Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) Pub24 Larsson

Prepared by Sabrina Chen
IMS Inc.
3901 Calverton Blvd, Suite 200 Calverton, MD 20705
August 7, 2018

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 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.

3 Archived Datasets

All the 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_24_hlarsson_niddk_30apr2014_1.sas7bdat" datasets.

4 Statistical Methods

Analyses were performed to duplicate results for the data published by Larsson et al [1] Diabetes in 2016. To verify the integrity of the dataset, descriptive statistics were computed.

5 Results

For Table 1 in the publication [1], Characteristics of the data analyzed, the results of the replication are similar to the published results.

6 Conclusions

The NIDDK repository is confident that the TEDDY M54 data files to be distributed are a true copy of the study data.

7 References

[1] Helena Elding Larsson, Kendra Vehik, Michael J. Haller, Xiang Liu, Beena Akolkar, William Hagopian, Jeffrey Krischer, Åke Lernmark, Jin-Xiong She, Olli Simell, Jorma Toppari, Anette-G. Ziegler, and Marian Rewers, for the TEDDY Study Group. "Growth and Risk for Islet Autoimmunity and Progression to Type 1 Diabetes in Early Childhood: The Environmental Determinants of Diabetes in the Young Study". Diabetes 2016;65:1988–1995.

Table A: Variables used to replicate Table 1: Characteristics of the data analyzed

Table Variable	**dataset.variable**
Developed IA/Did not develop IA	m_24_hlarsson_niddk_30apr2014_1.PERSIST_CONF_AB
Developed type 1 diabetes/Did not develop type 1 diabetes	m_24_hlarsson_niddk_30apr2014_1.T1D
Age at first autoantibody-postive visit/diagnosis of type 1 diabetes or most recent visit (months)	m_24_hlarsson_niddk_30apr2014_1.AGEPERSIST
Country	m_24_hlarsson_niddk_30apr2014_1.COUNTRY
Family History of type 1 diabetes	m_24_hlarsson_niddk_30apr2014_1.FDR
High-risk HLA-DR or –DQ genotype DR3/4	m_24_hlarsson_niddk_30apr2014_1.HLA_DR34
Sex	m_24_hlarsson_niddk_30apr2014_1.FEMALE

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

	Developed IA			Did not develop IA			Developed type 1 diabetes			Did not develop type 1 diabetes		
	Manuscript	DSIC	Diff	Manuscript	DSIC	Diff	Manuscript	DSIC	Diff	Manuscript	DSIC	Diff
	(n = 575)	(n = 575)	0	(n = 6,893)	(n = 6,893)	0	(n = 169)	(n = 169)	0	(n = 7,299)	(n = 7,299)	0
Age at first autoantibody-positive visit/diagnosis of type 1 diabetes or most recent visit (months)	31 (21)	31 (21)	0(0)	61 (26)	61 (26)	0(0)	45 (23)	20 (13)	25(10)	64 (26)	60 (27)	4(-1)
Country												
Finland	25 (144)	25 (144)	0(0)	22 (1,513)	22 (1,513)	0(0)	31 (52)	31 (52)	0(0)	22 (1,605)	22 (1,605)	0(0)
Germany	8 (46)	8 (46)	0(0)	7 (462)	7 (462)	0(0)	12 (21)	12 (21)	0(0)	7 (487)	7 (487)	0(0)
Sweden	34 (194)	34 (194)	0(0)	30 (2,069)	30 (2,069)	0(0)	27 (45)	27 (45)	0(0)	30 (2,218)	30 (2,218)	0(0)
U.S.	33 (191)	33 (191)	0(0)	41 (2,849)	41 (2,849)	0(0)	30 (51)	30 (51)	0(0)	41 (2,989)	41 (2,989)	0(0)
Family history of type 1 diabetes												
Yes	22 (124)	22 (124)	0(0)	10 (714)	10 (714)	0(0)	32 (54)	32 (54)	0(0)	11 (784)	11 (784)	0(0)
High-risk HLA-DR or -DQ genotype												
DR3/4	50 (290)	50 (290)	0(0)	38 (2,642)	38 (2,642)	0(0)	55 (93)	55 (93)	0(0)	39 (2,839)	39 (2,839)	0(0)
Sex												
Female	46 (250)	43 (250)	3(0)	49 (3,407)	49 (3,407)	0(0)	46 (77)	46 (77)	0(0)	49 (3,580)	49 (3,580)	0(0)

