TEDDY

The Environmental Determinants of Diabetes in the Young

Health Informatics Institute

NIDDK DMS Webinar Series

Metadata and Data Standards

Kenneth Young, PhD

June 27, 2023

Health Informatics Institute
University of South Florida

Introduction

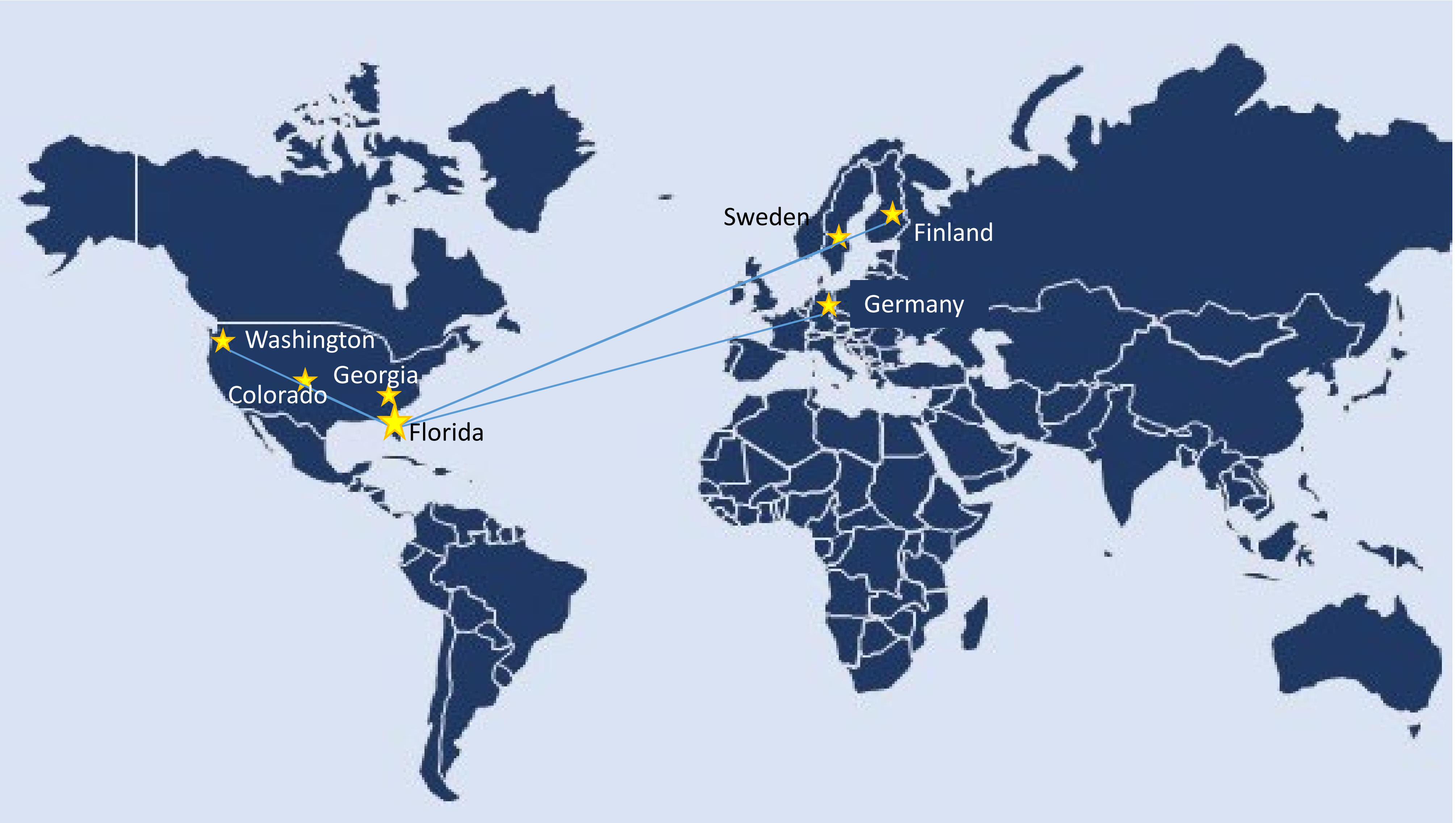
Health
Informatics
Institute

- Ken Young
  - Chief Information Officer
  - Assistant Professor
- Health Informatics Institute
- University of South Florida – Morsani College of Medicine
- TEDDY Data Coordinating Center (DCC) – Tampa, FL

TEDDY Study

- The Environmental Determinants of Diabetes in the Young (TEDDY) study.
- Primary objective(s) of this multi-center, multi-national, epidemiological study:
  - Identification of infectious agents, dietary factors, or other environmental exposures that are associated with increased risk of autoimmunity and T1DM.
  - Factors affecting specific phenotypic manifestations such as early age of onset or rate of progression, or with protection from the development of T1DM will also be identified.
- The TEDDY study investigates:
  - Genetic and genetic-environmental interactions, including gestational infection or other gestational events.
  - Childhood infections or other environmental factors after birth in relation to the development of prediabetes autoimmunity and T1D.

TEDDY Study Centers



TEDDY Study Centers

- Clinical Centers:
  - Colorado Barbara Davis Center, Univ. CO, Denver, CO
  - Finland University of Turku, Turku, Finland
  - Georgia/Florida Augusta University, Augusta, GA
  - Germany Diabetes Research Institute, Munich, Germany
  - Sweden Lund University, Malmö, Sweden
  - Washington Pacific Northwest Diabetes Research Institute, Seattle, WA
- Data Coordinating Center (DCC):
  - University of South Florida Health Informatics Institute, Tampa, FL

TEDDY Data Assets

- A variety of different data types are collected as a part of the TEDDY study, including clinical metadata and laboratory test result data across various 'omics analytes.
- The TEDDY DCC manages, curates, integrates, and provisions these data assets for analysis by TEDDY and approved external investigators.

TEDDY Clinical Metadata

- Case-Control Indicators
- Demographics
- Diet
- Family History
- Genotypes
- Household Exposures
- Medical History
- Medications
- Physical Activity
- Pre and Perinatal Exposures
- Psychosocial Stressors
- Test Results

TEDDY ‘Omics Analytes

- Dietary Biomarkers
- Epigenetics
- Exome
- Gene Expression
- Inflammatory Biomarkers
- Lipidomics
- Metabolomics
- Microbiome and Metagenomics
- Proteomics
- RNA Sequencing
- ImmunoChip SNPs
- Urinary Biomarkers
- Whole Genome Sequencing

