Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) M78 Endesfelder

Prepared by NIDDK-CR
June 6, 2023

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 M78 study sought to understand the differences in T1D progression among TEDDY children based on autoantibody patterns among select children from the TEDDY cohort.

3 Archived Datasets

A full listing of 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 M78 folder in the data package. For this replication, variables were taken from the "m_78_dendesfel_niddk_31dec2014_1.sas7bdat", "m_78_dendesfel_niddk_31dec2014_2.sas7bdat", and "m_78_dendesfel_niddk_31dec2014_4.sas7bdat" datasets.

4 Statistical Methods

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

5 Results

For Table 1 in the publication [1], Distribution of features among the multiple autoantibody clusters, Table A lists the variables that were used in the replication, and Tables B1 through B5 compare the results calculated from the archived data files to the results in Table 1. The results of the replication are within expected variation of the published results.

6 Conclusions

The NIDDK Central Repository is confident that the TEDDY M78 data files to be distributed are a true copy of the study data.

7 References

[1] Endesfelder D, Castell WZ, Bonifacio E, Rewers M, Hagopian WA, She JX, Lernmark Å, Toppari J, Vehik K, Williams A, Yu L, Akolkar B, Krischer JP, Ziegler AG, Achenbach P. Time-Resolved Autoantibody Profiling Facilitates Stratification of Preclinical Type 1 Diabetes in Children. Diabetes, 68(1), 119-130, January 2019. doi: https://doi.org/10.2337/db18-0594

Table A: Variables used to replicate Table 1 – Distribution of features among the multiple autoantibody clusters

Table Variable	**dataset.variable**
Age at seroconversion	m_78_dendesfel_niddk_31dec2014_4.cluster_multiple_ m_78_dendesfel_niddk_31dec2014_2.age_lastneg_samp_yr
Maternal T1D	m_78_dendesfel_niddk_31dec2014_4.cluster_multiple_ m_78_dendesfel_niddk_31dec2014_2.family
HLA	m_78_dendesfel_niddk_31dec2014_4.cluster_multiple_ m_78_dendesfel_niddk_31dec2014_2.HLA_Category
ZnT8A	m_78_dendesfel_niddk_31dec2014_4.cluster_multiple_ m_78_dendesfel_niddk_31dec2014_1.znt8a
Born by cesarean delivery	m_78_dendesfel_niddk_31dec2014_4.cluster_multiple_ m_78_dendesfel_niddk_31dec2014_2.delivery
Male sex	m_78_dendesfel_niddk_31dec2014_4.cluster_multiple_ m_78_dendesfel_niddk_31dec2014_2.sex

Table B1: Comparison of values computed in integrity check to reference article Table 1 (All clusters, mC1, and mC2)

Characteristic	Pub: All clusters	DSIC: All clusters	Diff.	Pub: mC1	DSIC: mC1	Diff.	Pub: mC2	DSIC: mC2	Diff.
Children, n	370	370	0	35	35	0	30	30	0
Age (years) at seroconversion, median (IQR)	2.0 (1.1-3.1)	1.8 (1.0-3.1)	0.2 (0.1-0)	4.0 (3.0-4.7)	3.7 (2.8-4.5)	0.3 (0.2-0.2)	4.2 (2.8-5.5)	4.5 (3.7-5.1)	0.3 (0.9-0.4)
Maternal T1D (%)	6	6	0	14	14	0	10	10	0
HLA (%)									
HLA-DR3/DR3	7	7	0	9	9	0	3	3	0
HLA-DR3/DR4	57	57	0	69	69	0	63	63	0
HLA-DR4/DR4	18	18	0	17	17	0	23	23	0
HLA-DR4/DRx	18	18	0	6	6	0	10	10	0
ZnT8A (%)	62	62	0	51	51	0	73	73	0
Born by cesarean delivery (%)	22	22	0	26	26	0	20	20	0
Male sex (%)	56	56	0	57	57	0	50	50	0

Table B2: Comparison of values computed in integrity check to reference article Table 1 (mC3, mC4, and mC5)

Characteristic	Pub: mC3	DSIC: mC3	Diff.	Pub: mC4	DSIC: mC4	Diff.	Pub: mC5	DSIC: mC5	Diff.
Children, n	12	12	0	21	21	0	27	27	0
Age (years) at seroconversion, median (IQR)	1.0 (0.8-2.0)	0.9 (0.75-2.0)	0.1 (0.05-0)	7.2 (3.3-7.5)	6.2 (4.0-7.0)	1.0 (0.7-0.5)	1.0 (0.8-1.1)	0.7 (0.6-1.0)	0.3 (0.2-0.1)
Maternal T1D (%)	8	8	0	0	0	0	7	7	0
HLA (%)									
HLA-DR3/DR3	8	8	0	5	5	0	0	0	0
HLA-DR3/DR4	33	33	0	48	48	0	44	44	0
HLA-DR4/DR4	33	33	0	14	14	0	15	15	0
HLA-DR4/DRx	25	25	0	33	33	0	41	41	0
ZnT8A (%)	83	83	0	52	52	0	44	44	0
Born by cesarean delivery (%)	50	50	0	24	24	0	37	37	0
Male sex (%)	83	83	0	52	52	0	74	74	0

Table B3: Comparison of values computed in integrity check to reference article Table 1 (mC6, mC7, and mC8)

