Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) M113 Törn

Prepared by Sabrina Chen
IMS Inc.
3901 Calverton Blvd, Suite 200 Calverton, MD 20705
February 28, 2019

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_113_ctorn_niddk_31dec2014.sas7bdat" dataset.

4 Statistical Methods

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

5 Results

For Table 1 in the publication [1], Characteristics of subjects by the status of islet autoimmunity (IA) and type 1 diabetes (T1D) in The Environmental Determinants of Diabetes in the Young (TEDDY) study, 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 similar to the published results.

6 Conclusions

The NIDDK repository is confident that the TEDDY M113 data files to be distributed are a true copy of the study data.

7 References

Törn, C. *et al.* Complement gene variants in relation to autoantibodies to beta cell specific antigens and type 1 diabetes in the TEDDY Study. *Sci. Rep.* 6, 27887; doi: 10.1038/srep27887 (2016).

Table A: Variables used to replicate Table 1: Characteristics of subjects by the status of islet autoimmunity (IA) and type 1 diabetes (T1D) in The Environmental Determinants of Diabetes in the Young (TEDDY) study.

Table Variable	**dataset.variable**
Age at first IA, T1D or the most recent visit	m_113_ctorn_niddk_31dec2014.agepersist
Age at first IA, T1D or the most recent visit	m_113_ctorn_niddk_31dec2014.aget1d
Country	m_113_ctorn_niddk_31dec2014.country
High-risk HLA-DR-DQ genotype	m_113_ctorn_niddk_31dec2014.hla_category
Gender	m_113_ctorn_niddk_31dec2014.female

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

Variable	**Category**	**IA (N)**		
		Manuscript	**DSIC**	**Diff**
Number of subjects (n)		413	413	0
Age at first IA, T1D or the most recent visit (months)	Median	27.9	27.9	0
	IQR	15.4	15.4	0
Country n (%)	Finland	125	125	0
	Germany	16	16	0
	Sweden	159	159	0
	US	113	113	0
High-risk HLA-DR-DQ	DR3/4	218	218	0
	DR4/4	78	78	0
	DR4/8	66	66	0
	DR3/3	51	51	0
Gender n (%)	Female	170	170	0

Variable	**Category**	**IA (%)**		
		Manuscript	**DSIC**	**Diff**
Number of subjects (n)				
Age at first IA, T1D or the most recent visit (months)	Median			
	IQR	47.8	47.8	0
Country n (%)	Finland	30.3	30.3	0
	Germany	3.9	3.9	0
	Sweden	38.5	38.5	0
	US	27.3	27.4	-0.1
High-risk HLA-DR-DQ	DR3/4	52.8	52.8	0
	DR4/4	18.9	18.9	0
	DR4/8	16.0	16	0
	DR3/3	12.3	12.3	0
Gender n (%)	Female	41.2	41.2	0

Variable	Category	No IA (N)		
		Manuscript	DSIC	Diff
Number of subjects (n)		5061	5061	0
Age at first IA, T1D or the most recent visit (months)	Median	69.9	69.9	0
	IQR	49.7	49.7	0
Country n (%)	Finland	1327	1327	0
	Germany	235	235	0
	Sweden	1746	1746	0
	US	1753	1753	0
High-risk HLA-DR-DQ	DR3/4	1986	1986	0
	DR4/4	1008	1008	0
	DR4/8	900	900	0
	DR3/3	1167	1167	0
Gender n (%)	Female	2487	2487	0

Variable	Category	No IA (%)		
		Manuscript	DSIC	Diff
Number of subjects (n)				
Age at first IA, T1D or the most recent visit (months)	Median			
	IQR	89.5	89.5	0
Country n (%)	Finland	26.2	26.2	0
	Germany	4.6	4.6	0
	Sweden	34.5	34.5	0
	US	34.7	34.6	0
High-risk HLA-DR-DQ	DR3/4	39.2	39.2	0
	DR4/4	19.9	19.9	0
	DR4/8	17.8	17.8	0
	DR3/3	23.1	23.1	0
Gender n (%)	Female	49.1	49.1	0

