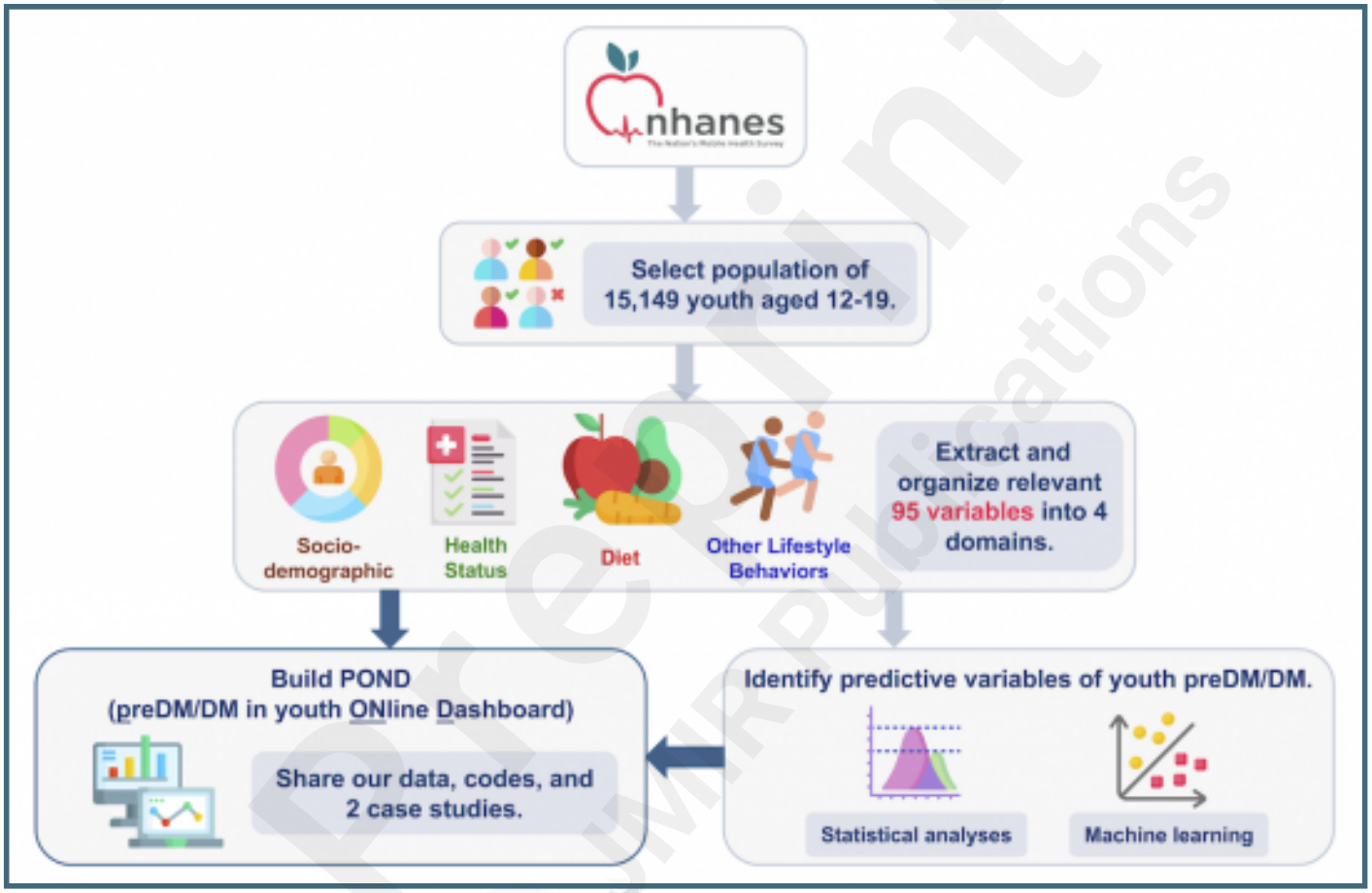
Physical Fitness in Type 2 Diabetic Patients. Diabetes Care 2013 Feb;36(2):228–236. PMID:23002086
52. Karstoft K, Christensen CS, Pedersen BK, Solomon TPJ. The acute effects of interval- Vs continuous-walking exercise on glycemic control in subjects with type 2 diabetes: a crossover, controlled study. J Clin Endocrinol Metab 2014 Sep;99(9):3334–3342. PMID:24905068
53. R Markdown Format for Flexible Dashboards. Available from: https://pkgs.rstudio.com/flexdashboard/ [accessed May 18, 2023]
54. Shiny - Welcome to Shiny. Available from: https://shiny.posit.co/r/getstarted/shiny-basics/lesson1/index.html [accessed May 18, 2023]
55. Kwak SG, Kim JH. Central limit theorem: the cornerstone of modern statistics. Korean J Anesthesiol 2017 Apr;70(2):144–156. PMID:28367284
56. Tomczak M, Tomczak E. The need to report effect size estimates revisited. An overview of some recommended measures of effect size. Trends Sport Sci 2014 Feb 15;1(21):19–25.
