Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) M233 Xhonneux

Prepared by NIDDK-CR
May 21, 2024

Contents

1 Standard Disclaimer

The intent of this DSIC is to provide confidence that the data distributed by the NIDDK repository is a true copy of the study data. Our intent is not to assess the integrity of the statistical analyses reported by study investigators. As with all statistical analyses of complex datasets, complete replication of a set of statistical results should not be expected in secondary analysis. This occurs for a number of reasons including differences in the handling of missing data, restrictions on cases included in samples for a particular analysis, software coding used to define complex variables, etc. Experience suggests that most discrepancies can ordinarily be resolved by consultation with the study data coordinating center (DCC), however this process is labor-intensive for both DCC and Repository staff. It is thus not our policy to resolve every discrepancy that is observed in an integrity check. Specifically, we do not attempt to resolve minor or inconsequential discrepancies with published results or discrepancies that involve complex analyses, unless NIDDK Repository staff suspect that the observed discrepancy suggests that the dataset may have been corrupted in storage, transmission, or processing by repository staff. We do, however, document in footnotes to the integrity check those instances in which our secondary analyses produced results that were not fully consistent with those reported in the target publication.

2 Study Background

The TEDDY study was designed to follow children with and without a family history of type 1 diabetes (T1D) to understand the environmental factors that contribute to the disease. Newborn children younger than 4 months were screened for high-risk HLA alleles, and those with qualifying haplotypes were eligible for follow-up. Data were collected on medical information (infections, medication, immunizations), exposure to dietary and other environmental factors, negative life events, family history, tap water, and measurements of psychological stress. Specimens, including blood, stool, urine, and nail clippings, were taken at baseline and follow-up study visits. The primary outcome measures include two endpoints—the first appearance of one or more islet cell autoantibodies (GADA, IAA, or IA-2A), confirmed at two consecutive visits, and development of T1D. The cohort was followed for 15 years, or until the occurrence of one of the primary endpoints.

The M233 study identified and interpreted age-associated gene expression changes in healthy infancy and age-independent changes tracking with progression to both T1D and islet autoimmunity, beginning before other evidence of islet autoimmunity was present.

3 Archived Datasets

A full listing of the archived datasets included in the package can be found in the Roadmap document. All data files, as provided by the Data Coordinating Center (DCC), are located in the TEDDY folder in the data package. For this replication, variables were taken from the “traits” tab in the “M_233_Xhonneux_NIDDK_TEDDYmetadata_05Aug2021.xlsm” spreadsheet.

4 Statistical Methods

Analyses were performed to replicate the TEDDY cohort characteristics data file S2 results in the publication by Xhonneux et al. [1]. To verify the integrity of the data, only descriptive statistics were computed.

5 Results

For the TEDDY cohort characteristics data file S2 in the publication [1], Table A lists the variables that were used in the replication, and Table B compares the results calculated from the archived data files to the publication results in data file S2. The results of the replication are within expected variation to the published results.

6 Conclusions

The NIDDK Central Repository is confident that the TEDDY M233 data files to be distributed are a true copy of the study data.

7 References

[1] Xhonneux LP, Knight O, Lernmark Å, Bonifacio E, Hagopian WA, Rewers MJ, She JX, Toppari J, Parikh H, Smith KGC, Ziegler AG, Akolkar B, Krischer JP, McKinney EF. Transcriptional Networks in At-risk Individuals Identify Signatures of Type 1 Diabetes Progression. Science Translational Medicine, 13(587), eabd5666, March 2021. doi: https://doi.org/10.1126/scitranslmed.abd5666

Table A: Variables used to replicate select variables from the TEDDY cohort characteristics data file S2 publication

Table Variable	**dataset.variable**
Clinical Center	traits.cc
Country	traits.country
Type 1 Diabetes	traits.t1d
Sex	traits.sex
Race	traits.race
Ethnicity	traits.race_ethnicity

Table B: Comparison of values computed in integrity check to reference article TEDDY cohort characteristics data file S2

