# Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) M162 Ziegler

**Prepared by NIDDK-CR**
**March 7, 2022**

# Contents

1 Standard Disclaimer ..... 2
2 Study Background ..... 2
3 Archived Datasets ..... 2
4 Statistical Methods ..... 2
5 Results ..... 3
6 Conclusions ..... 3
7 References ..... 3
Table A: Variables used to replicate Table 1 – Study characteristics by first-degree relative status ..... 4
Table B: Comparison of values computed in integrity check to reference article Table 1 values ..... 5
Attachment A: SAS Code ..... 6

# 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 M162 study sought to determine how much of the risk for autoimmunity and subsequent T1D, among those with high-risk HLA genotypes in the TEDDY cohort, can be attributable to genetic enrichment in affected families.

# 3 Archived Datasets

All SAS 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_162_AZiegler_NIDDK_18JUL2017.sas7bdat" dataset.

# 4 Statistical Methods

Analyses were performed to replicate results for the data published by Hippich et al. [1] for Genetic Contribution to the Divergence in Type 1 Diabetes Risk Between Children From the General Population and Children From Affected Families. To verify the integrity of the dataset, descriptive statistics were computed. No genetic risk scores were calculated in this replication.

# 5 Results

For Table 1 in the publication [1], Study characteristics by first-degree relative status, 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 published in Table 1. The results of the replication are within expected variation to the published results.

# 6 Conclusions

The NIDDK Central Repository is confident that the TEDDY M162 data files to be distributed are a true copy of the study data.

# 7 References

[1] Hippich M, Beyerlein A, Hagopian WA, Krischer JP, Vehik K, Knoop J, Winker C, Toppari J, Lernmark Å, Rewers MJ, Steck AK, She JX, Akolkar B, Robertson CC, Onengut-Gumuscu S, Rich SS, Bonifacio E, Ziegler AG. Genetic Contribution to the Divergence in Type 1 Diabetes Risk Between Children From the General Population and Children From Affected Families. Diabetes, 68(4), 847-857, April 2019. doi: https://doi.org/10.2337/db18-0882

**Table A: Variables used to replicate Table 1 – Study characteristics by first-degree relative status**

| **Table Variable** | **dataset.variable** |
|---|---|
| Males | M_162_AZiegler_NIDDK_18JUL2017.sex |
| HLA genotype | M_162_AZiegler_NIDDK_18JUL2017.HLA_Category |
| Country | M_162_AZiegler_NIDDK_18JUL2017.Country |
| First-degree relative with T1D | M_162_AZiegler_NIDDK_18JUL2017.family_mem |
| Outcome events | M_162_AZiegler_NIDDK_18JUL2017.persist_conf_gad<br>M_162_AZiegler_NIDDK_18JUL2017.persist_conf_miaa<br>M_162_AZiegler_NIDDK_18JUL2017.persist_conf_multiple<br>M_162_AZiegler_NIDDK_18JUL2017.t1d |

