Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) M115 Virtanen

Prepared by NIDDK-CR
December 2, 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 M115 study investigated the association between erythrocyte fatty acids and the risk of islet autoimmunity in children.

3 Archived Datasets

All data files, as provided by the Data Coordinating Center (DCC), are located in the TEDDY M115 folder in the data package. For this replication, variables were taken from the "m_115_virtanen_niddk_31may2012_1.sas7bdat" dataset.

4 Statistical Methods

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

5 Results

For Table 1 in the publication [1], Characteristics of TEDDY children with islet autoimmunity and control children, 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 Table 1. The results of the replication are an exact match to the published results.

6 Conclusions

The NIDDK Central Repository is confident that the TEDDY M115 data files to be distributed are a true copy of the study data.

7 References

[1] Niinistö S, Erlund I, Lee HS, Uusitalo U, Salminen I, Aronsson CA, Parikh HM, Liu X, Hummel S, Toppari J, She JX, Lernmark Å, Ziegler AG, Rewers M, Akolkar B, Krischer JP, Galas D, Das S, Sakhanenko N, Rich SS, Hagopian W, Norris JM, Virtanen SM. Children's Erythrocyte Fatty Acids are Associated with the Risk of Islet Autoimmunity. Scientific Reports, 11(1), 3627, February 2021. doi: https://doi.org/10.1038/s41598-021-82200-9

Table A: Variables used to replicate Table 1 – Characteristics of TEDDY children with islet autoimmunity and control children

Table Variable	**dataset.variable**
Clinical center, n (%)	m_115_virtanen_niddk_31may2012_1.outcome m_115_virtanen_niddk_31may2012_1.cc
Sex, n (%)	m_115_virtanen_niddk_31may2012_1.outcome m_115_virtanen_niddk_31may2012_1.sex
Status regarding first degree relative	m_115_virtanen_niddk_31may2012_1.outcome m_115_virtanen_niddk_31may2012_1.fdr
HLA genotype, n (%)	m_115_virtanen_niddk_31may2012_1.outcome m_115_virtanen_niddk_31may2012_1.dr34
Ancestry, mean (SD)	m_115_virtanen_niddk_31may2012_1.outcome m_115_virtanen_niddk_31may2012_1.pc1 m_115_virtanen_niddk_31may2012_1.pc2
Breastfed, n (%)	m_115_virtanen_niddk_31may2012_1.outcome m_115_virtanen_niddk_31may2012_1.newbf3_rev m_115_virtanen_niddk_31may2012_1.newbf6
Weight z-score, mean (SD)	m_115_virtanen_niddk_31may2012_1.outcome m_115_virtanen_niddk_31may2012_1.waz_3 m_115_virtanen_niddk_31may2012_1.waz_6 m_115_virtanen_niddk_31may2012_1.avwaz

Table B: Comparison of values computed in integrity check to reference article Table 1

Characteristics	Pub: Case Children (n=398)	DSIC: Case Children (n=398)	Diff. (n=0)	Pub: Control Children (n=1178)	DSIC: Control Children (n=1178)	Diff. (n=0)
Clinical center, n (%)						
Colorado	56 (14.1)	56 (14.1)	0 (0)	162 (13.8)	162 (13.8)	0 (0)
Georgia	27 (6.8)	27 (6.8)	0 (0)	78 (6.6)	78 (6.6)	0 (0)
Washington	36 (9.1)	36 (9.1)	0 (0)	107 (9.1)	107 (9.1)	0 (0)
Finland	113 (28.4)	113 (28.4)	0 (0)	339 (28.8)	339 (28.8)	0 (0)
Germany	35 (8.8)	35 (8.8)	0 (0)	105 (8.9)	105 (8.9)	0 (0)
Sweden	131 (32.9)	131 (32.9)	0 (0)	387 (32.9)	387 (32.9)	0 (0)
Sex, n (%)						
Female	178 (44.7)	178 (44.7)	0 (0)	530 (45.0)	530 (45.0)	0 (0)
Male	220 (55.3)	220 (55.3)	0 (0)	648 (55.0)	648 (55.0)	0 (0)
Status regarding first degree relative						
First degree relative with type 1 diabetes	88 (22.1)	88 (22.1)	0 (0)	259 (22.0)	259 (22.0)	0 (0)
General population	310 (77.9)	310 (77.9)	0 (0)	917 (78.0)	917 (78.0)	0 (0)
HLA genotype, n (%)						
High risk (DR3/4)	210 (52.8)	210 (52.8)	0 (0)	420 (35.7)	420 (35.7)	0 (0)
Moderate risk (other genotypes)	187 (47.0)	187 (47.0)	0 (0)	747 (63.4)	747 (63.4)	0 (0)
Missing	1 (0.2)	1 (0.2)	0 (0)	11 (0.9)	11 (0.9)	0 (0)
Ancestry, mean (SD)						
Principal component 1	0.0017 (0.0074)	0.0017 (0.0074)	0 (0)	0.0013 (0.0078)	0.0013 (0.0078)	0 (0)
Principal component 2	-0.0003 (0.0109)	-0.0003 (0.0109)	0 (0)	-0.0016 (0.0094)	-0.0016 (0.0094)	0 (0)
Breastfed, n (%)						
At 3 months	307 (77.1)	307 (77.1)	0 (0)	903 (76.7)	903 (76.7)	0 (0)
At 6 months	252 (63.3)	252 (63.3)	0 (0)	778 (66.0)	778 (66.0)	0 (0)
Missing information	1 (0.3)	1 (0.3)	0 (0)	4 (0.3)	4 (0.3)	0 (0)
Weight z-score, mean (SD)						
At 3 months	0.68 (0.95)	0.68 (0.95)	0 (0)	0.41 (1.03)	0.41 (1.03)	0 (0)
At 6 months	0.47 (1.00)	0.47 (1.00)	0 (0)	0.24 (1.01)	0.24 (1.01)	0 (0)
Over 1-6 years	0.20 (1.04)	0.20 (1.04)	0 (0)	0.01 (0.99)	0.01 (0.99)	0 (0)

Attachment A: SAS Code

```
libname m115 "X:\NIDDK\niddk-
dr_studies6\TEDDY\private_orig_data\M_115_Virtanen_NIDDK_Submission";

/*************/
/* M115 DSIC */
/*************/

data one; set m115.m_115_virtanen_niddk_31may2012_1;
run;

data two; set m115.m_115_virtanen_niddk_31may2012_2;
run;

*Clinical Center;
proc freq data=one;
tables cc*outcome/norow nopercent;
run;

*Sex;
proc freq data=one;
tables sex*outcome/norow nopercent;
run;

*Status of FDR;
proc freq data=one;
tables fdr*outcome/norow nopercent;
run;

*HLA genotype;
proc freq data=one;
tables dr34*outcome/norow nopercent missing;
run;

*Ancenstry, mean;
proc means data=one mean std;
var pc1 pc2;
class outcome;
run;

*Breastfed;
proc freq data=one;
tables (newbf3_rev newbf6)*outcome/norow nopercent missing;
run;

*weight z score;
```

```
proc means data=one mean std;
var waz_3 waz_6 avwaz;
class outcome;
run;
```