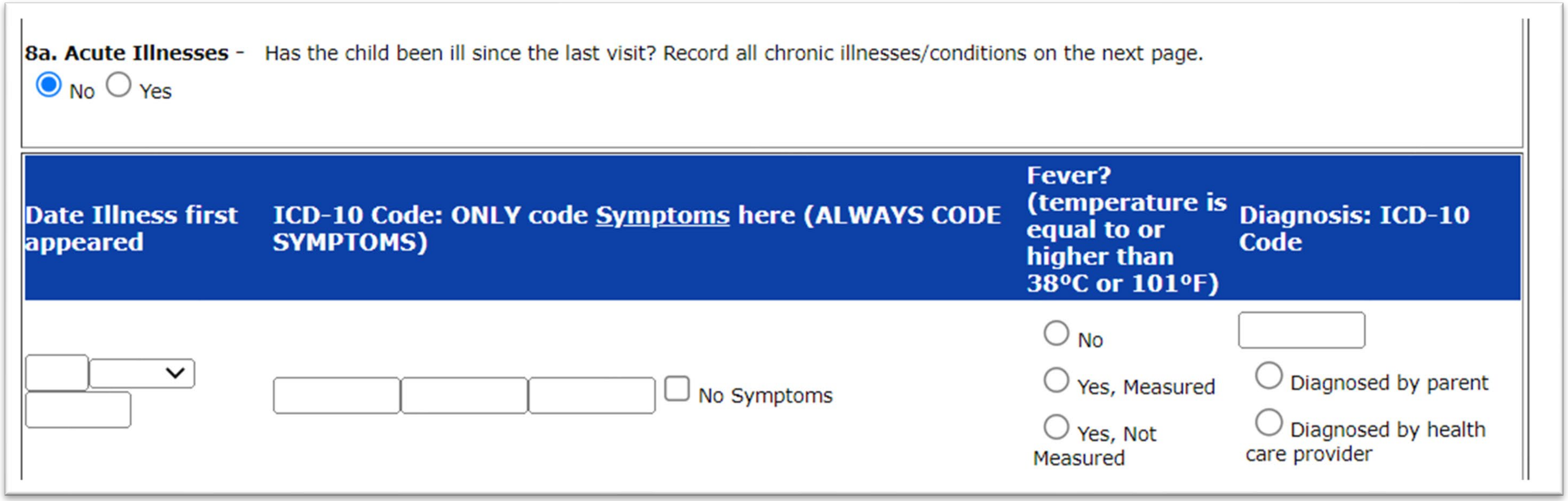
TEDDY Data Standards

- The TEDDY study has implemented the following biomedical ontologies to increase interoperability:

Type of Data	Standard/Ontology
Adverse Events/Reactions	CTCAE v5.0 (Common Terminology Criteria for Adverse Events)
Diagnoses	WHO ICD-10 (International Statistical Classification of Diseases and Related Health Problems) UMLS SNOMED CT (Systemized Nomenclature of Medicine Clinical Terms)
Medications	UMLS RxNorm

TEDDY Data Standards

- To improve the quality and (re)usability of the data, electronic case report forms (eCRFs) were designed to capture certain data standards directly.



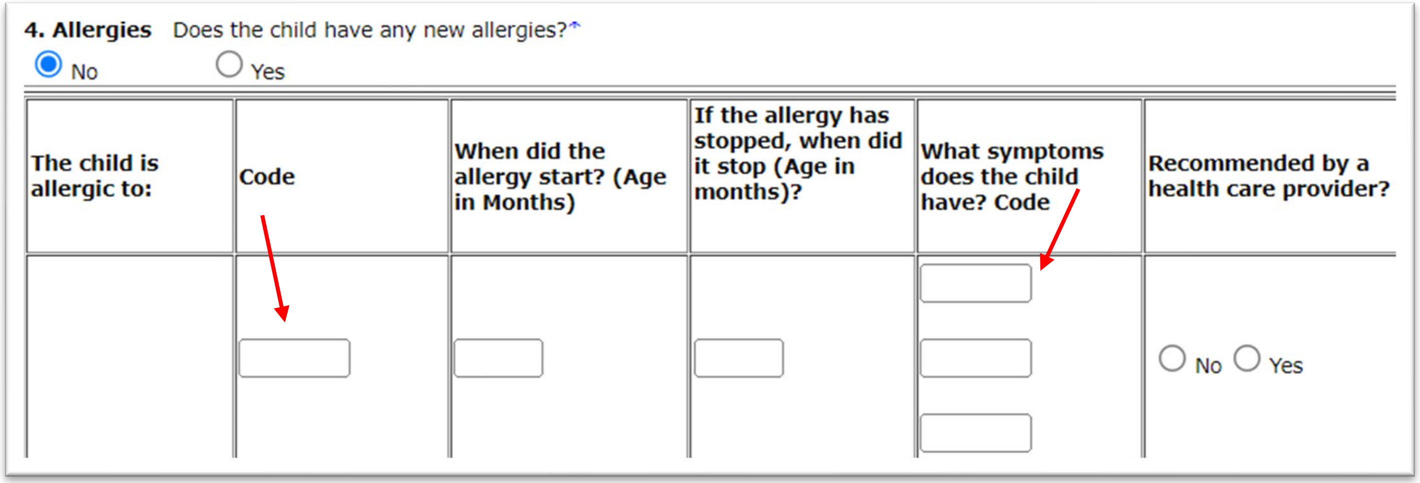
8a. Acute Illnesses - Has the child been ill since the last visit? Record all chronic illnesses/conditions on the next page.

● No ○ Yes

Date Illness first appeared	ICD-10 Code: ONLY code Symptoms here (ALWAYS CODE SYMPTOMS)	Fever? (temperature is equal to or higher than 38ºC or 101ºF)	Diagnosis: ICD-10 Code
	□ No Symptoms	○ No ○ Yes, Measured ○ Yes, Not Measured	○ Diagnosed by parent ○ Diagnosed by health care provider

TEDDY Data Standards

- TEDDY has also created unique “TEDDY codes” to capture other clinical data in a standardized way.
- Collecting codes in place of free-text fields improves data accuracy and provides consistency among reported values by restricting abbreviations, synonyms, and misspellings.



