Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) M208 Liu

Prepared by NIDDK-CR
April 17, 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. Information is 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. Biospecimens, including blood, stool, urine, and nail clippings, are 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 will be followed for 15 years, or until the occurrence of one of the primary endpoints.

The M208 study investigated two-phase growth patterns in early life and their association with development of islet autoimmunity and type 1 diabetes.

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 "m_208_xliu_niddk_28feb2018_1.sas7bdat" dataset.

4 Statistical Methods

Analyses were performed to replicate the results section for the data in the publication by Liu et al. [1]. To verify the integrity of the data, only descriptive statistics were computed.

5 Results

For the results 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 results section in the publication. The results of the replication are within expected variation to the published results.

6 Conclusions

The NIDDK Central Repository is confident that the TEDDY M208 data files to be distributed are a true copy of the study data.

7 References

[1] Liu X, Vehik K, Huang Y, Larsson HE, Toppari J, Ziegler AG, She JX, Rewers M, Hagopian WA, Akolkar B, Krischer JP. Distinct Growth Phases in Early Life Associated With the Risk of Type 1 Diabetes: The TEDDY Study. Diabetes Care, 43(3), 556-562, March 2020. doi: https://doi.org/10.2337/dc19-1670

Table A: Variables used to replicate publication results section

Table Variable	**dataset.variable**
Developed one or more persistent autoantibodies (GADA, IAA, or IA2A)	m_208_xliu_niddk_28feb2018_1.persist_conf_ab
Median age at development of persistent autoantibodies	m_208_xliu_niddk_28feb2018_1.persist_ab_age
Developed T1D	m_208_xliu_niddk_28feb2018_1.t1d
Median age at development of T1D	m_208_xliu_niddk_28feb2018_1.agemos_last_clinic m_208_xliu_niddk_28feb2018_1.t1d
Of the 761 participants with one or more persistent autoantibodies, the initial autoantibody at seroconversion	m_208_xliu_niddk_28feb2018_1.firstab
Of the 761 participants with one or more persistent autoantibodies, development of T1D	m_208_xliu_niddk_28feb2018_1.t1d m_208_xliu_niddk_28feb2018_1.persist_conf_ab

Table B: Comparison of values computed in integrity check to reference article results section

Characteristic	**Pub: Total (n=7,522)**	**DSIC: Total (n=7,522)**	**Diff. (n=0)**
Developed one or more persistent autoantibodies (GADA, IAA, or IA2A)	761 (10.1%)	761 (10.1%)	0 (0)
Median age at development of persistent autoantibodies (yrs), Median (IQR)	3.0 (1.5-6.0)	3.0 (1.5-6.0)	0 (0)
Developed T1D	290 (3.9%)	290 (3.9%)	0 (0)
Median age at development of T1D (yrs), Median (IQR)	5.5 (3.0-8.2)	6.2 (3.0-8.8)	0.7 (0-0.6)
Of the 761 participants with one or more persistent autoantibodies, the initial autoantibody at seroconversion:			
IAA only	287 (37.7%)	287 (37.7%)	0 (0)
GADA only	329 (43%)	329 (43%)	0 (0)
IA2A only	20 (2.6%)	20 (2.6%)	0 (0)
Multiple	125 (16.4%)	125 (16.4%)	0 (0)
Of the 761 participants with one or more persistent autoantibodies, development of T1D	272 (35.7%)	272 (35.7%)	0 (0)

Attachment A: SAS Code

```
libname m208 "X:\NIDDK\niddk-dr_studies6\TEDDY\private_created_data\M199, M208, and
M226\M_208_XLiu_NIDDK_Submission";

/*****************************/
/* TEDDY M208 DSIC Xiang Liu */
/*****************************/

*# and median age at persistent AB;
data one; set m208.m_208_xliu_niddk_28feb2018_1;
persist_ab_age_yr = persist_ab_age/12;
run;

proc freq data=one;
tables persist_conf_ab;
run;
*761 (10.1%);

proc means data=one n median q1 q3;
var persist_ab_age_yr;
run;
*3.0 (1.5, 6.0)

*developed T1d;
proc freq data=one;
tables t1d;
run;
*290 (3.9%);

data one_1; set one;
age_yrs_last_clinic = agemos_last_clinic/12;
run;

proc means data=one_1 n median q1 q3;
var age_yrs_last_clinic ;
class t1d;
run;
*6.2 (3.0, 8.8)

*among the 761, IAA only, GADA only, IA2A only, multiple;

proc freq data=one;
tables firstab;
run;
*match to pub;
```

```
*of the 761 >> progress to T1D;
proc freq data=one;
tables t1d*persist_conf_ab/norow nopercent;
run;
```