Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) M165b Beyerlein

Prepared by Sabrina Chen
IMS Inc.
3901 Calverton Blvd, Suite 200 Calverton, MD 20705
May 22, 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/private_orig_data/M_165b_ABeyerlein_NIDDK_Submission folder in the data package. For this replication, variables were taken from the “m_165b_abeyerlei_niddk_31may2016.sas7bdat” dataset.

4 Statistical Methods

Analyses were performed to duplicate results for the data published by Andreas Beyerlein et al [1] in Journal of Medical Genetics in 2018. 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 NIDDK repository is confident that the TEDDY data files to be distributed are a true copy of the study data.

7 References

[1] Andreas Beyerlein, Ezio Bonifacio, Kendra Vehik, Markus Hippich, Christiane Winkler, Brigitte I Frohnert, Andrea K Steck, William A Hagopian, Jeffrey P Krischer, Åke Lernmark, Marian J Rewers, Jin-Xiong She, Jorma Toppari, Beena Akolkar, Stephen S Rich, Anette-G Ziegler. Progression from islet autoimmunity to clinical type 1 diabetes is influenced by genetic factors: results from the prospective TEDDY study. Journal of Medical Genetics Published Online First: 04 October 2018. doi: 10.1136/jmedgenet-2018-105532.

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

Table Variable	**dataset.variable**
Gender	m_165b_abeyerlei_niddk_31may2016.gender
Haplotype	m_165b_abeyerlei_niddk_31may2016.haplotype
One or more autoantibodies	m_165b_abeyerlei_niddk_31may2016.any_ab
Merged TEDDY score	m_165b_abeyerlei_niddk_31may2016.rsmerged
Indicates if subject has developed T1D	m_165b_abeyerlei_niddk_31may2016.t1d
Indicates the subject is persistent confirmed for multiple autoantibodies - 1 = Yes, 0 = No	m_165b_abeyerlei_niddk_31may2016.multiple_persist_conf_ab
Age in days on date at which subject was diagnosed with T1D	m_165b_abeyerlei_niddk_31may2016.t1d_agedys
Follow-up to first confirmed autoantibody in years	m_165b_abeyerlei_niddk_31may2016.age_first_ab_y
Follow-up to second confirmed autoantibody in years	m_165b_abeyerlei_niddk_31may2016.age_second_ab_y

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

Variable	Manu script	DSIC	Diff.	Manuscri pt	DSIC	Diff.	Manuscri pt	DSIC	Diff.
Median/Q1/Q3									
Age - children who developed islet autoantibodies	2.7	2.8	-0.1	1.5	1.5	0.0	5	5	0.0
Age - children who developed multi autoantibodies	2.8	2.8	0.0	1.8	1.8	0.0	5.1	5.1	0.0
Age - children who developed clinical T1D	5	5	0.0	3	3	0.0	7.1	7.2	-0.1
	Manu script	DSIC	Diff.	Manuscri pt	DSIC	Diff.	Manuscri pt	DSIC	Diff.
n/percent									
children who developed islet autoantibodies	341	341	0.0						
subset who are female	141	141	0.0	41.3	41.4	-0.1			
subset who had the HLA DR3/DR4-DQ8 genotype	250	250	0.0						
subset who had the HLA DR4-DQ8/DR4-DQ8 genotype	91	91	0.0						
children who developed multi autoantibodies	214	214	0.0	62.8	62.8	0.0			
children who developed clinical T1D	107	107	0.0	31.4	31.4	0.0			
children with both multi autoab + T1D	96	96	0.0	28.2	28.2	0.0			

Attachment A: SAS Code

```
options nocenter validvarname=upcase;

title 'prj/niddk/ims_analysis/TEDDY/prog_initial_analysis/m_165b_dsic.sas';
run;

***********;
* INPUT    ;
***********;

libname sasfile '/prj/niddk/ims_analysis/TEDDY/private_orig_data/M_165b_ABeyerlein_NIDDK_Submission/';


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

  proc contents data=&ds;
  title3 "&ds";
  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 $nochar
  "#N/A"  = "#N/A"
  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"
  ;

run;
```

```
%readin(m_165b_abeyerlei_niddk_31may2016);

proc freq data=m_165b_abeyerlei_niddk_31may2016;
  tables t1d_agedys t1d progression_sd age_at_iaa_status haplotype age_first_ab_y age_second_ab_y/missing;
run;

data analy;
  set m_165b_abeyerlei_niddk_31may2016;

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

  if rsmerged ne "#N/A" then rsmergedn = input(rsmerged, best.);
  if T1d_Agedys      ne "#N/A" then T1d_Age_Y_n      =   input(T1d_Agedys     , best.)/365.25;
  if AGE_FIRST_AB_Y  ne "#N/A" then AGE_FIRST_AB_Y_n  =  input(AGE_FIRST_AB_Y , best.);
  if AGE_SECOND_AB_Y ne "#N/A" then AGE_SECOND_AB_Y_n =  input(AGE_SECOND_AB_Y, best.);

  * subset islet ab;
  if any_AB=1 and rsmergedn > 0 then isletab=1;

  * subset islet ab w/ T1D;
  if any_AB=1 and rsmergedn > 0 and t1d=1 then isletabt1d=1;

  * subset multi ab;
  if MULTIPLE_PERSIST_CONF_AB=1 and rsmergedn >0 then multiab=1;

  * subset multi ab w/ T1D;
  if MULTIPLE_PERSIST_CONF_AB=1 and rsmergedn >0 and t1d=1 then multiabt1d=1;

run;


proc freq data=analy;
  tables isletab* any_ab* rsmergedn
         isletabt1d* any_ab* rsmergedn *t1d
          multiab*multiple_persist_conf_ab* rsmergedn
          multiabt1d*multiple_persist_conf_ab* rsmergedn*t1d/list missing;
  tables isletab*isletabt1d*multiab*multiabt1d/list missing;
  format rsmerged $nochar. rsmergedn novalue.;
  title3 "check vars";
run;

proc freq data=analy;
```

```
  where isletab=1;
  tables gender haplotype/missing;
  tables ISLETAB*MULTIAB/list missing;
  tables ISLETAB*ISLETABT1D/list missing;
  tables ISLETAB*MULTIABT1D/list missing;
  title3 "Subset to 341 children who developed islet autoantibodies";
run;


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

%univ(1   , age_first_ab_y_n  , isletab , isletab);

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

%univ(6   , age_second_ab_y_n , multiab , multiab);

data prntunivisletabt1d;
  length _VAR_ $100;
  set _null_;
run;

%univ(7   , t1d_age_y_n       , isletabt1d , isletabt1d);

data alluniv;
  set prntunivisletab    (in=in1 keep = rownum _var_ covarf _nobs_ _median_ _q1_ _q3_)
      prntunivmultiab    (in=in1 keep = rownum _var_ covarf _nobs_ _median_ _q1_ _q3_)
      prntunivisletabt1d (in=in1 keep = rownum _var_ covarf _nobs_ _median_ _q1_ _q3_)
  ;
  _median_   = round(_median_  , 0.1);
  _q1_       = round(_q1_      , 0.1);
  _q3_      = round(_q3_      , 0.1);
run;

proc print data= alluniv noobs;
    var rownum _var_  covarf _nobs_ _median_ _q1_ _q3_ /*_min_ _max_ _std_*/;
    title3 "median, q1, q3 for each subset";
run;
```