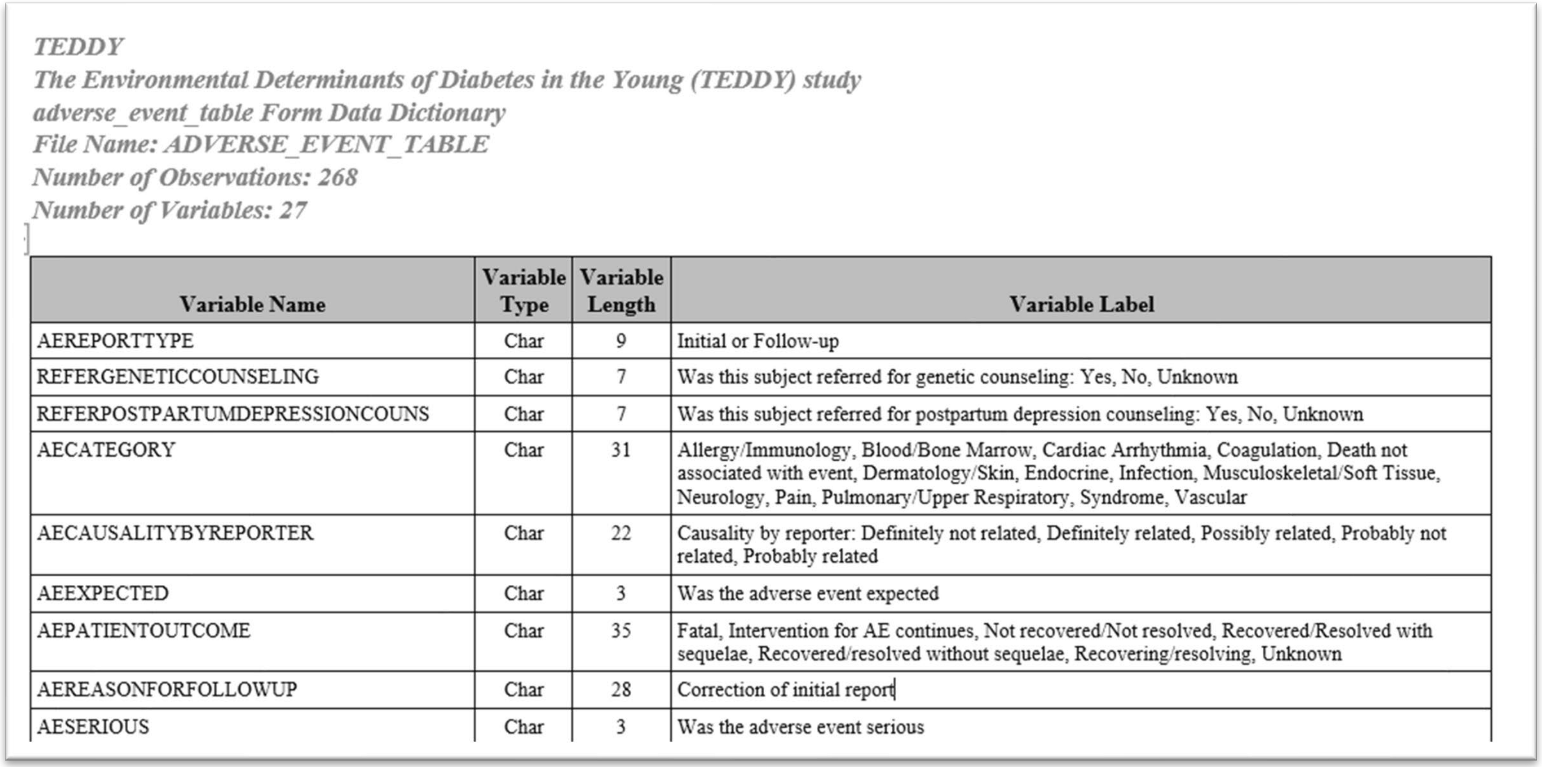
4. Allergies Does the child have any new allergies?*

● No ○ Yes

The child is allergic to:	Code	When did the allergy start? (Age in Months)	If the allergy has stopped, when did it stop (Age in months)?	What symptoms does the child have? Code	Recommended by a health care provider?
					○ No ○ Yes

TEDDY Data Dictionaries

- To ensure data are findable and reusable, each clinical data set TEDDY shares is accompanied by a data dictionary, which contains the variable name, type, length, and label.
- Data dictionaries can be provided in multiple formats including:
  - RTF
  - CSV
  - DOCX
  - XLSX