57. Herman WH, Smith PJ, Thompson TJ, Engelgau MM, Aubert RE. A new and simple questionnaire to identify people at increased risk for undiagnosed diabetes. Diabetes Care 1995 Mar;18(3):382–387. PMID:7555482
58. Bang H, Edwards AM, Bomback AS, Ballantyne CM, Brillon D, Callahan MA, Teutsch SM, Mushlin AI, Kern LM. Development and validation of a patient self-assessment score for diabetes risk. Ann Intern Med 2009 Dec 1;151(11):775–783. PMID:19949143
59. Poltavskiy E, Kim DJ, Bang H. Comparison of screening scores for diabetes and prediabetes. Diabetes Res Clin Pract 2016 Aug;118:146–153. PMID:27371780
60. Li YC, Wang L, Law JN, Murali TM, Pandey G. Integrating multimodal data through interpretable heterogeneous ensembles. Bioinforma Adv 2022 Jan 1;2(1):vbac065. doi: 10.1093/bioadv/vbac065
61. Bennett JJR, Li YC, Pandey G. eipy: An Open-Source Python Package for Multi-modal Data Integration using Heterogeneous Ensembles. arXiv; 2024. Available from: http://arxiv.org/abs/2401.09582 [accessed Jan 19, 2024]
62. Whalen S, Pandey OP, Pandey G. Predicting protein function and other biomedical characteristics with heterogeneous ensembles. Methods San Diego Calif 2016 Jan 15;93:92–102. PMID:26342255
63. Arslanian S, Bacha F, Grey M, Marcus MD, White NH, Zeitler P. Evaluation and Management of Youth-Onset Type 2 Diabetes: A Position Statement by the American Diabetes Association. Diabetes Care 2018 Nov 12;41(12):2648–2668. doi: 10.2337/dci18-0052
64. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proc 22nd ACM SIGKDD Int Conf Knowl Discov Data Min New York, NY, USA: Association for Computing Machinery; 2016. p. 785–794. doi: 10.1145/2939672.2939785
65. Shwartz-Ziv R, Armon A. Tabular data: Deep learning is not all you need. Inf Fusion 2022 May 1;81:84–90. doi: 10.1016/j.inffus.2021.11.011
66. Goyal K, Dumancic S, Blockeel H. Feature Interactions in XGBoost. arXiv; 2020. doi: 10.48550/arXiv.2007.05758
67. Feature Interaction Constraints — xgboost 2.0.3 documentation. Available from: https://xgboost.readthedocs.io/en/stable/tutorials/feature_interaction_constraint.html [accessed Feb 1, 2024]
68. Sesmero MP, Ledezma AI, Sanchis A. Generating ensembles of heterogeneous classifiers using Stacked Generalization. WIREs Data Min Knowl Discov 2015;5(1):21–34. doi:

10.1002/widm.1143
69. Whetzel PL, Noy NF, Shah NH, Alexander PR, Nyulas C, Tudorache T, Musen MA. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Res 2011 Jul;39(Web Server issue):W541-545. PMID:21672956
70. Bhattacharya S, Andorf S, Gomes L, Dunn P, Schaefer H, Pontius J, Berger P, Desborough V, Smith T, Campbell J, Thomson E, Monteiro R, Guimaraes P, Walters B, Wiser J, Butte AJ. ImmPort: disseminating data to the public for the future of immunology. Immunol Res 2014 May;58(2–3):234–239. PMID:24791905
71. Zhang J, Baran J, Cros A, Guberman JM, Haider S, Hsu J, Liang Y, Rivkin E, Wang J, Whitty B, Wong-Erasmus M, Yao L, Kasprzyk A. International Cancer Genome Consortium Data Portal—a one-stop shop for cancer genomics data. Database 2011 Jan 1;2011:bar026. doi: 10.1093/database/bar026
