Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) M72 Sharma

Prepared by NIDDK-CR
August 18, 2022

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 M72 study sought to understand the role of certain non-HLA genes in the development of tissue transglutaminase autoantibodies (tTGA) and celiac disease.

3 Archived Datasets

All data files, as provided by the Data Coordinating Center (DCC), are located in TEDDY M72 folder in the data package. For this replication, variables were taken from the "m_72_asharma_niddk_31aug2013_1.sas7bdat" dataset.

4 Statistical Methods

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

5 Results

For Supplemental Table S2 in the publication [1], Characteristics for Celiac Disease Autoimmunity, 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 in Supplemental Table S2. The results of the replication are within expected variation of the published results.

6 Conclusions

The NIDDK Central Repository is confident that the TEDDY M72 data files to be distributed are a true copy of the study data.

7 References

[1] Sharma A, Liu X, Hadley D, Hagopian W, Liu E, Chen WM, Onengut-Gumuscu S, Simell V, Rewers M, Ziegler AG, Lernmark Å, Simell O, Toppari J, Krischer JP, Akolkar B, Rich SS, Agardh D, She JX. Identification of Non-HLA Genes Associated with Celiac Disease and Country-Specific Differences in a Large, International Pediatric Cohort. PloS One, 11(3), e0152476, March 2016. doi: https://doi.org/10.1371/journal.pone.0152476

Table A: Variables used to replicate Supplemental Table S2 – Characteristics for Celiac Disease Autoimmunity

Table Variable	**dataset.variable**
Age	m_72_asharma_niddk_31aug2013_1.timetga m_72_asharma_niddk_31aug2013_1.tga
Country	m_72_asharma_niddk_31aug2013_1.country m_72_asharma_niddk_31aug2013_1.tga
Family history of celiac disease	m_72_asharma_niddk_31aug2013_1.fdr m_72_asharma_niddk_31aug2013_1.tga
HLA-DR, -DQ genotype	m_72_asharma_niddk_31aug2013_1.dr33 m_72_asharma_niddk_31aug2013_1.dr3x m_72_asharma_niddk_31aug2013_1.dr44 m_72_asharma_niddk_31aug2013_1.tga
HLA DPB1	m_72_asharma_niddk_31aug2013_1.hla_dpb1 m_72_asharma_niddk_31aug2013_1.tga
Gender	m_72_asharma_niddk_31aug2013_1.sex m_72_asharma_niddk_31aug2013_1.tga

Table B: Comparison of values computed in integrity check to reference article Supplemental Table S2

Characteristic	**Pub: Developed tTGA (n=703)**	**DSIC: Developed tTGA (n=703)**	**Diff. (n=0)**	**Pub: Did not develop tTGA (n=4676)**	**DSIC: Did not develop tTGA (n=4681)**	**Diff. (n=5)**
Age at first tTG+ visit or most recent visit (years)	3.06 (SD=1.33)	3.04 (1.32)	0.02 (0.01)	4.84 (SD=1.67)	4.81 (1.66)	0.03 (0.01)
Country						
U.S.	191 (27.2)	191 (27.2)	0 (0.0)	1594 (34.1)	1596 (34.1)	2 (0.0)
Finland	167 (23.7)	167 (23.7)	0 (0.0)	1247 (26.7)	1249 (26.7)	2 (0.0)
Germany	37 (5.3)	37 (5.3)	0 (0.0)	291 (6.2)	292 (6.2)	1 (0.0)
Sweden	308 (43.8)	308 (43.8)	0 (0.0)	1544 (33.0)	1544 (33.0)	0 (0.0)
Family history of celiac disease						
Yes	39 (5.6)	39 (5.6)	0 (0.0)	102 (2.2)	102 (2.2)	0 (0.0)
HLA-DR, -DQ genotype						
DR3-DQ2/DR3-DQ2	312 (44.4)	312 (44.4)	0 (0.0)	823 (17.6)	826 (17.6)	3 (0.0)
DR3-DQ2/X	272 (38.7)	272 (38.7)	0 (0.0)	1879 (40.2)	1879 (40.1)	0 (0.1)
DR4-DQ8/DR4-DQ8	93 (13.2)	93 (13.2)	0 (0.0)	943 (20.2)	943 (20.1)	0 (0.1)
Other	26 (3.7)	-	-	1031 (22.0)	-	-
HLA DPB1						
0	300 (42.7)	300 (42.7)	0 (0.0)	1537 (32.9)	1538 (32.9)	1 (0.0)
1	319 (45.4)	319 (45.4)	0 (0.0)	2279 (48.7)	2282 (48.7)	3 (0.0)
2	84 (11.9)	84 (11.9)	0 (0.0)	860 (18.4)	861 (18.4)	1 (0.0)
Gender						
Female	421 (59.9)	421 (59.9)	0 (0.0)	2215 (47.4)	2218 (47.4)	3 (0.0)

Attachment A: SAS Code

```
libname dsic "X:\NIDDK\niddk-dr_studies6\TEDDY\private_created_data\M72";

*DSIC for TEDDY M72 Sharma;

proc contents data=dsic.m_72_asharma_niddk_31aug2013_1;
run;

*create temp dataset;
data m72; set dsic.m_72_asharma_niddk_31aug2013_1;
run;

*Celiac;
proc freq data=m72;
tables celiac;
run;

*tTGA;
proc freq data=m72;
tables tga tga_tested / missing;
run;

*age;
data m72_1; set m72;
age = timetga/365.25;
run;

proc sort data=m72_1;
by tga;
run;

proc means data=m72_1 n mean std;
var age;
by tga;
run;

*Country;
proc freq data=m72;
tables country*tga/norow nopercent;
run;

*family history of celiac disease;
proc freq data=m72;
tables fdr*tga/norow nopercent;
run;
```

```
*HLA-DR, -DQ genotype;
proc freq data=m72;
tables dr33*tga/norow nopercent;
run;

proc freq data=m72;
tables dr3x*tga/norow nopercent;
run;

proc freq data=m72;
tables dr44*tga/norow nopercent;
run;

*HLA DPB1;
proc freq data=m72;
tables hla_dpb1*tga/norow nopercent;
run;

*Gender;
proc freq data=m72;
tables sex*tga/norow nopercent;
run;
```