Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) M62 LJacobsen

Prepared by Sabrina Chen
IMS Inc.
3901 Calverton Blvd, Suite 200 Calverton, MD 20705
Feb 2, 2021

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_62_LJacobsen_NIDDK_Submission folder in the data package. For this replication, variables were taken from the “m_62_ljacobsen_niddk_31aug2016.sas7bdat” dataset.

4 Statistical Methods

Analyses were performed to duplicate results for the data published by Laura Jacobsen et al [1] in Pediatric Diabetes 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] Laura M. Jacobsen, Helena E. Larsson, Roy N. Tamura, Kendra Vehik, Joanna Clasen, Jay Sosenko, William A. Hagopian, Jin-Xiong She, Andrea K. Steck, Marian Rewers, Olli Simell, Jorma Toppari, Riitta Veijola, Anette G. Ziegler, Jeffrey P. Krischer, Beena Akolkar, Michael J. Haller, the TEDDY Study Group. Predicting progression to type 1 diabetes from ages 3 to 6 in islet autoantibody positive TEDDY children. *Pediatr Diabetes*. 2019;1–8. https://doi.org/10.1111/pedi.12812.

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

Table Variable	**dataset.variable**
Gender	m_62_ljacobsen_niddk_31aug2016.sex
First Degree Relative	m_62_ljacobsen_niddk_31aug2016.fdr_
Country	m_62_ljacobsen_niddk_31aug2016.country
HLA	m_62_ljacobsen_niddk_31aug2016.hla_category
SNP rs12708716_G number of minor alleles	m_62_ljacobsen_niddk_31aug2016.rs12708716_g
Indicates whether or not the subject had a persistent confirmed GAD antibody by age 3 – 0=no, 1=yes (persistent confirmed is defined as the antibody being confirmed positive on 2 consecutive visits)	m_62_ljacobsen_niddk_31aug2016.age3_gad
Indicates whether or not the subject had a persistent confirmed IA-2A antibody by age 3 – 0=no, 1=yes (persistent confirmed is defined as the antibody being confirmed positive on 2 consecutive visits)	m_62_ljacobsen_niddk_31aug2016.age3_ia2a
Indicates whether or not the subject had a persistent confirmed mIAA antibody by age 3 – 0=no, 1=yes (persistent confirmed is defined as the antibody being confirmed positive on 2 consecutive visits)	m_62_ljacobsen_niddk_31aug2016.age3_miaa
Indicates if subject had developed diabetes (T1D) by age 3 – 0=no, 1=yes	m_62_ljacobsen_niddk_31aug2016.age3_t1d
Age 6 T1D	m_62_ljacobsen_niddk_31aug2016.t1d_age_6

Table B-1: Comparison of values computed in integrity check to reference article Table 1 values. Complete Cohort (N, %)

		Manuscript	DSIC	Diff	Manuscript	DSIC	Diff
Variable	**Category**	**Number**			**%**		
Age 6 T1D	Yes	76	76	0	21	21	0
	No	276	0	276	76	0	76
	Missing	11	287	-276	3	79	-76
Country	United States	124	124	0	34	34	0
	Finland	83	83	0	23	23	0
	Germany	25	25	0	7	7	0
	Sweden	131	131	0	36	36	0
Gender	Female	163	163	0	45	45	0
	Male	200	200	0	55	55	0
First-degree relative	No	290	290	0	80	80	0
	Yes	73	73	0	20	20	0
HLA genotype	DR3/DR4	177	177	0	49	49	0
	DR4/DR4	72	72	0	20	20	0
	DR4/DR8	56	56	0	15	15	0
	DR3/DR3	43	43	0	12	12	0
	Other	15	15	0	4	4	0
Number of autoantibodies	1	168	168	0	46	46	0
	1+	195	195	0	54	54	0
IA-2A Status	Positive	134	134	0	37	37	0
	Negative	229	229	0	63	63	0
SNP rs12708716_G number of minor alleles	0	167	167	0	46	46	0
	1	162	162	0	45	45	0
	2	32	32	0	9	9	0

Table B-2: Comparison of values computed in integrity check to reference article Table 1 values. Complete Cohort (Mean, SD, Range)

	Manuscript	DSIC	Diff	Manuscript	DSIC	Diff	Manuscript	DSIC	Diff	Manuscript	DSIC	Diff
Variable	**Mean**			**SD**			**Min**			**Max**		
Age in years at persistent autoantibody confirmation	1.86	1.86	0	0.87	0.87	0	0.25	0.25	0	3.5	3.5	0
HbA1c	5.14	5.12	0.02	0.28	0.27	0.01	4.4	4.4	0	6	5.8	0.2
BMI (Z-score)	0.26	0.26	0	0.99	0.99	0	-3.35	-3.35	0	2.33	3.1	-0.77

Table B-3: Comparison of values computed in integrity check to reference article Table 1 values. T1D subjects: Age 6 Cohort (N, %)