Characteristic	Pub: mC6	DSIC: mC6	Diff.	Pub: mC7	DSIC: mC7	Diff.	Pub: mC8	DSIC: mC8	Diff.
Children, n	88	88	0	27	27	0	24	24	0
Age (years) at seroconversion, median (IQR)	1.4 (0.8-1.8)	1.1 (0.6-1.5)	0.3 (0.2-0.3)	3.3 (3.0-3.9)	3.0 (2.5-3.7)	0.3 (0.5-0.2)	2.4 (1.8-2.9)	2.2 (1.4-2.8)	0.2 (0.4-0.1)
Maternal T1D (%)	5	5	0	4	4	0	4	4	0
HLA (%)									
HLA-DR3/DR3	3	3	0	4	4	0	0	0	0
HLA-DR3/DR4	64	64	0	56	56	0	38	38	0
HLA-DR4/DR4	10	10	0	26	26	0	42	42	0
HLA-DR4/DRx	22	22	0	15	15	0	21	21	0
ZnT8A (%)	58	58	0	74	74	0	71	71	0
Born by cesarean delivery (%)	17	17	0	26	26	0	17	17	0
Male sex (%)	55	55	0	52	52	0	67	67	0

Table B4: Comparison of values computed in integrity check to reference article Table 1 (mC9, mC10, and mC11)

Characteristic	Pub: mC9	DSIC: mC9	Diff.	Pub: mC10	DSIC: mC10	Diff.	Pub: mC11	DSIC: mC11	Diff.
Children, n	16	16	0	30	30	0	19	19	0
Age (years) at seroconversion, median (IQR)	2.5 (2.3-3.0)	2.1 (2.0-2.5)	0.4 (0.3-0.5)	1.7 (1.0-2.2)	1.3 (0.8-2.0)	0.4 (0.2-0.2)	1.6 (1.1-2.0)	1.5 (1.1-2.1)	0.1 (0-0.1)
Maternal T1D (%)	6	6	0	3	3	0	0	0	0
HLA (%)									
HLA-DR3/DR3	31	31	0	17	17	0	21	21	0
HLA-DR3/DR4	56	56	0	43	43	0	68	68	0
HLA-DR4/DR4	13	13	0	27	27	0	0	0	0
HLA-DR4/DRx	0	0	0	13	13	0	11	11	0
ZnT8A (%)	75	75	0	77	77	0	47	47	0
Born by cesarean delivery (%)	25	25	0	20	20	0	16	16	0
Male sex (%)	69	69	0	50	50	0	47	47	0

Table B5: Comparison of values computed in integrity check to reference article Table 1 (mC12)

Characteristic	**Pub: mC12**	**DSIC: mC12**	**Diff.**
Children, n	41	41	0
Age (years) at seroconversion, median (IQR)	1.5 (1.0-2.3)	1.2 (0.7-2.0)	0.3 (0.3-0.3)
Maternal T1D (%)	5	5	0
HLA (%)			
HLA-DR3/DR3	7	7	0
HLA-DR3/DR4	63	63	0
HLA-DR4/DR4	15	15	0
HLA-DR4/DRx	15	15	0
ZnT8A (%)	56	56	0
Born by cesarean delivery (%)	20	20	0
Male sex (%)	41	41	0

Attachment A: SAS Code

```
libname m78 "X:\NIDDK\niddk-dr_studies6\TEDDY\private_created_data\M78";

/**********************/
/* M78 DSIC */
/* Endesfelder */
/**********************/

*temp datasets;

data one; set m78.m_78_dendesfel_niddk_31dec2014_1;
run;

data two; set m78.m_78_dendesfel_niddk_31dec2014_2;
run;

data thr; set m78.m_78_dendesfel_niddk_31dec2014_3;
run;

data fou; set m78.m_78_dendesfel_niddk_31dec2014_4;
run;

*merging datasets;
proc sort data=two;
by mask_id;
run;

proc sort data=fou;
by mask_ID;
run;

data clust; merge
two (in=a)
fou (in=b);
by mask_id;
if b=1;
run;

*Number of children in the clusters;
proc freq data=clust;
tables Cluster_multiple_;
run;

*age at seroconversion;
data clust_1; set clust;
age_lastneg_samp_yr = (age_lastneg_samp_months/12);
```

```
run;

proc means data=clust_1 median q1 q3;
var age_lastneg_samp_yr;
class Cluster_multiple_;
run;

*maternal T1d;
proc freq data=clust_1;
tables family*Cluster_multiple_/missing norow /*nopercent*/;
run;

*HLA Category;
proc freq data=clust;
tables HLA_Category*Cluster_multiple_;
run;

data clust_2; set clust;
HLA = .;
if hla_category = 0 then HLA = .;
if hla_category = 1 then HLA = 2;
if hla_category = 2 then HLA = 3;
if hla_category = 4 OR hla_category = 5 OR hla_category = 6 OR hla_category = 8 then HLA = 4;
if hla_category = 9 then HLA = 1;
run;

proc freq data=clust_2;
tables hla*cluster_multiple_/norow /*nopercent*/;
run;

*ZnT8;
data pos; set one;
keep mask_id znt8a pos;
if znt8a = 1;
pos = znt8a;
run;

data neg; set one;
keep mask_id znt8a neg;
if znt8a = 0;
neg = znt8a;
run;

proc sort data=pos nodupkey;
by mask_id;
run;

proc sort data=neg nodupkey;
```

```
by mask_id;
run;

data znt8; merge
pos (in=a)
neg (in=b);
by mask_id;
run;

data znt8_1; set znt8;
final = .;
if pos = 1 then final = 1; else final = 0;
run;

data clust_3; merge
znt8_1 (in=a)
clust_2 (in=b);
by mask_id;
if b=1;
run;

proc freq data=clust_3;
tables final*Cluster_multiple_/norow /*nopercent*/;
run;

*delivery;
proc freq data=clust_3;
tables delivery*Cluster_multiple_/norow /*nopercent*/;
run;

*male sex;
proc freq data=clust_3;
tables Sex*Cluster_multiple_/norow /*nopercent*/;
run;
```