Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) M239 Salami

Prepared by NIDDK-CR
May 21, 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. Data were 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. Specimens, including blood, stool, urine, and nail clippings, were taken at baseline and follow-up study visits. The primary outcome measures included 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 was followed for 15 years, or until the occurrence of one of the primary endpoints.

The M239 study investigated whether changes in complete blood count (CBC) in islet autoantibody positive children with increased genetic risk for type 1 diabetes were associated with oral glucose tolerance tests (OGTT) and HbA1c over time.

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_239_fsalami_niddk_30apr2019.sas7bdat" dataset.

4 Statistical Methods

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

5 Results

For Table 2 in the publication [1], Characteristics of all children in the study cohort with one or multiple persistent confirmed islet autoantibodies (IA), 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 2. The results of the replication are within expected variation to the published results.

6 Conclusions

The NIDDK Central Repository is confident that the TEDDY M239 data files to be distributed are a true copy of the study data.

7 References

[1] Salami F, Tamura RN, Larsson HE, Lernmark Å, Törn C. Complete Blood Counts with Red Blood Cell Determinants Associate with Reduced Beta-cell Function in Seroconverted Swedish TEDDY Children. Endocrinology, Diabetes & Metabolism, 4(3), e00251, May 2021. doi: https://doi.org/10.1002/edm2.251

Table A: Variables used to replicate Table 2 – Characteristics of all children in the study cohort with one or multiple persistent confirmed islet autoantibodies (IA)

Table Variable	**dataset.variable**
Gender	m_239_fsalami_niddk_30apr2019.sex m_239_fsalami_niddk_30apr2019.mult_persist_conf_ab_this_visit
HLA-DR/DQ	m_239_fsalami_niddk_30apr2019.hla_category m_239_fsalami_niddk_30apr2019.mult_persist_conf_ab_this_visit
Number of IA at first CBC	m_239_fsalami_niddk_30apr2019.cbc_sampleage m_239_fsalami_niddk_30apr2019.mult_persist_conf_ab_this_visit
Age at first CBC	m_239_fsalami_niddk_30apr2019.cbc_sampleage m_239_fsalami_niddk_30apr2019.mult_persist_conf_ab_this_visit
CBC follow-up	m_239_fsalami_niddk_30apr2019.cbc_sampleage m_239_fsalami_niddk_30apr2019.mult_persist_conf_ab_this_visit

Table B: Comparison of values computed in integrity check to reference article Table 2

Characteristic	**Pub: Persistent confirmed IA (n=89)**	**DSIC: Persistent confirmed IA (n=89)**	**Diff. (n=0)**	**Pub: Single IA (n=34)**	**DSIC: Single IA (n=34)**	**Diff. (n=0)**	**Pub: Multiple IA (n=55)**	**DSIC: Multiple IA (n=55)**	**Diff. (n=0)**
Gender									
Girls	37 (42%)	37 (42%)	0 (0)	15 (44%)	15 (44%)	0 (0)	22 (40%)	22 (40%)	0 (0)
Boys	52 (58%)	52 (58%)	0 (0)	19 (56%)	19 (56%)	0 (0)	33 (60%)	33 (60%)	0 (0)
HLA-DR/DQ									
DR3-DQ2/DR4-DQ8	47 (53%)	47 (53%)	0 (0)	17 (50%)	17 (50%)	0 (0)	30 (55%)	30 (55%)	0 (0)
DR4-DQ8/DR4-DQ8	17 (19%)	17 (19%)	0 (0)	3 (8%)	3 (9%)	0 (1)	14 (25%)	14 (25%)	0 (0)
DR4-DQ8/DR8-DQ4	14 (16%)	14 (16%)	0 (0)	7 (21%)	7 (21%)	0 (0)	7 (13%)	7 (13%)	0 (0)
DR3-DQ2/DR3-DQ2	11 (12%)	11 (12%)	0 (0)	7 (21%)	7 (21%)	0 (0)	4 (7%)	4 (7%)	0 (0)
Number of IA at first CBC									
0	6 (7%)	0 (0)	6 (7%)	4 (12%)	0 (0)	4 (12%)	2 (4%)	0 (0)	2 (4%)
1	35 (39%)	39 (43%)	4 (4%)	30 (88%)	39 (100)	9 (12%)	5 (9%)	0 (0)	5 (9%)
2	21 (24%)	23 (26%)	2 (2%)	0	0	0	21 (38%)	23 (46%)	2 (8%)
3	27 30%)	27 (30%)	0 (0)	0	0	0	27 (49%)	27 (54%)	0 (5%)
Age at first CBC (years)									
Median (SD)	8.8 (1.8)	8.9 (1.8)	0.1 (0)	9.3 (1.5)	9.3 (1.6)	0 (0.1)	8.1 (1.8)	8.3 (1.9)	0.2 (0.1)
Min-Max	5.0-12.0	4.9-12.3	0.1-0.3	5.2-12.0	5.2-12.3	0-0.3	5.0-11.4	5.0-11.4	0-0
CBC follow-up (years)									
Median (SD)	2.3 (1.7)	2.0 (1.7)	0.3 (0)	2.5 (1.8)	2.0 (1.7)	0.5 (0.1)	2.0 (1.6)	2.0 (1.6)	0 (0)
Min-Max	0.0-4.9	0.0-4.8	0-0.1	0.0-4.7	0-4.7	0-0	0.0-4.9	0.0-4.8	0-0.1