Characteristic	**Pub: Data file S2 (n=400)**	**DSIC: TEDDY spreadsheet tab "Traits" (n=401)**	**Diff. (n=1)**
Clinical Center			
Colorado	51 (12.7)	51 (12.7)	0 (0)
Georgia	35 (8.7)	36 (9.0)	1 (0.3)
Washington	38 (9.5)	38 (9.5)	0 (0)
Finland	113 (28.2)	113 (28.2)	0 (0)
Germany	40 (10.0)	40 (10.0)	0 (0)
Sweden	123 (30.8)	123 (30.7)	0 (0.1)
Country			
USA	124 (31.0)	125 (31.2)	1 (0.2)
Finland	113 (28.2)	113 (28.2)	0 (0)
Germany	40 (10.0)	40 (10.0)	0 (0)
Sweden	123 (30.7)	123 (30.7)	0 (0)
Type 1 Diabetes			
Yes	62 (15.5)	63 (15.7)	1 (0.2)
No	338 (84.5)	338 (84.3)	0 (0.2)
Sex			
Female	192 (48.0)	192 (47.9)	0 (0.1)
Male	208 (52.0)	209 (52.1)	1 (0.1)
Race			
White	270 (67.5)	271 (67.6)	1 (0.1)
Black or African American	2 (0.5)	2 (0.5)	0 (0)
Asian	0 (0)	0 (0)	0 (0)
Native Hawaiian/Pacific Islander	0 (0)	0 (0)	0 (0)
American Indian/Alaska Native	0 (0)	0 (0)	0 (0)
Unknown or Not reported	124 (31.0)	124 (30.9)	0 (0.1)
More than one race	4 (1.0)	4 (1.0)	0 (0)
Ethnicity			
All Hispanics regardless of race	15 (3.7)	15 (3.7)	0 (0)
White Non-Hispanic	256 (64.0)	257 (64.1)	1 (0.1)
African American Non-Hispanic	2 (0.5)	2 (0.5)	0 (0)
All other races Non-Hispanic	3 (0.8)	3 (0.8)	0 (0)
Missing or Unknown race and ethnicity	124 (31.0)	124 (30.9)	0 (0.1)

Attachment A: SAS Code

```
libname m233 "\\gshare.bah.com\MD Baltimore\NIDDK\niddk-
dr_studies6\TEDDY\private_created_data\M233 & M239\M_233_Xhonneux_NIDDK_Submission";

/*****************************/
/* TEDDY M233 DSIC ***********/
/*****************************/

*Imported "Traits" sheet from "M_233_Xhonneux_NIDDK_TEDDYmetadata_05Aug2021.xlsm" and the
TEDDY Cohort Characteristics Dataset S2 downloadable from the publication;

data submitted; set work.m_233_xhonneux_niddk_teddymetada;
run; *2013 observations;

data downld; set work.'abd5666_data_file_s2 (1)'n;
run; *400 observations;

*Sorting and removing duplicates;
proc sort data=submitted nodupkey;
by MaskID;
run; *400 observations;

*Keeping only certain demographics for comparison from the submitted dataset;
data sub1; set submitted;
keep MaskID cc country race race_ethnicity Sex t1d;
run;

data down1; set downld;
keep cc country Sex race race_ethnicity t1d individual_id ;
run;

*Comparing the two datasets;

*CC;
proc freq data=sub1;
tables cc;
run;

proc freq data=down1;
tables cc;
run;

*Country;
proc freq data=down1;
tables country;
run;
```

```
proc freq data=sub1;
tables country;
run;

*T1D;
proc freq data=down1;
tables t1d;
run;

proc freq data=sub1;
tables t1d;
run;

*Sex;
proc freq data=down1;
tables sex;
run;

proc freq data=sub1;
tables sex;
run;

*Race;
proc freq data=down1;
tables race;
run;

proc freq data=sub1;
tables race;
run;

*Race/Ethnicity;
proc freq data=down1;
tables race_ethnicity;
run;

proc freq data=sub1;
tables race_ethnicity;
run;
```