TEDDY

The Environmental Determinants of Diabetes in the Young (TEDDY) study

adverse_event_table Form Data Dictionary

File Name: ADVERSE_EVENT_TABLE

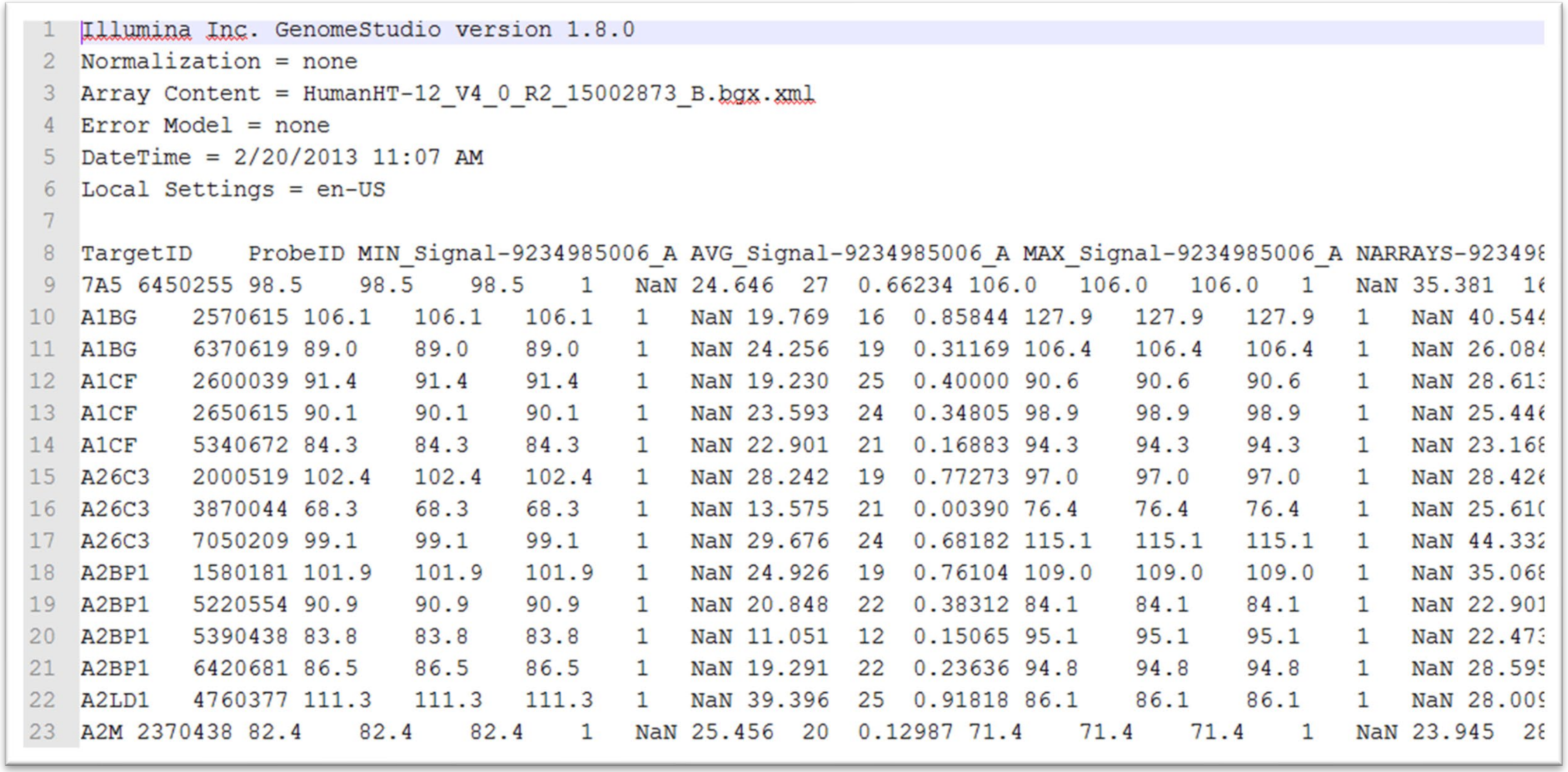
Number of Observations: 268

Number of Variables: 27

Variable Name	Variable Type	Variable Length	Variable Label
AEREPORTTYPE	Char	9	Initial or Follow-up
REFERGENETICCOUNSELING	Char	7	Was this subject referred for genetic counseling: Yes, No, Unknown
REFERPOSTPARTUMDEPRESSIONCOUNS	Char	7	Was this subject referred for postpartum depression counseling: Yes, No, Unknown
AECATEGORY	Char	31	Allergy/Immunology, Blood/Bone Marrow, Cardiac Arrhythmia, Coagulation, Death not associated with event, Dermatology/Skin, Endocrine, Infection, Musculoskeletal/Soft Tissue, Neurology, Pain, Pulmonary/Upper Respiratory, Syndrome, Vascular
AECAUSALITYBYREPORTER	Char	22	Causality by reporter: Definitely not related, Definitely related, Possibly related, Probably not related, Probably related
AEEXPECTED	Char	3	Was the adverse event expected
AEPATIENTOUTCOME	Char	35	Fatal, Intervention for AE continues, Not recovered/Not resolved, Recovered/Resolved with sequelae, Recovered/resolved without sequelae, Recovering/resolving, Unknown
AEREASONFORFOLLOWUP	Char	28	Correction of initial report
AESERIOUS	Char	3	Was the adverse event serious

TEDDY Data Dictionaries

- Dictionaries for omics data vary by data type.
- Examples of dictionaries include:
  - Manifest files *(Generated by Instrument Manufacturer)*
  - Annotation files *(Generated by Lab/TEDDY)*



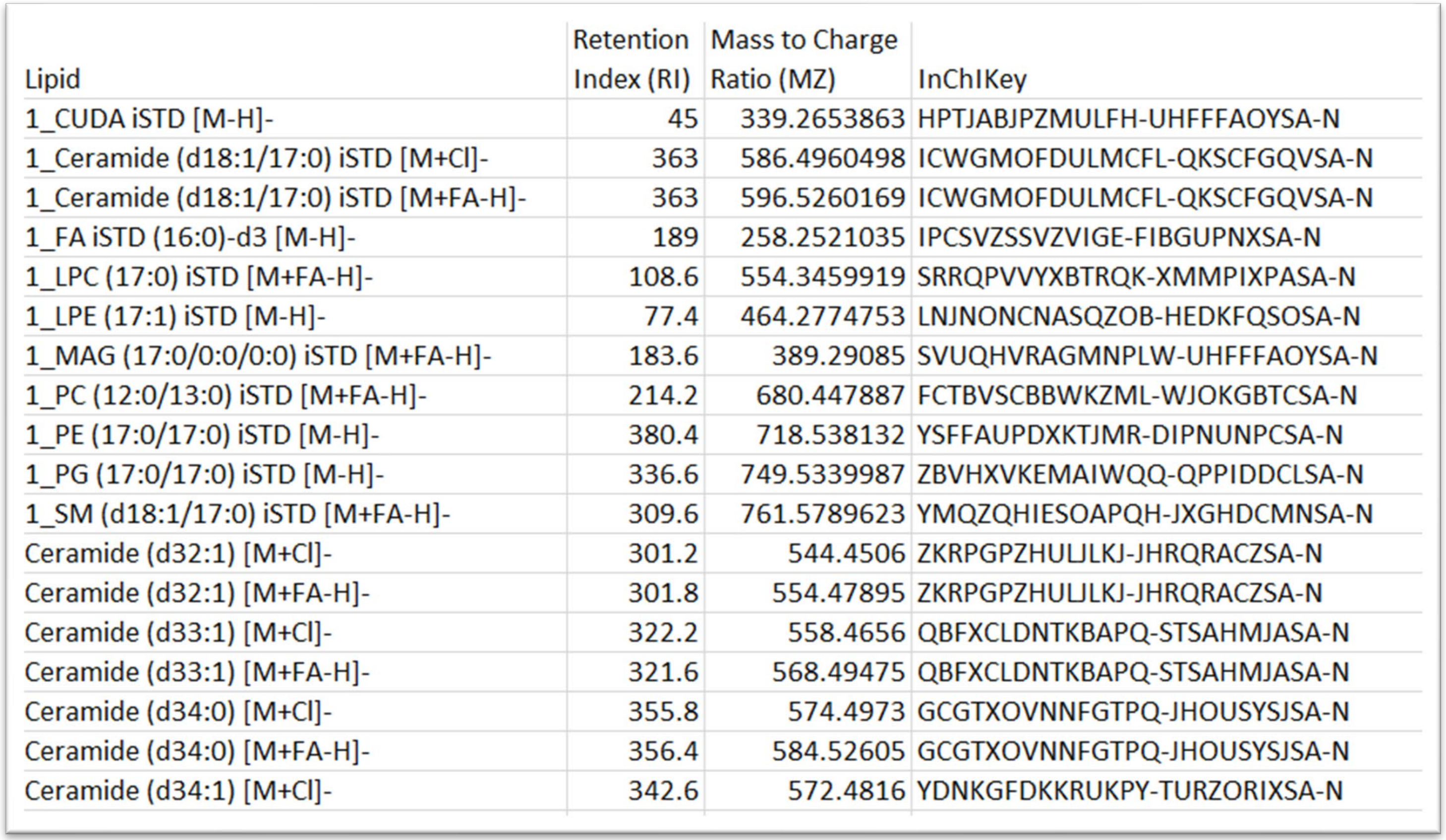
```
Illumina Inc. GenomeStudio version 1.8.0
Normalization = none
Array Content = HumanHT-12_V4_0_R2_15002873_B.bgx.xml
Error Model = none
DateTime = 2/20/2013 11:07 AM
Local Settings = en-US

TargetID    ProbeID MIN_Signal-9234985006_A AVG_Signal-9234985006_A MAX_Signal-9234985006_A NARRAYS-923498
7A5 6450255 98.5    98.5    98.5    1   NaN 24.646  27  0.66234 106.0   106.0   106.0   1   NaN 35.381  16
A1BG    2570615 106.1   106.1   106.1   1   NaN 19.769  16  0.85844 127.9   127.9   127.9   1   NaN 40.544
A1BG    6370619 89.0    89.0    89.0    1   NaN 24.256  19  0.31169 106.4   106.4   106.4   1   NaN 26.084
A1CF    2600039 91.4    91.4    91.4    1   NaN 19.230  25  0.40000 90.6    90.6    90.6    1   NaN 28.613
A1CF    2650615 90.1    90.1    90.1    1   NaN 23.593  24  0.34805 98.9    98.9    98.9    1   NaN 25.446
A1CF    5340672 84.3    84.3    84.3    1   NaN 22.901  21  0.16883 94.3    94.3    94.3    1   NaN 23.168
A26C3   2000519 102.4   102.4   102.4   1   NaN 28.242  19  0.77273 97.0    97.0    97.0    1   NaN 28.426
A26C3   3870044 68.3    68.3    68.3    1   NaN 13.575  21  0.00390 76.4    76.4    76.4    1   NaN 25.610
A26C3   7050209 99.1    99.1    99.1    1   NaN 29.676  24  0.68182 115.1   115.1   115.1   1   NaN 44.332
A2BP1   1580181 101.9   101.9   101.9   1   NaN 24.926  19  0.76104 109.0   109.0   109.0   1   NaN 35.068
A2BP1   5220554 90.9    90.9    90.9    1   NaN 20.848  22  0.38312 84.1    84.1    84.1    1   NaN 22.901
A2BP1   5390438 83.8    83.8    83.8    1   NaN 11.051  12  0.15065 95.1    95.1    95.1    1   NaN 22.473
A2BP1   6420681 86.5    86.5    86.5    1   NaN 19.291  22  0.23636 94.8    94.8    94.8    1   NaN 28.595
A2LD1   4760377 111.3   111.3   111.3   1   NaN 39.396  25  0.91818 86.1    86.1    86.1    1   NaN 28.009
A2M 2370438 82.4    82.4    82.4    1   NaN 25.456  20  0.12987 71.4    71.4    71.4    1   NaN 23.945  28
```