72. Rayner L, McGovern A, Creagh-Brown B, Woodmansey C, de Lusignan S. Type 2 Diabetes and Asthma: Systematic Review of the Bidirectional Relationship. Curr Diabetes Rev 2019;15(2):118–126. PMID:29992891
73. Black MH, Anderson A, Bell RA, Dabelea D, Pihoker C, Saydah S, Seid M, Standiford DA, Waitzfelder B, Marcovina SM, Lawrence JM. Prevalence of Asthma and Its Association With Glycemic Control Among Youth With Diabetes. Pediatrics 2011 Oct;128(4):e839–e847. PMID:21949144
74. Wu TD. Diabetes, insulin resistance, and asthma: a review of potential links. Curr Opin Pulm Med 2021 Jan;27(1):29–36. PMID:33002990
75. Vartanian LR, Schwartz MB, Brownell KD. Effects of Soft Drink Consumption on Nutrition and Health: A Systematic Review and Meta-Analysis. Am J Public Health 2007 Apr;97(4):667–675. PMID:17329656
76. Greenwood DC, Threapleton DE, Evans CEL, Cleghorn CL, Nykjaer C, Woodhead C, Burley VJ. Association between sugar-sweetened and artificially sweetened soft drinks and type 2 diabetes: systematic review and dose-response meta-analysis of prospective studies. Br J Nutr 2014 Sep 14;112(5):725–734. PMID:24932880
77. Malik VS, Popkin BM, Bray GA, Després J-P, Willett WC, Hu FB. Sugar-sweetened beverages and risk of metabolic syndrome and type 2 diabetes: a meta-analysis. Diabetes Care 2010 Nov;33(11):2477–2483. PMID:20693348