Variable	Category	T1D (N)		
		Manuscript	DSIC	Diff
Number of subjects (n)		115	115	0
Age at first IA, T1D or the most recent visit (months)	Median	51.4	51.4	0
	IQR	29.2	29.2	0
Country n (%)	Finland	41	41	0
	Germany	6	6	0
	Sweden	39	39	0
	US	29	29	0
High-risk HLA-DR-DQ	DR3/4	65	65	0
	DR4/4	19	19	0
	DR4/8	16	16	0
	DR3/3	15	15	0
Gender n (%)	Female	50	50	0

Variable	Category	T1D (%)		
		Manuscript	DSIC	Diff
Number of subjects (n)				
Age at first IA, T1D or the most recent visit (months)	Median			
	IQR	67.5	67.5	0
Country n (%)	Finland	35.7	35.7	0
	Germany	5.2	5.2	0
	Sweden	33.9	33.9	0
	US	25.2	25.2	0
High-risk HLA-DR-DQ	DR3/4	56.5	56.5	0
	DR4/4	16.5	16.5	0
	DR4/8	13.9	13.9	0
	DR3/3	13.1	13	0
Gender n (%)	Female	43.5	43.5	0

Variable	Category	No T1D (N)		
		Manuscript	DSIC	Diff
Number of subjects (n)		5359	5359	0
Age at first IA, T1D or the most recent visit (months)	Median	72.9	72.9	0
	IQR	54.7	54.7	0
Country n (%)	Finland	1411	1411	0
	Germany	245	245	0
	Sweden	1866	1866	0
	US	1837	1837	0
High-risk HLA-DR-DQ	DR3/4	2139	2139	0
	DR4/4	1067	1067	0
	DR4/8	950	950	0
	DR3/3	1203	1203	0
Gender n (%)	Female	2607	2607	0

Variable	Category	No T1D (%)		
		Manuscript	DSIC	Diff
Number of subjects (n)				
Age at first IA, T1D or the most recent visit (months)	Median			
	IQR	91.1	91.1	0
Country n (%)	Finland	26.3	26.3	0
	Germany	4.6	4.6	0
	Sweden	34.8	34.8	0
	US	34.3	34.3	0
High-risk HLA-DR-DQ	DR3/4	39.9	39.9	0
	DR4/4	19.9	19.9	0
	DR4/8	17.7	17.7	0
	DR3/3	22.5	22.4	0.1
Gender n (%)	Female	48.7	48.6	0.1

