Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) M128 Aronsson

Prepared by Sabrina Chen
IMS Inc.
3901 Calverton Blvd, Suite 200 Calverton, MD 20705
February 12, 2020

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/private_orig_data/M_128_Aronsson_NIDDK_Submission folder in the data package. For this replication, variables were taken from the “m_128_aronsson_niddk_30sep2017_1.sas7bdat” dataset.

4 Statistical Methods

Analyses were performed to duplicate results for the data published by Carin Andrén Aronsson et al [1] in the JAMA in 2019. To verify the integrity of the dataset, descriptive statistics were computed.

5 Results

For **Comparison of Data in the publication**, 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.

6 Conclusions

The results of the replication are almost an exact match to the published results.

7 References

[1] Carin Andrén Aronsson, PhD; Hye-Seung Lee, PhD; Elin M. Hård af Segerstad, MSc; Ulla Uusitalo, PhD; Jimin Yang, PhD; Sibylle Koletzko, MD, PhD; Edwin Liu, MD, PhD; Kalle Kurppa, MD, PhD; Polly J. Bingley, MD; Jorma Toppari, MD, PhD; Anette G. Ziegler, MD; Jin-Xiong She, PhD; William A. Hagopian, MD, PhD; Marian Rewers, MD, PhD; Beena Akolkar, PhD; Jeffrey P. Krischer, PhD; Suvi M. Virtanen, MD, PhD; Jill M. Norris, MPH, PhD; Daniel Agardh, MD, PhD; for the TEDDY Study Group. Association of Gluten Intake During the First 5 Years of Life With Incidence of Celiac Disease Autoimmunity and Celiac Disease Among Children at Increased Risk. JAMA (2019);322(6):1-10.

Table A: Variables used to replicate data in the publication.

Table Variable	**dataset.variable**
Sex of the subject - Male, Female	m_128_aronsson_niddk_30sep2017_1.sex
Country	m_128_aronsson_niddk_30sep2017_1.country
First degree relative status for Celiac Disease	m_128_aronsson_niddk_30sep2017_1.celiac_fdr
TG single positivity - 1 = Yes, 0 = No	m_128_aronsson_niddk_30sep2017_1.htg_pos
HLA - 1 = 'DR3/DR3', 2 = 'DR3/4', 3 = 'DR4/4 or DR4/8', 4 = 'FDR HLAs'	m_128_aronsson_niddk_30sep2017_1.egehla
Duration of breastfeeding (months)	m_128_aronsson_niddk_30sep2017_1.mbfdur
Age of gluten introduction (months)	m_128_aronsson_niddk_30sep2017_1.mgluten

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

	Always Negative for tTG Autoantibodies					
	Manuscript	**DSIC**	**Diff**	**Manuscript**	**DSIC**	**Diff**
	COUNT			PERCENT		
Female	2453	2453	0	47.2	47.2	0
Male	2741	2741	0	52.8	52.8	0
(0) Total	1218	1218	0	23.5	23.5	0
(1) DR3/DR3	124	124	0	10.2	10.2	0
(2) DR3/4	376	376	0	30.9	30.9	0
(3) Other HLA antigen genotypes	718	718	0	58.9	58.9	0
(0) Total	314	314	0	6.1	6	0.1
(1) DR3/DR3	50	50	0	15.9	15.9	0
(2) DR3/4	131	131	0	41.7	41.7	0
(3) Other HLA antigen genotypes	133	133	0	42.4	42.4	0
(0) Total	1554	1554	0	29.9	29.9	0
(1) DR3/DR3	225	225	0	14.5	14.5	0
(2) DR3/4	690	690	0	44.4	44.4	0
(3) Other HLA antigen genotypes	639	639	0	41.1	41.1	0
(0) Total	2108	2108	0	40.5	40.6	-0.1
(1) DR3/DR3	391	391	0	18.5	18.5	0
(2) DR3/4	849	849	0	40.3	40.3	0
(3) Other HLA antigen genotypes	868	868	0	41.2	41.2	0
Mother, father or sibling has CD	129	129	0	2.5	2.5	0

	Manuscript	**DSIC**	**Diff**	**Manuscript**	**DSIC**	**Diff**	**Manuscript**	**DSIC**	**Diff**
	Median			Q1			Q3		
Breastfeeding duration, median (IQR), mo	7.8	7.9	-0.1	3.5	3.4	0.1	12	12	0
	Manuscript	**DSIC**	**Diff**	**Manuscript**	**DSIC**	**Diff**			
	Mean			SD					
Age at gluten introduction, mean (SD), mo	6.2	6.2	0	1.9	1.9	0			