78. Muraki I, Imamura F, Manson JE, Hu FB, Willett WC, van Dam RM, Sun Q. Fruit consumption and risk of type 2 diabetes: results from three prospective longitudinal cohort studies. BMJ 2013 Aug 28;347:f5001. PMID:23990623
79. McDonough C, Li YC. Youth preDM/DM dataset and Case Studies. Zenodo; 2024. doi: 10.5281/zenodo.10531245

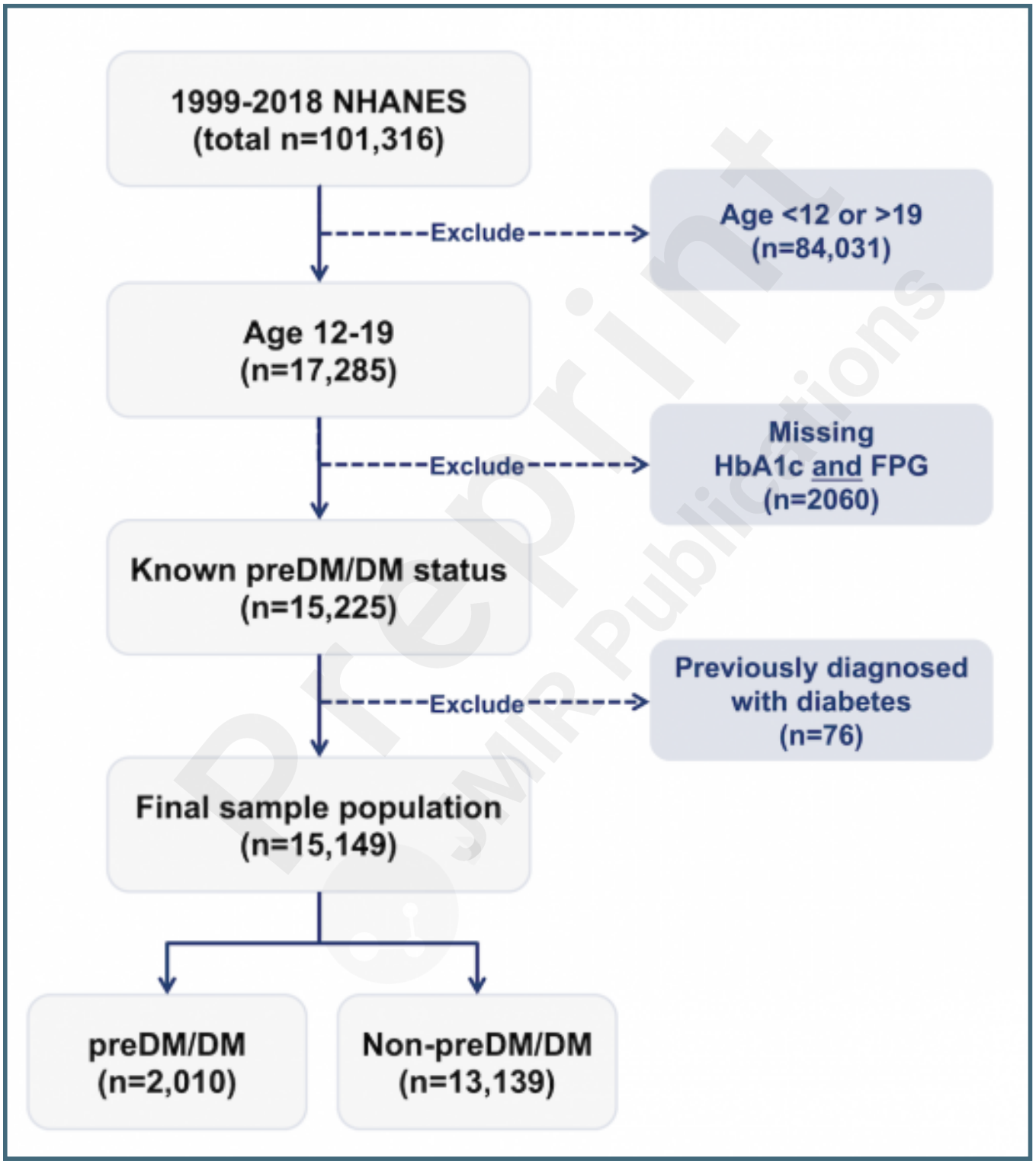
Supplementary Files

Figures

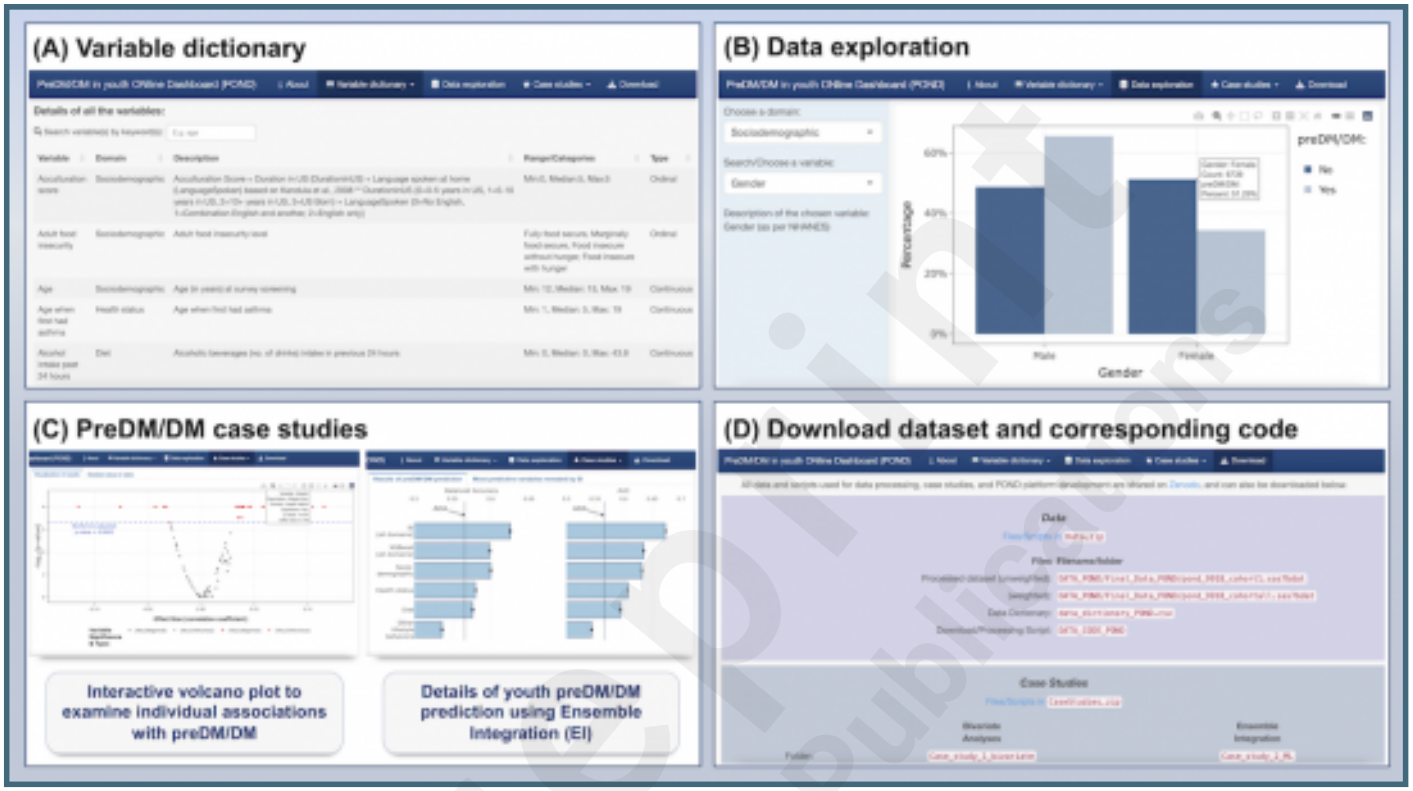
Study design and workflow. We processed data from 10 survey cycles (1999-2018) from the National Health and Nutrition Examination Survey (NHANES), which yielded 15,149 youth with known prediabetes/diabetes (preDM/DM) status. We extracted 95 variables that were relevant to preDM/DM and organized them into 4 domains: sociodemographic, health status, diet, and other lifestyle behaviors. We made the dataset easily accessible to the public through the user-friendly POND (Prediabetes/diabetes in youth ONline Dashboard) web portal, enabling users to navigate, visualize, and download the data. Additionally, we conducted two case studies with complementary statistical and machine learning methods that are designed to illustrate the translation potential of our dataset and point. Both analyses identified predictive variables associated with youth diabetes, and the results can be explored in POND. (Some images in this figure were obtained from the open-source collection at https://www.flaticon.com and were made by Freepik.).