		Manuscript	DSIC	Diff	Manuscript	DSIC	Diff
Variable	**Category**	**Number**			**%**		
Age 6 T1D	Yes	76	76	0	100	100	0
	No	0	0	0	0	0	0
	Missing	0	0	0	0	0	0
Country	United States	20	20	0	26	26	0
	Finland	24	24	0	32	32	0
	Germany	6	6	0	8	8	0
	Sweden	26	26	0	34	34	0
Gender	Female	40	40	0	53	53	0
	Male	36	36	0	47	47	0
First-degree relative	No	62	62	0	82	82	0
	Yes	14	14	0	18	18	0
HLA genotype	DR3/DR4	43	43	0	57	57	0
	DR4/DR4	12	12	0	16	16	0
	DR4/DR8	10	10	0	13	13	0
	DR3/DR3	7	7	0	9	9	0
	Other	4	4	0	5	5	0
Number of autoantibodies	1	8	8	0	11	11	0
	1+	68	68	0	89	89	0
IA-2A Status	Positive	57	57	0	75	75	0
	Negative	19	19	0	25	25	0
SNP rs12708716_G number of minor alleles	0	31	31	0	41	41	0
	1	35	35	0	46	46	0
	2	11	10	1	13	13	0

Table B-4: Comparison of values computed in integrity check to reference article Table 1 values. T1D subjects: Age 6 Cohort (Mean, SD, Range)

	Manuscript	DSIC	Diff	Manuscript	DSIC	Diff	Manuscript	DSIC	Diff	Manuscript	DSIC	Diff
Variable	**Mean**			**SD**			**Min**			**Max**		
Age in years at persistent autoantibody confirmation	1.53	1.53	0	0.72	0.72	0	0.36	0.36	0	3.3	3.3	0
HbA1c	5.31	5.3	0.01	0.27	0.26	0.01	4.7	4.7	0	5.8	5.8	0
BMI (Z-score)	0.42	0.42	0	1.06	1.06	0	-2.45	-2.45	0	1.78	3.1	-1.32