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

| Variable | FDR children – Manuscript (n=423) | FDR children – DSIC (n=423) | Diff. (n=0) | GP children – Manuscript (n=4,149) | GP children – DSIC (n=4,149) | Diff. (n=0) |
|---|---|---|---|---|---|---|
| Males | 200 (47.3) | 200 (47.3) | 0 (0) | 2082 (50.2) | 2082 (50.2) | 0 (0) |
| HLA genotype | | | | | | |
| DR3/4-DQ8 | 280 (66.2) | 280 (66.2) | 0 (0) | 2755 (66.4) | 2755 (66.4) | 0 (0) |
| DR4-DQ8/DR4-DQ8 | 143 (33.8) | 143 (33.8) | 0 (0) | 1394 (33.6) | 1394 (33.6) | 0 (0) |
| Country | | | | | | |
| U.S. | 194 (45.9) | 194 (45.9) | (0) | 1750 (42.2) | 1750 (42.2) | 0 (0) |
| Finland | 51 (12.1) | 51 (12.1) | 0 (0) | 792 (19.1) | 792 (19.1) | 0 (0) |
| Germany | 92 (21.7) | 92 (21.7) | 0 (0) | 209 (5.0) | 209 (5.0) | 0 (0) |
| Sweden | 86 (20.3) | 86 (20.3) | 0 (0) | 1398 (33.7) | 1398 (33.7) | 0 (0) |
| First-degree relative with T1D | | | | | | |
| None | 0 (0.0) | 0 (0) | 0 (0) | 4149 (100.0) | 4149 (100.0) | 0 (0) |
| Mother | 146 (34.5) | 146 (34.5) | 0 (0) | 0 (0.0) | 0 (0.0) | 0 (0) |
| Father | 180 (42.6) | 180 (42.6) | 0 (0) | 0 (0.0) | 0 (0.0) | 0 (0) |
| Sibling | 79 (18.7) | 79 (18.7) | 0 (0) | 0 (0.0) | 0 (0.0) | 0 (0) |
| Multiplex | 18 (4.3) | 18 (4.3) | 0 (0) | 0 (0.0) | 0 (0.0) | 0 (0) |
| Outcome events | | | | | | |
| One or more islet autoantibodies | 85 (20.1) | 85 (20.1) | 0 (0) | 415 (10.0) | 415 (10.0) | 0 (0) |
| Multiple islet autoantibodies | 69 (16.3) | 69 (16.3) | 0 (0) | 255 (6.1) | 255 (6.1) | 0 (0) |
| First-appearing IAA | 51 (12.1) | 73 (17.3) | 22 (5.2) | 227 (5.5) | 289 (7.0) | 62 (1.5) |
| First-appearing GADA | 46 (10.9) | 70 (16.5) | 24 (5.6) | 250 (6.0) | 332 (8.0) | 82 (2.0) |
| Diabetes | 47 (11.1) | 47 (11.1) | 0 (0) | 145 (3.5) | 145 (3.5) | 0 (0) |

# Attachment A: SAS Code

```
libname dsic "X:\NIDDK\niddk-
dr_studies6\TEDDY\private_orig_data\M_162_AZiegler_NIDDK_Submission";

/*************************************/
/* Dataset Integrity Check for TEDDY */
/* M162 */
/*************************************/

*data: limiting to the high risk HLA categories outlined in the publication;
data m162; set dsic.m_162_aziegler_niddk_18jul2017;
where HLA_Category = 1 OR HLA_Category = 2;
run;

*Sex variable by FDR status among participants with high risk HLA genotypes outlined in pub;
proc freq data=m162;
tables sex*fdr/norow nopercent;
run;

*HLA genotype variable by FDR status;
proc freq data=m162;
tables HLA_Category*fdr/norow nopercent;
run;

*Country variable by FDR status;
proc freq data=m162;
tables Country*fdr/norow nopercent;
run;

*First degree relative with T1D by FDR statust;
*making new variable for family memebers to match publication;
data m162_2; set m162;
if family_mem = 0 then fam = 0;
if family_mem = 2 OR family_mem = 3 OR family_mem = 4 then fam = 1;
if family_mem = 5 then fam = 2;
if family_mem = 6 then fam = 3;
if family_mem = 7 then fam = 4;
run;

proc freq data=m162_2;
tables Fam*FDR/norow nopercent;
run;

*Outcomes variable;
*making a new variable to match the publication;
data m162_3; set m162_2;
```

```
if persist_conf_gad = 1 OR persist_conf_ia2a = 1 OR
        persist_conf_miaa = 1 then outcome1 = 1; else outcome1 = 0;
run;

proc freq data=m162_3;
tables (outcome1 persist_conf_multiple persist_conf_miaa persist_conf_gad t1d)*FDR/ norow
nopercent;
run;

proc freq data=m162_3;
tables Persist_Conf_Gad*fdr/norow nopercent;
where Persist_Conf_Gad_Age ^= .;
run;
```