Attachment A: SAS Code

```
options nocenter validvarname=upcase;

title '/prj/niddk/ims_analysis/TEDDY/prog_initial_analysis/m_128_dsic.sas';
run;

libname dat '/prj/niddk/ims_analysis/TEDDY/private_orig_data/M_128_Aronsson_NIDDK_Submission';


proc format;
 value val
 .     = "no value"
 other = "   value"
 ;

 value oneplus
 . = "no value"
 0 = "0"
 1-high = "1+"
 ;

 value zerohi
 . = "no value"
 0-high = "0-high"
 ;


 value country
 1 = "USA"
 2 = "Finland"
 3 = "Germany"
 4 = "Sweden"
 ;

 value sexf
 0 = 'Male'
 1 = 'Female'
 ;

 value hlaf
 1   = '(1) DR3/DR3'
 2   = '(2) DR3/4'
```

```
 3,4 = '(3) Other HLA antigen genotypes'
 ;

 value fdrcd
 0 = 'General Pop'
 1 = 'Mother, father or sibling has CD'
 ;

run;


* produce n and %;
%macro npercent(rownum, var, varf, subset,  subsetname);
  proc freq data=analy noprint;
    where HTG_POS=0 and &subset = 1;
    tables &var/list missing out=tbl1&subsetname;
    format &var &varf..;
  run;

  data tbl1&subsetname;
    length covar covarf $100;
    set tbl1&subsetname;
    covar = "&var";
    covarf = put(&var,&varf..);
    rownum = &rownum;
  run;

  data prnt&subsetname;
    set prnt&subsetname tbl1&subsetname;
  run;

%mend;



data analy;
  set dat.m_128_aronsson_niddk_30sep2017_1;
run;

proc contents data=analy;
run;

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

```
data dups1;
  set analy;
  by maskid;
  if not (first.maskid and last.maskid);
run;

proc print data=dups1;
  by maskid;
  id maskid;
run;

proc freq data=analy;
  where HTG_POS=0;
  format maskid TIME_CONFTGA MTIME_CONFTGA TIMETOCD MTIMETOCD MBFDUR MGLUTEN FU_YEAR val.;
run;


data analy2;
  set dat.m_128_aronsson_niddk_30sep2017_2;
run;

proc contents data=analy2;
run;

proc sort data=analy2;
  by maskid;
run;

data dups2;
  set analy2;
  by maskid;
  if not (first.maskid and last.maskid);
run;

proc print data=dups2 (obs=25);
  by maskid;
  id maskid;
run;




data analy;
   set analy;
```

```
   * create subset flag for each row to use in macro call;
   all = 1;
   * country;
   if country=1 then usa=1;
   else if country=2 then Finland=1;
   else if country=3 then Germany=1;
   else if country=4 then Sweden=1;

   if sex = 'Male' then sexnum = 0;
   else if sex = 'Female' then sexnum=1;

   if EGEHLA in(3,4) then EGEHLA_gp = 3;
   else EGEHLA_gp = EGEHLA;
run;


proc freq data=analy;
  tables all * country * usa*Finland*Germany*Sweden/list missing;
  tables sexnum*sex/list missing;
  tables EGEHLA_gp*EGEHLA/list missing;
  tables HTG_POS*MBFDUR/list missing;
  tables HTG_POS*MGLUTEN/list missing;
  format MBFDUR MGLUTEN val.;
run;

* n and percent;
data prntall;
  set _null_;
run;

%npercent(1, sexnum         , sexf    , all       ,all);
%npercent(2, country        , country , all       ,all);
%npercent(3, CELIAC_FDR     , fdrcd   , all       ,all);



data prntusa;
  set _null_;
run;

%npercent(4, EGEHLA_gp         , hlaf , usa       , usa);
```

```
data prntfinland;
  set _null_;
run;

%npercent(4, EGEHLA_gp         , hlaf , finland       , finland);



data prntgermany;
  set _null_;
run;

%npercent(4, EGEHLA_gp         , hlaf , germany      , germany);



data prntsweden;
  set _null_;
run;

%npercent(4, EGEHLA_gp         , hlaf , sweden       , sweden);



* Table 1;
data npercent;
  length subgroup $10;
  set prntall (in=in1) prntusa (in=in2) prntfinland (in=in3) prntgermany (in=in4) prntsweden (in=in5);

  if in1 then do;
    if      covarf = "Finland" then do;
      subgroup = "finland";
      covarf = "(0) Total";
    end;
    else if covarf = "Germany" then do;
      subgroup = "germany";
      covarf = "(0) Total";
    end;
    else if covarf = "Sweden"  then do;
      subgroup = "sweden" ;
      covarf = "(0) Total";
    end;
    else if covarf = "USA"     then do;
      subgroup = "usa"    ;
      covarf = "(0) Total";
```

```
    end;
    else if covarf in("Male","Female") then subgroup = "1 gender";
    else if covarf in('General Pop', 'Mother, father or sibling has CD') then subgroup = "z FDR";
  end;

  if in2 then subgroup = "usa";
  if in3 then subgroup = "finland";
  if in4 then subgroup = "germany";
  if in5 then subgroup = "sweden";

  percent = put(percent,8.1);
run;

proc sort data=npercent;
  by  subgroup covarf;
run;

proc print data=npercent;
  var rownum subgroup covar covarf count percent;
  title3 "Table 1 - n, percent";
run;

ods listing close;
ods phtml file="/prj/niddk/ims_analysis/TEDDY/private_created_data/TEDDY.m128.Table1.xls";

proc print data=npercent;
  var rownum subgroup covar covarf count percent;
  title3 "Table 1 - n, percent";
run;

ods phtml close;
ods listing;


proc univariate data=analy;
  where HTG_POS=0 ;
  var MBFDUR;
title3 "Breastfeeding duration";
run;

proc means data=analy;
  where HTG_POS=0 ;
  var MGLUTEN;
title3 "Age at gluten introduction";
run;
```