Illumina HumanHT-12 v4.0 Manifest File

TEDDY Data Dictionaries

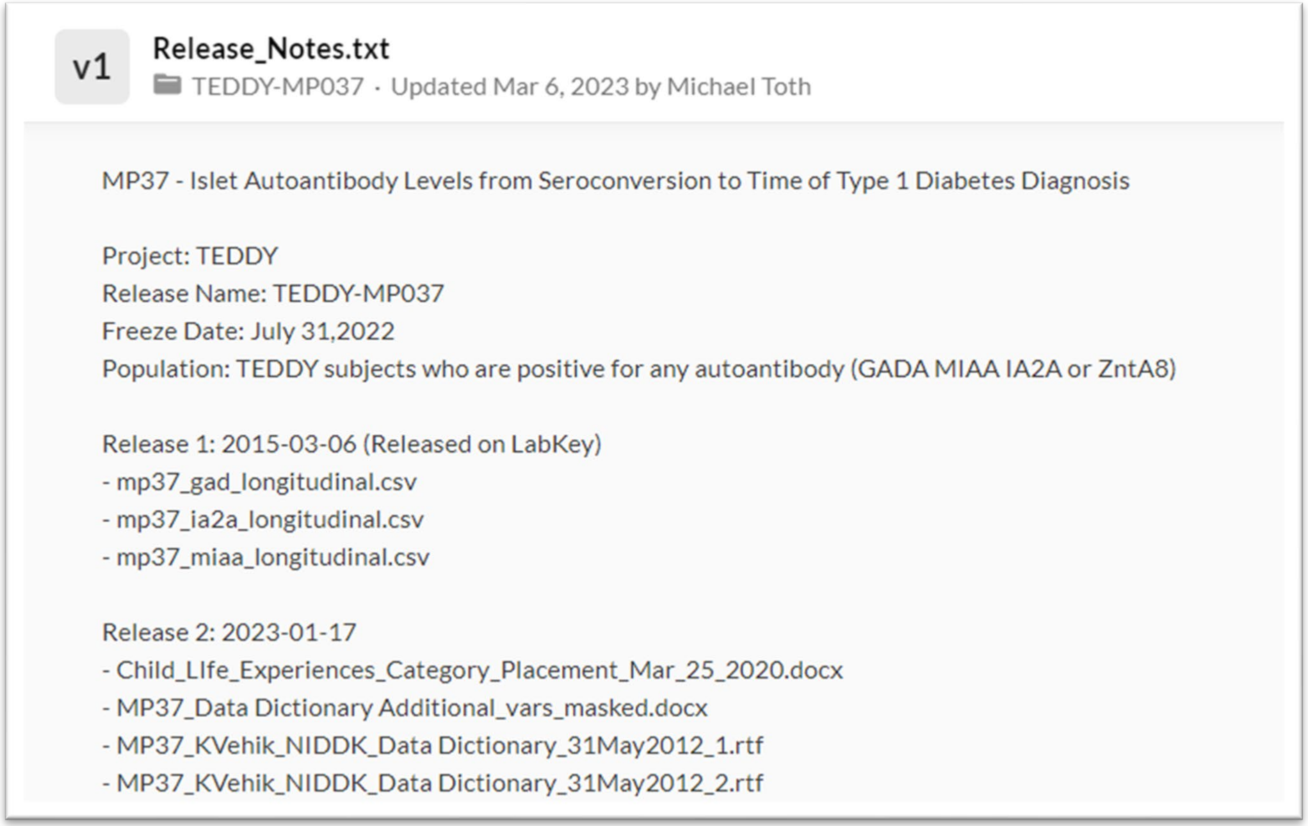
- Dictionaries for omics data vary by data type.
- Examples of dictionaries include:
  - Manifest files *(Generated by Instrument Manufacturer)*
  - Annotation files *(Generated by Lab/TEDDY)*



Lipid	Retention Index (RI)	Mass to Charge Ratio (MZ)	InChIKey
1_CUDA iSTD [M-H]-	45	339.2653863	HPTJABJPZMULFH-UHFFFAOYSA-N
1_Ceramide (d18:1/17:0) iSTD [M+Cl]-	363	586.4960498	ICWGMOFDULMCFL-QKSCFGQVSA-N
1_Ceramide (d18:1/17:0) iSTD [M+FA-H]-	363	596.5260169	ICWGMOFDULMCFL-QKSCFGQVSA-N
1_FA iSTD (16:0)-d3 [M-H]-	189	258.2521035	IPCSVZSSVZVIGE-FIBGUPNXSA-N
1_LPC (17:0) iSTD [M+FA-H]-	108.6	554.3459919	SRRQPVVYXBTRQK-XMMPIXPASA-N
1_LPE (17:1) iSTD [M-H]-	77.4	464.2774753	LNJNONCNASQZOB-HEDKFQSOSA-N
1_MAG (17:0/0:0/0:0) iSTD [M+FA-H]-	183.6	389.29085	SVUQHVRAGMNPLW-UHFFFAOYSA-N
1_PC (12:0/13:0) iSTD [M+FA-H]-	214.2	680.447887	FCTBVSCBBWKZML-WJOKGBTCSA-N
1_PE (17:0/17:0) iSTD [M-H]-	380.4	718.538132	YSFFAUPDXKTJMR-DIPNUNPCSA-N
1_PG (17:0/17:0) iSTD [M-H]-	336.6	749.5339987	ZBVHXVKEMAIWQQ-QPPIDDCLSA-N
1_SM (d18:1/17:0) iSTD [M+FA-H]-	309.6	761.5789623	YMQZQHIESOAPQH-JXGHDCMNSA-N
Ceramide (d32:1) [M+Cl]-	301.2	544.4506	ZKRPGPZHULJLKJ-JHRQRACZSA-N
Ceramide (d32:1) [M+FA-H]-	301.8	554.47895	ZKRPGPZHULJLKJ-JHRQRACZSA-N
Ceramide (d33:1) [M+Cl]-	322.2	558.4656	QBFXCLDNTKBAPQ-STSAHMJASA-N
Ceramide (d33:1) [M+FA-H]-	321.6	568.49475	QBFXCLDNTKBAPQ-STSAHMJASA-N
Ceramide (d34:0) [M+Cl]-	355.8	574.4973	GCGTXOVNNFGTPQ-JHOUSYSJSA-N
Ceramide (d34:0) [M+FA-H]-	356.4	584.52605	GCGTXOVNNFGTPQ-JHOUSYSJSA-N
Ceramide (d34:1) [M+Cl]-	342.6	572.4816	YDNKGFDKKRUKPY-TURZORIXSA-N