Attachment A: SAS Code

```
options nocenter validvarname=upcase;

title '/prj/niddk/ims_analysis/TEDDY/prog_initial_analysis/m_113_dsic.sas';
run;

***********;
* INPUT    ;
***********;
libname sasfile1 '/prj/niddk/ims_analysis/TEDDY/private_orig_data/M_113_CTorn_NIDDK_ Submission/';

************;
* MACROS    ;
************;
%macro readin(lib, ds);
  data &ds;
    set sasfile&lib..&ds;
  run;

  proc contents data=&ds;
  title3 "&ds";
  run;
%mend;


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


%macro univ(rownum, var, subset, subsetname);

  proc univariate data=analy outtable= univ&subsetname noprint;
    where &subset=1;
    var &var
       ;
  run;

  data univ&subsetname;
    length covarf $100;
    set univ&subsetname;
        covarf = "&subset";
        rownum = &rownum;
  run;

  data prntuniv&subsetname;
    set prntuniv&subsetname univ&subsetname;
  run;

%mend;



*************;
* FORMATS    ;
*************;
proc format;
  value novalue
      . = "No Value"
  other = "   Value"
  ;

  value negpos
  0 = "Negative"
  1 = "Positive"
  ;

  value yesno
  0 = "No"
  1 = "Yes"
```

```
  ;

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

  value country
  1="US"
  2="FIN"
  3="GER"
  4="SWE"
  ;

  value hlaf
  1='DR3/4'
  2='DR4/4'
  4='DR4/8'
  9='DR3/3'
  ;

run;


%readin(1, m_113_ctorn_niddk_31dec2014);


proc freq data=m_113_ctorn_niddk_31dec2014;
  tables persist_conf_ab*t1d/list missing;
run;

data analy;
  set m_113_ctorn_niddk_31dec2014;
  * create subset flag for each row to use in macro call;
  all = 1;

  if persist_conf_ab=0 then subset_no_ia = 1;
  else if persist_conf_ab=1 then subset_ia = 1;

  if t1d=1 then subset_t1d = 1;
  else if t1d=0 then subset_no_t1d = 1;

run;

proc freq data=analy;
```

```
  tables persist_conf_ab*t1d* subset_ia*subset_no_ia*subset_no_t1d*subset_t1d/list missing;
  tables subset_ia subset_no_ia subset_no_t1d subset_t1d/list missing;
run;



* med, min and max;
data prntunivia;
  length _VAR_ $100;
  set _null_;
run;

%univ(1   , AGEPERSIST  , subset_ia , ia);

proc print data= prntunivia noobs;
    var rownum _var_  covarf /*_nobs_*/ _median_ _q1_ _q3_ /*_min_ _max_ _std_*/;
run;

data prntunivnoia;
  length _VAR_ $100;
  set _null_;
run;

%univ(1   , AGEPERSIST  , subset_no_ia , noia);

proc print data= prntunivnoia noobs;
    var rownum _var_  covarf /*_nobs_*/ _median_ _q1_ _q3_ /*_min_ _max_ _std_*/;
run;

data prntunivt1d;
  length _VAR_ $100;
  set _null_;
run;

%univ(1   , aget1d      , subset_t1d , t1d);

proc print data= prntunivt1d noobs;
    var rownum _var_  covarf /*_nobs_*/ _median_ _q1_ _q3_ /*_min_ _max_ _std_*/;
run;

data prntunivnot1d;
  length _VAR_ $100;
  set _null_;
run;
```

```
%univ(1   , aget1d     , subset_no_t1d , not1d);

proc print data= prntunivnot1d noobs;
    var rownum _var_  covarf /*_nobs_*/ _median_ _q1_ _q3_ /*_min_ _max_ _std_*/;
run;

* combine;
proc sort data=prntunivia;
  by rownum ;
run;

proc sort data=prntunivnoia (rename=(_median_ = noia_median_
                                     _q1_     = noia_q1_
                                    _q3_     = noia_q3_ ))
                                    ;
  by rownum ;
run;

proc sort data=prntunivt1d (rename=(_median_ = t1d_median_
                                     _q1_    = t1d_q1_
                                    _q3_    = t1d_q3_ ))
                                    ;
  by rownum ;
run;

proc sort data=prntunivnot1d (rename=(_median_ = not1d_median_
                                     _q1_      = not1d_q1_
                                    _q3_      = not1d_q3_ ))
                                    ;
  by rownum ;
run;


data alluniv;
  merge prntunivia   (in=in1 keep = rownum _var_  _median_  _q1_      _q3_)
