Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) M165a Bonifacio

Prepared by Sabrina Chen
IMS Inc.
3901 Calverton Blvd, Suite 200 Calverton, MD 20705
May 8, 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_165a_abeyerlei_niddk_31may2016.sas7bdat" dataset.

4 Statistical Methods

Analyses were performed to duplicate results for the data published by Ezio Bonifacio et al [1] in PLoS Med 15(4):e1002548. To verify the integrity of the dataset, descriptive statistics were computed.

5 Results

For Table S2 in the publication [1], **Frequencies of risk alleles in TEDDY children with the HLA DR3/DR4-DQ8 or DR4-DQ8/DR4-DQ8 genotype**, 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 an exact match to the published results.

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] Ezio Bonifacio, Andreas Beyerlein, Markus Hippich, Christiane Winkler, Kendra Vehik, Michael N. Weedon, Michael Laimighofer, Andrew T. Hattersley, Jan Krumsiek, Brigitte I. Frohnert, Andrea K. Steck, William A. Hagopian, Jeffrey P. Krischer, Åke Lernmark, Marian J. Rewers, Jin-Xiong She, Jorma Toppari, Beena Akolkar, Richard A. Oram, Stephen S. Rich, Anette-G. Ziegler, for the TEDDY Study Group. Genetic scores to stratify risk of developing multiple islet autoantibodies and type 1 diabetes: A prospective study in children. PLoS Med 15(4):e1002548.

Table A: **Variables used to replicate Table 1**: **Frequencies of risk alleles in TEDDY children with the HLA DR3/DR4-DQ8 or DR4-DQ8/DR4-DQ8 genotype.**

Table Variable	dataset.variable
continent	m_165a_abeyerlei_niddk_31may2016.continent
RS3087243_A	m_165a_abeyerlei_niddk_31may2016.RS3087243_A
RS2476601_A	m_165a_abeyerlei_niddk_31may2016.RS2476601_A
rs4788084_A	m_165a_abeyerlei_niddk_31may2016.rs4788084_A
rs2069763_A	m_165a_abeyerlei_niddk_31may2016.rs2069763_A
rs3757247_A	m_165a_abeyerlei_niddk_31may2016.rs3757247_A
rs1738074_A	m_165a_abeyerlei_niddk_31may2016.rs1738074_A
rs45450798_C	m_165a_abeyerlei_niddk_31may2016.rs45450798_C
rs9388489_G	m_165a_abeyerlei_niddk_31may2016.rs9388489_G
rs2292239_A	m_165a_abeyerlei_niddk_31may2016.rs2292239_A
rs7804356_G	m_165a_abeyerlei_niddk_31may2016.rs7804356_G
rs3184504_A	m_165a_abeyerlei_niddk_31may2016.rs3184504_A
rs2664170_G	m_165a_abeyerlei_niddk_31may2016.rs2664170_G
rs5753037_A	m_165a_abeyerlei_niddk_31may2016.rs5753037_A
rs3788013_A	m_165a_abeyerlei_niddk_31may2016.rs3788013_A
rs1990760_G	m_165a_abeyerlei_niddk_31may2016.rs1990760_G
rs6897932_A	m_165a_abeyerlei_niddk_31may2016.rs6897932_A
rs6920220_A	m_165a_abeyerlei_niddk_31may2016.rs6920220_A
rs1465788_A	m_165a_abeyerlei_niddk_31may2016.rs1465788_A
rs2816316_C	m_165a_abeyerlei_niddk_31may2016.rs2816316_C
rs229541_A	m_165a_abeyerlei_niddk_31may2016.rs229541_A
rs2395029_C	m_165a_abeyerlei_niddk_31may2016.rs2395029_C

Table Variable	dataset.variable
rs7020673_G	m_165a_abeyerlei_niddk_31may2016.rs7020673_G
rs7202877_C	m_165a_abeyerlei_niddk_31may2016.rs7202877_C
rs10509540_G	m_165a_abeyerlei_niddk_31may2016.rs10509540_G

Table B: Comparison of values computed in integrity check to reference article Table S2 values

Table S2						
SNP	**Europe**			**USA**		
	Manuscript	**DSIC**	**Diff**	**Manuscript**	**DSIC**	**Diff**
rs3087243	37.28	37.28	0	45.28	45.28	0
rs2476601	12.11	12.11	0	8.8	8.8	0
rs4788084	45.77	45.77	0	40.52	40.52	0
rs2069763	40	40	0	35.5	35.5	0
rs3757247	41.08	41.08	0	45.73	45.73	0
rs1738074	40.59	40.59	0	44.98	44.98	0
rs45450798	17.17	17.17	0	16.12	16.12	0
rs9388489	44.65	44.65	0	48.77	48.77	0
rs2292239	33.19	33.19	0	29.45	29.45	0
rs7804356	21.64	21.64	0	25.11	25.11	0
rs3184504	45.92	45.92	0	42.72	42.72	0
rs2664170	30.28	30.28	0	33.62	33.62	0
rs5753037	34.93	34.93	0	37.96	37.96	0
rs3788013	39.81	39.81	0	41.26	41.26	0

rs1990760	39.98	39.98	0	43.04	43.04	0
rs6897932	29.3	29.3	0	26.83	26.83	0
rs6920220	21.38	21.38	0	19.13	19.13	0
rs1465788	29.39	29.39	0	27.22	27.22	0
rs2816316	17.65	17.65	0	19.09	19.09	0
rs229541	40.47	40.47	0	42.35	42.35	0
rs2395029	0.96	0.96	0	1.46	1.46	0
rs7020673	49.15	49.15	0	48.35	48.35	0
rs7202877	11.74	11.74	0	10.45	10.45	0
rs10509540	27.54	27.54	0	25.7	25.7	0