TEDDY Release Notes

- Direct to Investigator Data Releases also receive release notes describing the data freeze date, population, datasets provided, and any relevant notes for the investigator



Release_Notes.txt

TEDDY-MP037 · Updated Mar 6, 2023 by Michael Toth

MP37 - Islet Autoantibody Levels from Seroconversion to Time of Type 1 Diabetes Diagnosis

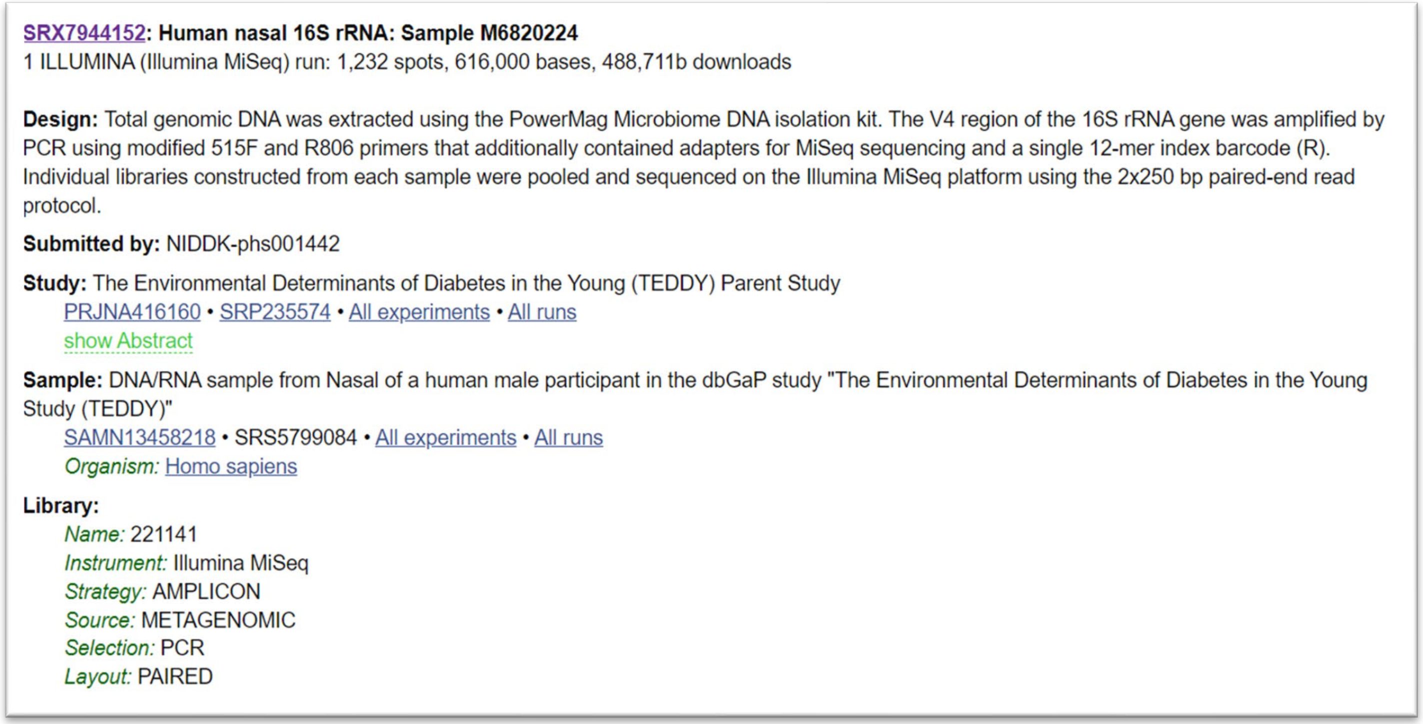
Project: TEDDY
Release Name: TEDDY-MP037
Freeze Date: July 31,2022
Population: TEDDY subjects who are positive for any autoantibody (GADA MIAA IA2A or ZntA8)

Release 1: 2015-03-06 (Released on LabKey)
- mp37_gad_longitudinal.csv
- mp37_ia2a_longitudinal.csv
- mp37_miaa_longitudinal.csv

Release 2: 2023-01-17
- Child_LIfe_Experiences_Category_Placement_Mar_25_2020.docx
- MP37_Data Dictionary Additional_vars_masked.docx
- MP37_KVehik_NIDDK_Data Dictionary_31May2012_1.rtf
- MP37_KVehik_NIDDK_Data Dictionary_31May2012_2.rtf

TEDDY Omics Metadata

- TEDDY omics metadata is also shared with data repositories, such as dbGaP and Metabolomics Workbench.



SRX7944152: Human nasal 16S rRNA: Sample M6820224
1 ILLUMINA (Illumina MiSeq) run: 1,232 spots, 616,000 bases, 488,711b downloads

Design: Total genomic DNA was extracted using the PowerMag Microbiome DNA isolation kit. The V4 region of the 16S rRNA gene was amplified by PCR using modified 515F and R806 primers that additionally contained adapters for MiSeq sequencing and a single 12-mer index barcode (R). Individual libraries constructed from each sample were pooled and sequenced on the Illumina MiSeq platform using the 2x250 bp paired-end read protocol.