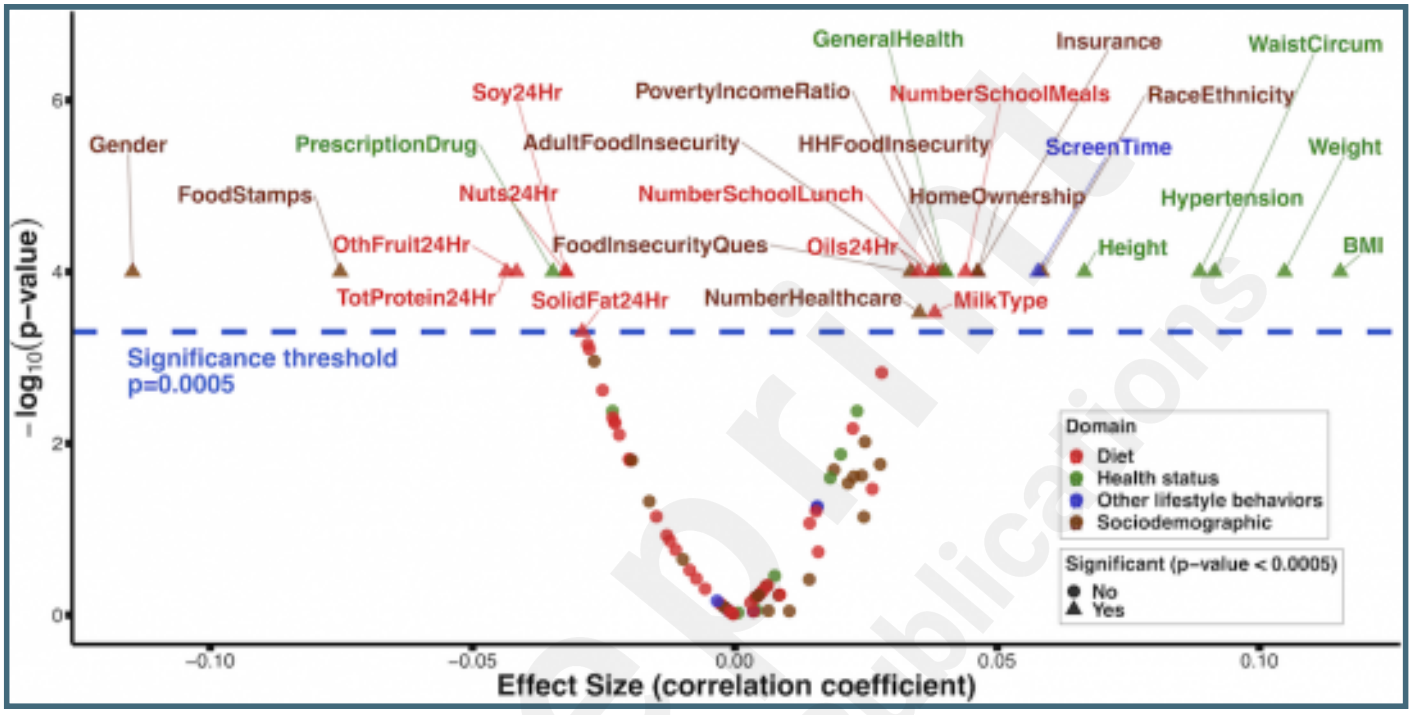
Flow chart showing the inclusion and exclusion criteria applied to 1999-2018 NHANES participants that yielded the study population included in our youth preDM/DM dataset. PreDM/DM status was defined by the current American Diabetes Association (ADA) biomarker criteria, i.e., elevated levels of one of two preDM/DM biomarkers (fasting plasma glucose (FPG) ?100 mg/dL or hemoglobin A1c (HbA1c) ? 5.7%). NHANES = National Health and Nutrition Examination Survey.



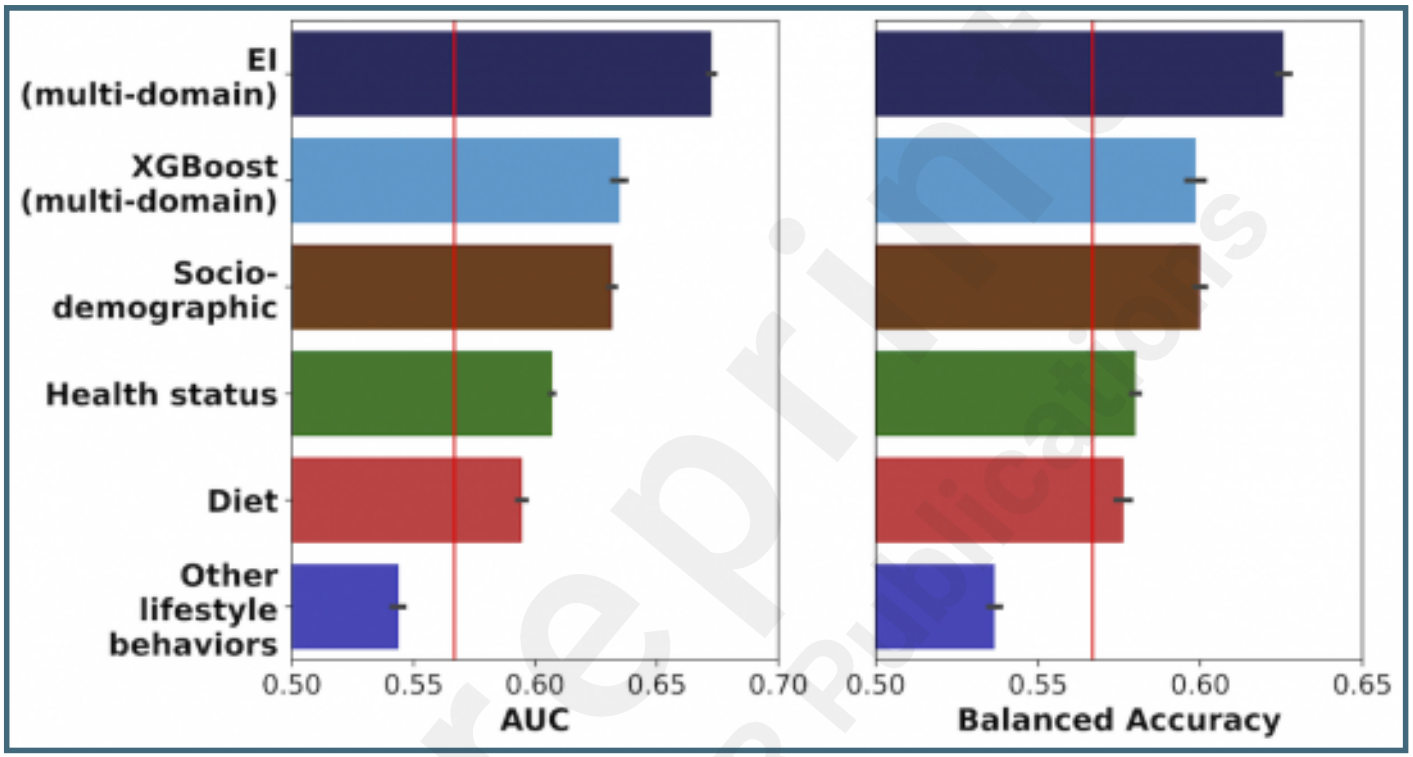
Screenshots of different functionalities available in POND (Prediabetes/diabetes in youth ONline Dashboard). (A) Detailed dictionary of the 95 variables included in our youth preDM/DM database organized by four domains, (B) Data exploration section showing the distribution of user selectable variables by preDM/DM status, (C) Case study section detailing the results of bivariate association analyses and the prediction of youth preDM/DM status from machine learning approaches and (D) Download section, where the dataset and the code used in the current study are publicly available to facilitate reproducibility and further exploration for interested users.