Attachment A: SAS Code

```
options nocenter validvarname=upcase;

title 'prj/niddk/ims_analysis/TEDDY/prog_initial_analysis/m_165a_dsic.sas';
run;

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

libname sasfile '/prj/niddk/ims_analysis/TEDDY/private_orig_data/M_165a_ABeyerlein_NIDDK_Submission/';


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

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


%macro maf(rownum, snp, strat);

proc freq data = analy;
  where &snp in("0","1","2") and continent="&strat";
  tables &snp/out=freqout noprint;
run;

data freqout;
  set freqout;
  subject = 1;
  if &snp = "0" then VAR=1;
  else if &snp = "1" then VAR=2;
  else if &snp = "2" then VAR=3;
run;

*Nucleotide_change;
```

```
data oneobs_temp;
  set freqout;
  by subject;
  array xx[*] AA AB BB;
  retain AA AB BB;
  if first.subject then do i = 1 to 3;
   xx[i] = .;
  end;
 xx[VAR] = count;
 if last.subject then output;
run;

data oneobs_temp(drop = var    COUNT       subject i);
  length var_label $60;
  set oneobs_temp;
  var_label = vlabel(&snp);

  rownum = input("&rownum", best.);

  if AA = . then AA = 0;
  if AB = . then AB = 0;
  if BB = . then BB = 0;

  sum=AA+AB+BB;

  * calculate the MAF (minor allele frequency);
  * Minor allele frequency (MAF) = (1*frequency of the heterozygote + 2*frequency of the homozygote variant) / (2*total N for that
SNP)  ;
  maf = (1*AB + 2*BB)/(2*(AA+AB+BB));
run;

data prnt&strat;
  set prnt&strat oneobs_temp;
run;

%mend;



%readin(m_165a_abeyerlei_niddk_31may2016);

proc freq data=m_165a_abeyerlei_niddk_31may2016;
  tables continent/missing;
run;
```

```
data analy;
  set m_165a_abeyerlei_niddk_31may2016;
run;


data prntus;
  set _null_;
run;

%maf(1  ,RS3087243_A  , US);
%maf(2  ,RS2476601_A  , US);
%maf(3  ,rs4788084_A  , US);
%maf(4  ,rs2069763_A  , US);
%maf(5  ,rs3757247_A  , US);
%maf(6  ,rs1738074_A  , US);
%maf(7  ,rs45450798_C , US);
%maf(8  ,rs9388489_G  , US);
%maf(9  ,rs2292239_A  , US);
%maf(10 ,rs7804356_G  , US);
%maf(11 ,rs3184504_A  , US);
%maf(12 ,rs2664170_G  , US);
%maf(13 ,rs5753037_A  , US);
%maf(14 ,rs3788013_A  , US);
%maf(15 ,rs1990760_G  , US);
%maf(16 ,rs6897932_A  , US);
%maf(17 ,rs6920220_A  , US);
%maf(18 ,rs1465788_A  , US);
%maf(19 ,rs2816316_C  , US);
%maf(20 ,rs229541_A   , US);
%maf(21 ,rs2395029_C  , US);
%maf(22 ,rs7020673_G  , US);
%maf(23 ,rs7202877_C  , US);
%maf(24 ,rs10509540_G , US);

proc print data=prntus;
  var rownum var_label maf;
title3 "USA";
run;



data prnteurope;
  set _null_;
run;
```

```
%maf(1  ,RS3087243_A  , Europe);
%maf(2  ,RS2476601_A  , Europe);
%maf(3  ,rs4788084_A  , Europe);
%maf(4  ,rs2069763_A  , Europe);
%maf(5  ,rs3757247_A  , Europe);
%maf(6  ,rs1738074_A  , Europe);
%maf(7  ,rs45450798_C , Europe);
%maf(8  ,rs9388489_G  , Europe);
%maf(9  ,rs2292239_A  , Europe);
%maf(10 ,rs7804356_G  , Europe);
%maf(11 ,rs3184504_A  , Europe);
%maf(12 ,rs2664170_G  , Europe);
%maf(13 ,rs5753037_A  , Europe);
%maf(14 ,rs3788013_A  , Europe);
%maf(15 ,rs1990760_G  , Europe);
%maf(16 ,rs6897932_A  , Europe);
%maf(17 ,rs6920220_A  , Europe);
%maf(18 ,rs1465788_A  , Europe);
%maf(19 ,rs2816316_C  , Europe);
%maf(20 ,rs229541_A   , Europe);
%maf(21 ,rs2395029_C  , Europe);
%maf(22 ,rs7020673_G  , Europe);
%maf(23 ,rs7202877_C  , Europe);
%maf(24 ,rs10509540_G , Europe);

proc print data=prnteurope;
  var rownum var_label maf;
title3 "Europe";
run;


* combine sites;
proc sort data=prntus (rename=(maf=maf_us));
  by rownum;
run;

proc sort data=prnteurope (rename=(maf=maf_europe));
  by rownum;
run;

data table_s2;
  merge prntus    (in=in1 keep=rownum var_label maf_us)
        prnteurope(in=in2 keep=rownum maf_europe);
  by rownum;
  if in1 or in2;
```

```
  maf_us     =maf_us    *100;
  maf_europe =maf_europe*100;

  format maf_us
         maf_europe 6.2;
run;

proc print data=table_s2;
  var rownum var_label maf_europe maf_us;
title3 "Table S2";
run;
```