Submitted by: NIDDK-phs001442

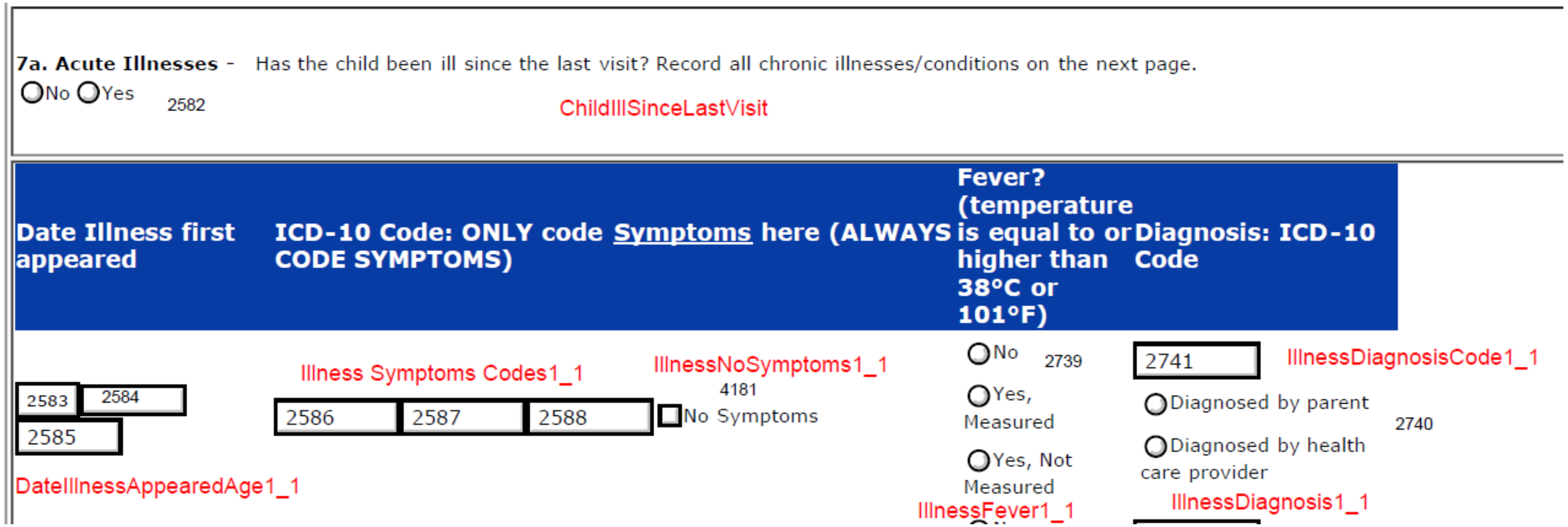
Study: The Environmental Determinants of Diabetes in the Young (TEDDY) Parent Study
PRJNA416160 • SRP235574 • All experiments • All runs
show Abstract

Sample: DNA/RNA sample from Nasal of a human male participant in the dbGaP study "The Environmental Determinants of Diabetes in the Young Study (TEDDY)"
SAMN13458218 • SRS5799084 • All experiments • All runs
Organism: Homo sapiens

Library:
Name: 221141
Instrument: Illumina MiSeq
Strategy: AMPLICON
Source: METAGENOMIC
Selection: PCR
Layout: PAIRED

TEDDY Annotated Forms

- TEDDY eCRFs annotated with the data set name at the top and variable name by each field have been shared with the NIDDK Central Repository as a searchable PDF.
- Annotated forms allow investigators to find data of interest, see how it was collected, and identify related variables.



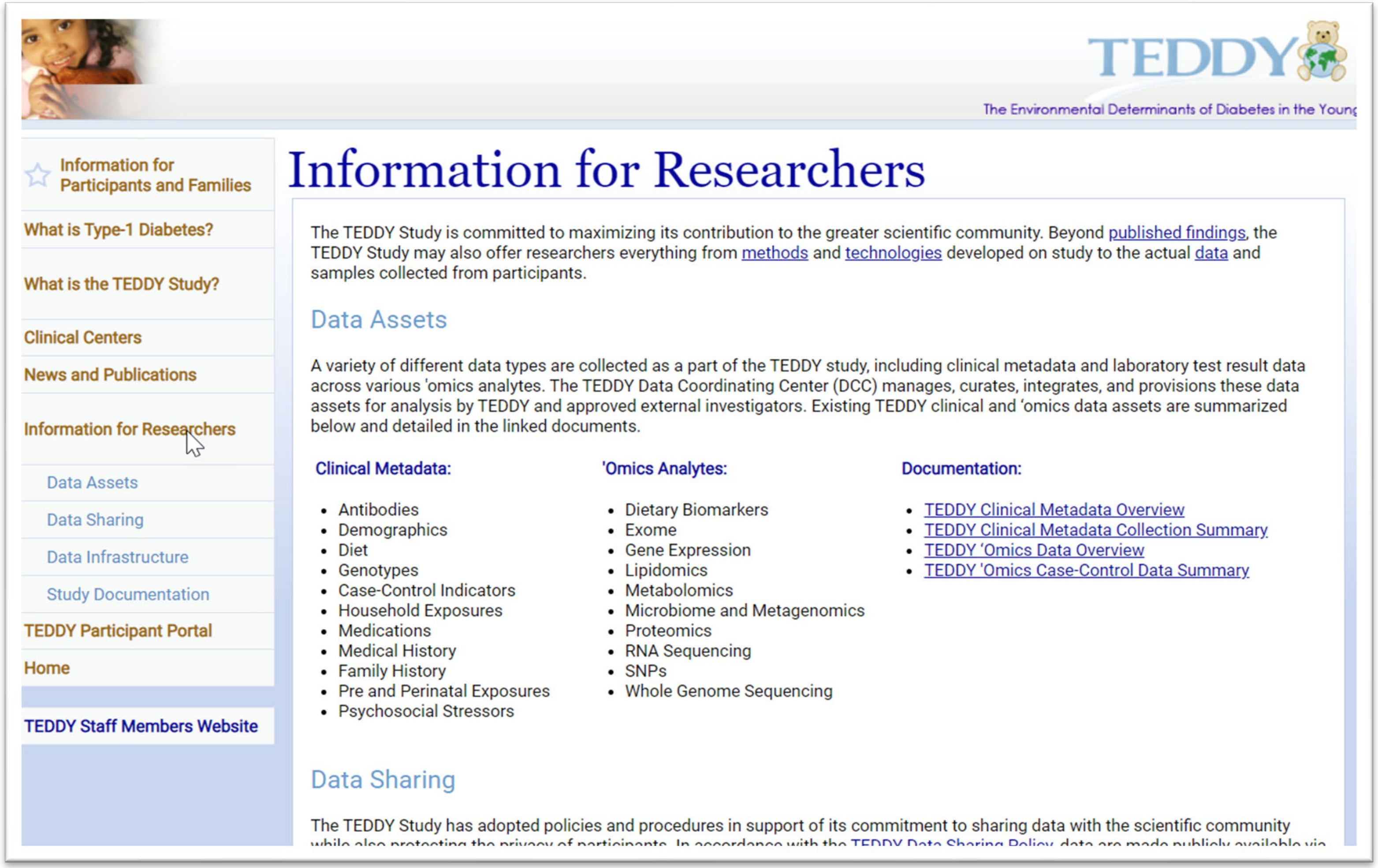
7a. Acute Illnesses - Has the child been ill since the last visit? Record all chronic illnesses/conditions on the next page.