Attachment A: SAS Code

```
libname m239 "X:\NIDDK\niddk-dr_studies6\TEDDY\private_created_data\M233 &
M239\M_239_FSalami_NIDDK_Submission";

proc freq data=m239.m_239_fsalami_niddk_30apr2019;
run;

/**********************************/
/* DSIC for TEDDY M239 */
/**********************************/

data one; set m239.m_239_fsalami_niddk_30apr2019;
if persist_conf_ab_this_visit=1 OR persist_conf_gad_this_visit=1 OR
persist_conf_ia2a_this_visit=1 OR persist_conf_miaa_this_visit=1 OR mult_persist_conf_ab_this_visit=1;
run;

proc sort data=one;
by maskid due_num;
run;

data two;
set one;
by maskid due_num;
if last.maskid;
run;

*sex;
proc freq data=two;
tables sex*mult_persist_conf_ab_this_visit/norow;
run;

*HLA;
proc freq data=two;
tables hla_category*mult_persist_conf_ab_this_visit/norow;
run;

*number of IA at first CBC;
proc sort data= one;
by maskid cbc_sampleage;
run ;

data ia; set one;
by maskid cbc_sampleage;
if first.maskid;
run;
```

```
data ia2; set ia;
ia_firstcbc = 0;
if persist_conf_gad_this_visit = 0 AND persist_conf_ia2a_this_visit = 0 AND persist_conf_miaa_this_visit
= 0 then ia_firstcbc = 0;
if persist_conf_gad_this_visit = 1 AND persist_conf_ia2a_this_visit = 0 AND persist_conf_miaa_this_visit
= 0 then ia_firstcbc = 1;
if persist_conf_gad_this_visit = 0 AND persist_conf_ia2a_this_visit = 1 AND persist_conf_miaa_this_visit
= 0 then ia_firstcbc = 1;
if persist_conf_gad_this_visit = 0 AND persist_conf_ia2a_this_visit = 0 AND persist_conf_miaa_this_visit
= 1 then ia_firstcbc = 1;
if persist_conf_gad_this_visit = 1 AND persist_conf_ia2a_this_visit = 1 AND persist_conf_miaa_this_visit
= 0 then ia_firstcbc = 2;
if persist_conf_gad_this_visit = 1 AND persist_conf_ia2a_this_visit = 0 AND persist_conf_miaa_this_visit
= 1 then ia_firstcbc = 2;
if persist_conf_gad_this_visit = 0 AND persist_conf_ia2a_this_visit = 1 AND persist_conf_miaa_this_visit
= 1 then ia_firstcbc = 2;
if persist_conf_gad_this_visit = 1 AND persist_conf_ia2a_this_visit = 1 AND persist_conf_miaa_this_visit
= 1 then ia_firstcbc = 3;
run;

proc freq data=ia2;
tables ia_firstcbc*mult_persist_conf_ab_this_visit;
run;

*Age at first CBC;
proc sort data=one;
by maskid cbc_sampleage;
run;

data three; set one;
by maskid cbc_sampleage;
if first.maskid;
run;

data four; set three;
cbc_year = cbc_sampleage/12;
run;

proc means data=four n median mean std min max;
var cbc_year;
class mult_persist_conf_ab_this_visit;
run;

*follow-up;
data five; set one;
cbc_year = cbc_sampleage/12;
run;
```

```
data first; set five;
by maskid cbc_year;
        if first.maskid;
run;

data last; set five;
by maskid cbc_year;
        if last.maskid;
run;

data last1; set last;
last_year = cbc_year;
keep maskid last_year mult_persist_conf_ab_this_visit;
run;

data first1; set first;
first_year = cbc_year;
keep maskid first_year mult_persist_conf_ab_this_visit;
run;

proc sort data=last1;
by maskid;
run;

proc sort data=first1;
by maskid;
run;

data follow; merge
first1 (in=a)
last1 (in=b);
by maskid;
run;

data follow1; set follow;
follow = last_year - first_year;
run;

proc means data=follow1 n median std min max;
var follow;
class mult_persist_conf_ab_this_visit;
run;
```