Attachment A: SAS Code

```
*** TEDDY M54 DSIC;
*** Programmer: Sabrina Chen
*** Date: 6/27/18;

options nocenter validvarname=upcase;

title '/prj/niddk/ims_analysis/TEDDY/prog_initial_analysis/m_24_dsic.sas';
run;


libname dat '/prj/niddk/ims_analysis/TEDDY/private_orig_data/M_24_HLarsson_NIDDK_Submission';


proc format;
 value noyes
 .  = "no"
 other = "yes"
 ;
run;


data m24_1;
        set dat.m_24_hlarsson_niddk_30apr2014_1;
run;

proc contents data = m24_1;
title3 'm24_1';
run;

data m24_2;
        set dat.m_24_hlarsson_niddk_30apr2014_2;
run;

proc contents data = m24_2;
title3 'm24_2';
run;


data m24_3;
        set dat.m_24_hlarsson_niddk_30apr2014_3;
run;

proc contents data = m24_3;
title3 'm24_3';
run;


data m24_4;
        set dat.m_24_hlarsson_niddk_30apr2014_4;
run;

proc contents data = m24_4;
title3 'm24_4';
run;


** check for dups;
proc sort data=m24_1;
  by maskid;
  run;

data dup1;
  set m24_1;
  by maskid;
  if not (first.maskid and last.maskid);
run;
```

```
proc sort data=m24_2;
  by maskid;
  run;

data dup2;
  set m24_2;
  by maskid;
  if not (first.maskid and last.maskid);
run;

proc sort data=m24_3;
  by maskid;
  run;

data dup3;
  set m24_3;
  by maskid;
  if not (first.maskid and last.maskid);
run;

proc sort data=m24_4;
  by maskid;
  run;

data dup4;
  set m24_4;
  by maskid;
  if not (first.maskid and last.maskid);
run;

data dupcheck;
  merge dup1 (in=in1 keep=maskid)
        dup2 (in=in2 keep=maskid)
        dup3 (in=in3 keep=maskid)
        dup4 (in=in4 keep=maskid);
  by maskid;
  if in1 then in_1=1;
  if in2 then in_2=1;
  if in3 then in_3=1;
  if in4 then in_4=1;
run;

proc freq data=dupcheck;
  tables in_1*in_2*in_3*in_4/list missing;
title3 'check overlap of dups';
run;


proc print data=m24_2 (obs=25);
  by maskid;
  id maskid;
  run;


* check overlap at subject level;
proc sort data=m24_1 out=per1 nodupkey;
  by maskid;
  run;

proc sort data=m24_2 out=per2 nodupkey;
  by maskid;
  run;

proc sort data=m24_3 out=per3 nodupkey;
  by maskid;
  run;

proc sort data=m24_4 out=per4 nodupkey;
  by maskid;
  run;
```

```
data overlap;
  merge per1 (in=in1 keep=maskid)
        per2 (in=in2 keep=maskid)
        per3 (in=in3 keep=maskid)
        per4 (in=in4 keep=maskid);
  by maskid;
  if in1 then in_1=1;
  if in2 then in_2=1;
  if in3 then in_3=1;
  if in4 then in_4=1;
run;

proc freq data=overlap;
  tables in_1*in_2*in_3*in_4/list missing;
  title3 'Checking overlap at subject level';
run;


proc freq data=m24_1;
  tables /*AGEPERSIST
AGET1D      */
COUNTRY
FDR
FEMALE
HLA_DR34
/*HTZ_0Y
HTZ_1Y
HTZ_2Y
HTZ_3Y       */
MULTI_PERSIST_AB
PERSIST_CONF_AB
T1D
/*TIME_AB
TIME_MULTIAB
TIME_P_T1D
TIME_T1D
WTZ_0Y
WTZ_1Y
WTZ_2Y
WTZ_3Y */     /missing;
title3 'File 1 - raw freq';
run;


* Table 1;
proc univariate data=m24_1;
  where PERSIST_CONF_AB=1;
  var AGEPERSIST;
title3 "Table 1 column: Developed IA";
run;

proc freq data=m24_1;
  where PERSIST_CONF_AB=1;
  tables country fdr HLA_DR34 female/missing;
run;

proc univariate data=m24_1;
  where PERSIST_CONF_AB=0;
  var AGEPERSIST;
title3 "Table 1 column: Did not develop IA";
run;

proc freq data=m24_1;
  where PERSIST_CONF_AB=0;
  tables country fdr HLA_DR34 female/missing;
run;


proc univariate data=m24_1;
  where T1D=1;
  var AGEPERSIST;
```

```
title3 "Table 1 column: Developed type 1 diabetes";
run;

proc freq data=m24_1;
  where T1D=1;
  tables country fdr HLA_DR34 female/missing;
run;

proc univariate data=m24_1;
  where T1D=0;
  var AGEPERSIST;
title3 "Table 1 column: Did not develop type 1 diabetes";
run;

proc freq data=m24_1;
  where T1D=0;
  tables country fdr HLA_DR34 female/missing;
run;
```