        prntunivnoia (in=in2 keep = rownum _var_   noia_median_  noia_q1_  noia_q3_)
        prntunivt1d (in=in3 keep = rownum _var_    t1d_median_   t1d_q1_   t1d_q3_)
        prntunivnot1d (in=in4 keep = rownum _var_  not1d_median_ not1d_q1_ not1d_q3_)
        ;
  by rownum;
  if in1 or in2 or in3 or in4;
  _median_     = round(_median_      , 0.1);
  _q1_         = round(_q1_          , 0.1);
  _q3_         = round(_q3_          , 0.1);
  noia_median_          = round(noia_median_ , 0.1);
```

```
  t1d_median_   = round(t1d_median_  , 0.1);
  not1d_median_  = round(not1d_median_, 0.1);
  noia_q1_      = round(noia_q1_     , 0.1);
  t1d_q1_       = round(t1d_q1_      , 0.1);
  not1d_q1_     = round(not1d_q1_    , 0.1);
  noia_q3_      = round(noia_q3_     , 0.1);
  t1d_q3_       = round(t1d_q3_      , 0.1);
  not1d_q3_     = round(not1d_q3_    , 0.1);
run;


* n percent;
data prntia;
  set _null_;
run;

%npercent(2   , country            , country,        subset_ia , ia);
%npercent(3   , hla_category , hlaf   ,  subset_ia , ia);
%npercent(4   , female     , sexf   ,         subset_ia , ia);

proc print data=prntia;
  var rownum covar covarf count    percent;
run;


data prntnoia;
  set _null_;
run;

%npercent(2   , country            , country,        subset_no_ia , noia);
%npercent(3   , hla_category , hlaf   ,  subset_no_ia , noia);
%npercent(4   , female     , sexf   ,         subset_no_ia , noia);

proc print data=prntnoia;
  var rownum covar covarf count    percent;
run;


data prntt1d;
  set _null_;
run;

%npercent(2   , country            , country,        subset_t1d , t1d);
%npercent(3   , hla_category , hlaf   ,  subset_t1d , t1d);
%npercent(4   , female     , sexf   ,         subset_t1d , t1d);
```

```
proc print data=prntt1d;
  var rownum covar covarf count    percent;
run;


data prntnot1d;
  set _null_;
run;

%npercent(2   , country            , country,         subset_no_t1d , not1d);
%npercent(3   , hla_category , hlaf   ,  subset_no_t1d , not1d);
%npercent(4   , female      , sexf   ,         subset_no_t1d , not1d);

proc print data=prntnot1d;
  var rownum covar covarf count    percent;
run;

* combine ;
proc sort data=prntia;
  by rownum covar covarf;
run;

proc sort data=prntnoia (rename=(count    = count_noia
                                 percent  = percent_noia))
                                 ;
  by rownum covar covarf;
run;

proc sort data=prntt1d  (rename=(count    = count_t1d
                                 percent  = percent_t1d))
                                 ;
  by rownum covar covarf;
run;

proc sort data=prntnot1d  (rename=(count    = count_not1d
                                    percent  = percent_not1d))
                                 ;
  by rownum covar covarf;
run;


data allnpercent;
  merge prntia    (in=in1 keep = rownum covar covarf  count        percent   )
        prntnoia  (in=in2 keep = rownum covar covarf  count_noia   percent_noia)
```

```
       prntt1d   (in=in3 keep = rownum covar covarf  count_t1d    percent_t1d)
       prntnot1d (in=in4 keep = rownum covar covarf  count_not1d  percent_not1d)
       ;
  by rownum covar covarf;
  if in1 or in2 or in3 or in4;
  * round percentages;
  percent      = round(percent      , 0.1);
  percent_noia  = round(percent_noia , 0.1);
  percent_t1d   = round(percent_t1d  , 0.1);
  percent_not1d = round(percent_not1d, 0.1);
run;


* Table 1;
data table1;
  set allnpercent alluniv;
run;

proc sort  data=table1;
  by rownum covarf;
run;

proc print data=table1;
  var rownum covar covarf _var_ count percent  _median_ _q1_ _q3_ count_noia   percent_noia noia_median_  noia_q1_  noia_q3_
                          count_t1d    percent_t1d t1d_median_   t1d_q1_   t1d_q3_  count_not1d  percent_not1d not1d_median_
not1d_q1_ not1d_q3_ ;
title3 'Table 1';
run;
```