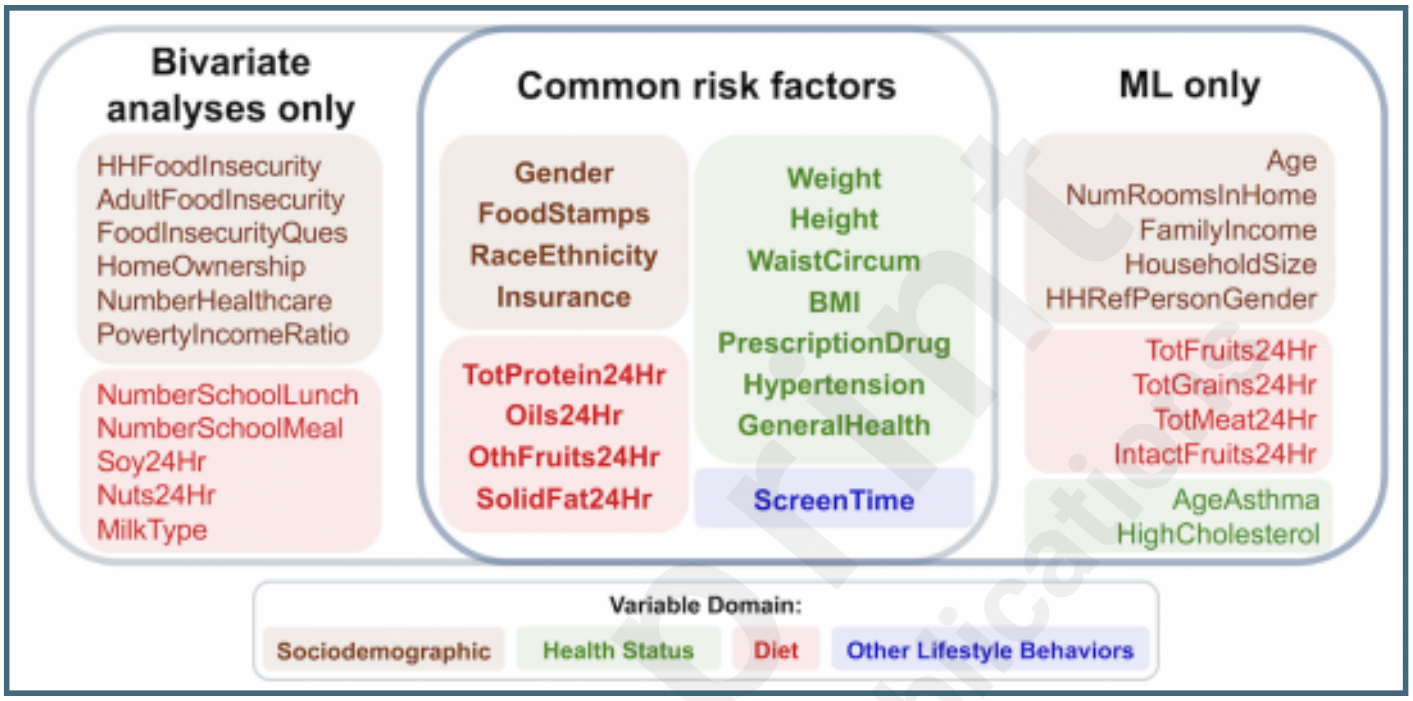
Individual variables associated with youth preDM/DM status based on bivariate analyses. This volcano plot shows the p-values and the effect sizes of the associations between the individual variables and youth preDM/DM status. Categorical and continuous variables were tested for association using Chi-square and Wilcoxon rank sum tests, respectively. Effect size was measured by Cramer's V for categorical variables and Wilcoxon's r-value [53] for continuous ones. After Bonferroni adjustment for multiple hypothesis testing, we found 27 variables to be significantly (p?0.0005; blue dotted line) associated with youth preDM/DM status. These are named above the blue dotted line in this plot, and colored by the domain they belong to.



Comparison of the performance of multiple approaches for predicting youth preDM/DM status based on machine learning approaches. We compared the performance of the multi-domain Ensemble Integration (EI) approach with three alternative prediction approaches. The alternative approaches were: (i) a modified form of the American Diabetes Association (ADA) screening guideline (vertical red line), (ii) single-domain EI-based prediction based on each of the four individual domains, and (iii) the commonly used eXtreme Gradient Boosting (XGBoost) algorithm applied to our whole dataset. Performance was measured in terms of the Area Under the ROC Curve (AUC) and Balanced Accuracy (average of sensitivity and specificity) measures. For each machine learning approach, the horizontal bar shows the average of the corresponding scores and the error bar indicates the corresponding standard error measured over ten rounds of five-fold cross-validation.



Variables associated with youth preDM/DM selected by bivariate analyses and the multi-domain EI approaches. Venn diagram summarizing the overlap between the 27 significant variables identified in the bivariate analyses and the 27 most predictive variables identified from the multidomain EI model. We found 16 variables overlapped between the two methods (Fisher's p=7.06x10^-6), and were drawn from all four domains (shown in different colors), indicating the multifactorial nature of youth preDM/DM.



Multimedia Appendixes

Supplementary Materials.
URL: http://asset.jmir.pub/assets/983db9dcdf638588937ab97f9b6983f3.doc