○No ○Yes 2582 ChildIllSinceLastVisit

Date Illness first appeared	ICD-10 Code: ONLY code Symptoms here (ALWAYS CODE SYMPTOMS)		Fever? (temperature is equal to or higher than 38°C or 101°F)	Diagnosis: ICD-10 Code	
2583 2584 2585 DateIllnessAppearedAge1_1	Illness Symptoms Codes1_1 2586 2587 2588	IllnessNoSymptoms1_1 4181 □No Symptoms	○No 2739 ○Yes, Measured ○Yes, Not Measured IllnessFever1_1	2741 ○Diagnosed by parent ○Diagnosed by health care provider IllnessDiagnosis1_1	IllnessDiagnosisCode1_1 2740

TEDDY Data Documentation

- On the TEDDY public website, investigators can find documents detailing TEDDY data collection procedures (e.g., protocol, MOO, NCC Design), data availability, and data sharing policy.
- Data documentation available on the public website and across repositories has been developed to make TEDDY data more FAIR.



Information for Researchers

The TEDDY Study is committed to maximizing its contribution to the greater scientific community. Beyond published findings, the TEDDY Study may also offer researchers everything from methods and technologies developed on study to the actual data and samples collected from participants.

Data Assets

A variety of different data types are collected as a part of the TEDDY study, including clinical metadata and laboratory test result data across various 'omics analytes. The TEDDY Data Coordinating Center (DCC) manages, curates, integrates, and provisions these data assets for analysis by TEDDY and approved external investigators. Existing TEDDY clinical and 'omics data assets are summarized below and detailed in the linked documents.

Clinical Metadata:

- Antibodies
- Demographics
- Diet
- Genotypes
- Case-Control Indicators
- Household Exposures
- Medications
- Medical History
- Family History
- Pre and Perinatal Exposures
- Psychosocial Stressors

'Omics Analytes:

- Dietary Biomarkers
- Exome
- Gene Expression
- Lipidomics
- Metabolomics
- Microbiome and Metagenomics
- Proteomics
- RNA Sequencing
- SNPs
- Whole Genome Sequencing

Documentation:

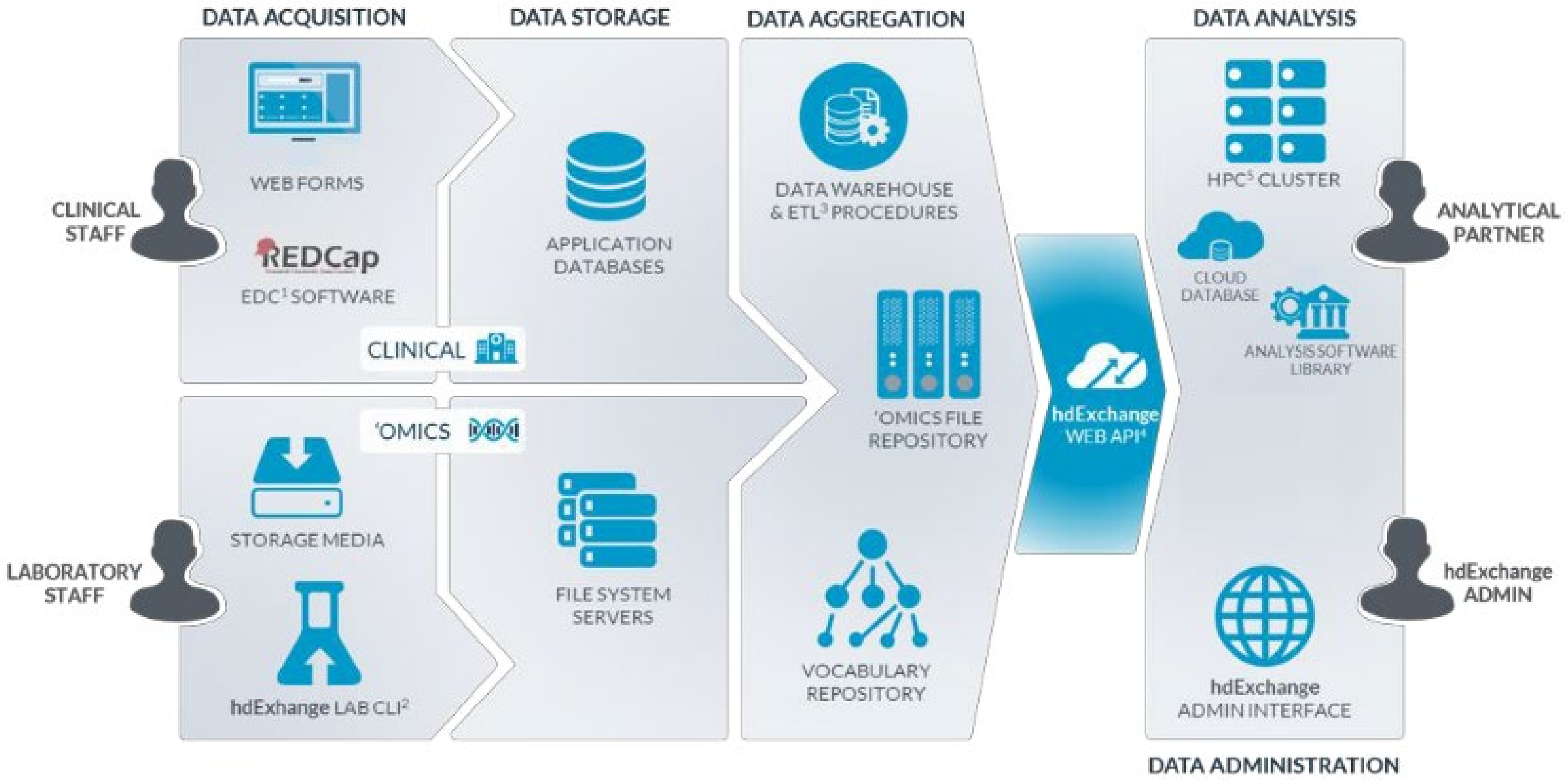
- TEDDY Clinical Metadata Overview
- TEDDY Clinical Metadata Collection Summary
- TEDDY 'Omics Data Overview
- TEDDY 'Omics Case-Control Data Summary

Data Sharing

The TEDDY Study has adopted policies and procedures in support of its commitment to sharing data with the scientific community while also protecting the privacy of participants. In accordance with the TEDDY Data Sharing Policy, data are made publicly available via

HII Data Infrastructure

- As TEDDY DCC, HII developed infrastructure to support large scale data and analysis.



TEDDY Data Sharing

- The TEDDY Study has adopted policies and procedures in support of its commitment to sharing data with the scientific community while also protecting the privacy of participants.
- Data releases have been submitted at different time points and to various repositories, depending on NIH requirements and the nature of the data.
- Each submission is treated as an independent release, possessing uniquely masked subject and sample identifiers.
- Researchers may desire to combine data across these releases for analysis but are unable to do so as a result of the independently masked identifiers. The NIDDK repository can provide repository data release identifier mapping materials to satisfy this demand once the investigators have received approval to access the data.

Acknowledgements

- Health Informatics Institute at the University of South Florida (USF)
  - Dr. Jeffrey Krischer
  - Dena Tewey
  - Chris Shaffer
- TEDDY Study Group:
  - NIH, TEDDY DCC, clinical center investigators, clinical center staff, and TEDDY study participants.
- NIDDK

Health Informatics Institute

University of South Florida

TEDDY

The Environmental Determinants of Diabetes in the Young