Attachment A: SAS Code

```
options nocenter validvarname=upcase fmtsearch=(formats) nofmterr;

title '/prj/niddk/ims_analysis/TEDDY/prog_initial_analysis/m_62_dsic.sas';
run;

* Peds primary outcome.pdf ;

**********;
* INPUT   ;
**********;
libname orig   '/prj/niddk/ims_analysis/TEDDY/private_orig_data/M_62_LJacobsen_NIDDK_Submission/';


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


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

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

  data univ&subsetname;
    length covarf $100 _var_ $25;
    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 countryf
 1 = 'US'
 2 = 'Finland'
 3 = 'Germany'
 4 = 'Sweden'
 ;

value fdr
0= 'GEN POP (also includes unknown)'
1= 'FDR'
;

value sexnumf
```

```
1='Female'
2='Male'
;

value oneplus
1 = "1"
2 = ">1"
;

value posneg
0 = "neg"
1 = "pos"
;

  value hlagenof
-1='HLA*Results*Pending'
 0='Not*Eligible'
 1='DR4*030X/0302*DR3*0501/0201'
 2='DR4*030X/0302*DR4*030X/0302'
 4='DR4*030X/0302*DR8*0401/0402'
 9='DR3*0501/0201*DR3*0501/0201'
 3,5,6,7,8,10 = 'Other'
/* 3='DR4*030X/0302*DR4*030X/020X'
 5='DR4*030X/0302*DR1*0101/0501'
 6='DR4*030X/0302*DR13*0102/0604'
 7='DR4*030X/0302*DR4*030X/0304'
 8='DR4*030X/0302*DR9*030X/0303'
10='DR3*0501/0201*DR9*030X/0303'*/
99='Results*Under*Review'
;

 value hlagpf
-1='HLA*Results*Pending'
 0='Not*Eligible'
 1='DR4*030X/0302*DR3*0501/0201'
 2='DR4*030X/0302*DR4*030X/0302'
 4='DR4*030X/0302*DR8*0401/0402'
 9='DR3*0501/0201*DR3*0501/0201'
 999 = 'Other'
/* 3='DR4*030X/0302*DR4*030X/020X'
 5='DR4*030X/0302*DR1*0101/0501'
 6='DR4*030X/0302*DR13*0102/0604'
 7='DR4*030X/0302*DR4*030X/0304'
 8='DR4*030X/0302*DR9*030X/0303'
10='DR3*0501/0201*DR9*030X/0303'*/
99='Results*Under*Review'
;

value yesno
```

```
0 = 'no'
1 = 'yes'
;

value minorf
 0 = "0 minor alleles"
 1 = "1 minor allele"
 2 = "2 minor alleles"
 ;
run;


%readin(orig, m_62_ljacobsen_niddk_31aug2016 );


data analy;
  set m_62_ljacobsen_niddk_31aug2016;

  in_analy=1;

  if sex = 'Female' then sexnum=1;
  else if sex= 'Male' then sexnum=2;

  * number of autoantibodies;
  if sum(AGE3_GAD,    AGE3_IA2A,    AGE3_MIAA,    AGE3_T1D) = 1 then num_autoab=1;
  else if sum(AGE3_GAD,    AGE3_IA2A,    AGE3_MIAA,    AGE3_T1D) > 1 then num_autoab=2;

  * group HLA;
  if hla_category in(3,5,6,7,8,10) then hla_category_gp = 999;
  else hla_category_gp = hla_category;

  * convert to years;
  persist_draw_yrs = persist_draw_dys/365.24;

run;

proc freq data=analy;
  tables T1D_AGE_6 sex*sexnum/list missing;
  tables hla_category_gp*hla_category/list missing;
title3 "checking";
run;

proc freq data=analy;
  tables sex FDR_ COUNTRY HLA_CATEGORY   AGE3_IA2A   RS12708716_G  /missing;
  tables num_autoab*AGE3_GAD  *         AGE3_IA2A *           AGE3_MIAA  *   AGE3_T1D    /list missing;
  format hla_category hlagenof.  FDR_ fdr. country countryf.;
title3 "Table 1.";
run;
```

```
proc means data=analy;
  var HBA1C BMIZ;
run;



* Cohort: All;

* med, q1, q3;
data prntunivall;
 * length _VAR_ $100;
  set _null_;
run;

%univ(9    , analy, HBA1C          , in_analy , all);
%univ(10   , analy, BMIZ           , in_analy , all);
%univ(11   , analy, persist_draw_yrs , in_analy , all);

data prntunivall;
  set prntunivall;
   _median_  = round(_median_ , 0.1);
   _q1_      = round(_q1_     , 0.1);
   _q3_      = round(_q3_     , 0.1);
   _mean_    = round(_mean_   , 0.01);
   _std_     = round(_std_    , 0.01);
   _min_     = round(_min_    , 0.01);
   _max_     = round(_max_    , 0.01);
run;

proc print data=prntunivall;
 var rownum _var_ covarf _nobs_ /*_median_ _q1_ _q3_ */  _mean_ _std_ _min_ _max_;
 title3 "Complete cohort";
run;


data prntall;
 * length _VAR_ $100;
  set _null_;
run;

%npercent(1, T1D_AGE_6     , yesno     , in_analy , all);
%npercent(2, country       , countryf  , in_analy , all);
%npercent(3, sexnum        , sexnumf   , in_analy , all);
%npercent(4, fdr_          , fdr       , in_analy , all);
%npercent(5, hla_category_gp  , hlagpf    , in_analy , all);
%npercent(6, num_autoab    , oneplus   , in_analy , all);
%npercent(7, age3_ia2a     , posneg    , in_analy , all);
%npercent(8, rs12708716_g  , minorf       , in_analy , all);
```

```
data prntall;
  set prntall;
  percent = round(percent);
run;

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

proc print data=prntall;
  var rownum covar covarf count percent;
 title3 "Complete cohort";
run;
```

```
* Cohort: T1D;

* med, q1, q3;
data prntunivT1D;
 * length _VAR_ $100;
  set _null_;
run;

%univ(9    , analy, HBA1C          , T1D_AGE_6 , T1D);
%univ(10   , analy, BMIZ           , T1D_AGE_6 , T1D);
%univ(11   , analy, persist_draw_yrs , T1D_AGE_6 , T1D);

data prntunivT1D;
  set prntunivT1D;
   _median_  = round(_median_ , 0.1);
   _q1_      = round(_q1_     , 0.1);
   _q3_      = round(_q3_     , 0.1);
   _mean_    = round(_mean_   , 0.01);
   _std_     = round(_std_    , 0.01);
   _min_     = round(_min_    , 0.01);
   _max_     = round(_max_    , 0.01);
run;

proc print data=prntunivT1D;
 var rownum _var_ covarf _nobs_ /*_median_ _q1_ _q3_ */  _mean_ _std_ _min_ _max_;
 title3 "T1D cohort";
run;
```

```
data prntT1D;
 * length _VAR_ $100;
  set _null_;
run;

%npercent(1, T1D_AGE_6     , yesno     , T1D_AGE_6 , T1D);
%npercent(2, country       , countryf  , T1D_AGE_6 , T1D);
%npercent(3, sexnum        , sexnumf   , T1D_AGE_6 , T1D);
%npercent(4, fdr_          , fdr       , T1D_AGE_6 , T1D);
%npercent(5, hla_category_gp  , hlagpf    , T1D_AGE_6 , T1D);
%npercent(6, num_autoab    , oneplus   , T1D_AGE_6 , T1D);
%npercent(7, age3_ia2a     , posneg    , T1D_AGE_6 , T1D);
%npercent(8, rs12708716_g  , minorf        , T1D_AGE_6 , T1D);

data prntT1D;
  set prntT1D;
  percent = round(percent);
run;

proc sort data=prntT1D;
  by rownum covarf;
  run;

proc print data=prntT1D;
  var rownum covar covarf count percent;
 title3 "T1D cohort";
run;
```