Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) M237 EBonifacio

Prepared by Sabrina Chen
IMS Inc.
3901 Calverton Blvd, Suite 200 Calverton, MD 20705
Jan 12, 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_237_EBonifacio_NIDDK_Submission folder in the data package. For this replication, variables were taken from the “m_237_ebonifa_niddk_30june2019_1.sas7bdat” and “m_237_ebonifa_niddk_30june2019_2.sas7bdat " datasets.

4 Statistical Methods

Analyses were performed to duplicate results for the data published by Ezio Bonifacio et al [1] in Diabetes Care 2021. 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 an exact match to the published results.

7 References

[1] Bonifacio E, Weiß A, Winkler C, Hippich M, Rewers MJ, Toppari J, Lernmark Å, She JX, Hagopian WA, Krischer JP, Vehik K, Schatz DA, Akolkar B, Ziegler AG; TEDDY Study Group. An Age-Related Exponential Decline in the Risk of Multiple Islet Autoantibody Seroconversion During Childhood. Diabetes Care. 2021 Feb 24:dc202122. doi: 10.2337/dc20-2122. Epub ahead of print. PMID: 33627366.

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

Table Variable	**dataset.variable**
Sex	m_237_ebonifa_niddk_30june2019_1.sex
First-degree Relative	m_237_ebonifa_niddk_30june2019_1.fdr
Country	m_237_ebonifa_niddk_30june2019_1.country
HLA Genotype	m_237_ebonifa_niddk_30june2019_1.hla_category
Genetic Risk Score	m_237_ebonifa_niddk_30june2019_2.rs_mp165

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

		Manuscript	DSIC	Diff	Manuscript	DSIC	Diff
Variable	**Category**	**Number**			**%**		
Sex	Girls	4,226	4,226	0	49.4	49.4	0
	Boys	4,330	4,330	0	50.6	50.6	0
First-degree relative	Yes	955	955	0	11.2	11.2	0
with type 1 diabetes	No	7,601	7,601	0	88.8	88.8	0
Site	Europe	4,895	4,895	0	57.2	57.2	0
	U.S.A.	3,661	3,661	0	42.8	42.8	0
HLA genotype	DR3/DR4-DQ8	3,339	3,339	0	39	39	0
	DR4-DQ8/DR4-DQ8	1,674	1,674	0	19.6	19.6	0
	DR4-DQ8/DR8	1,474	1,474	0	17.2	17.2	0
	DR3/DR3	1,791	1,791	0	20.9	20.9	0
	Other*	278	278	0	3.3	3.2	0
Genetic risk score**	Highest quartile	1,104	1,103	1	25	24.99	0
(n=4,413)	2nd and 3rd quartiles	2,206	2,207	-1	50	50.01	0
	Lowest quartile	1,103	1,103	0	25	24.99	0

Attachment A: SAS Code

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

title '/prj/niddk/ims_analysis/TEDDY/prog_initial_analysis/m_237_dsic.sas';
run;

* Peds primary outcome.pdf ;

**********;
* INPUT   ;
**********;
libname orig   '/prj/niddk/ims_analysis/TEDDY/private_orig_data/M_237_EBonifacio_NIDDK_Submission/';

/*
libname fmts '/prj/niddk/ims_analysis/LOGIC/private_created_data/';

PROC FORMAT CNTLIN = fmts.algsformats;

  */


************;
* 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=analy&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 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 sitef
1="US"
2,3,4="Europe"
;

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

value quartf
1 = 'Highest quartile'
2 = '2nd and 3rd quartiles'
3 = 'Lowest quartile'
;

value sexnumf
1='Female'
2='Male'
;

run;


%readin(orig, m_237_ebonifa_niddk_30june2019_1 );
%readin(orig, m_237_ebonifa_niddk_30june2019_2 );


proc sort data=m_237_ebonifa_niddk_30june2019_1;
  by MP237_MASKID;
run;

proc sort data=m_237_ebonifa_niddk_30june2019_2;
  by MP237_MASKID;
run;

data analy;
  merge m_237_ebonifa_niddk_30june2019_1 (in=in1 keep=mp237_maskid sex fdr T1D COUNTRY HLA_CATEGORY)
```

```
       m_237_ebonifa_niddk_30june2019_2 (in=in2 keep=mp237_maskid  RS_MP162  RS_MP165);
  by mp237_maskid;
  if in1 then in_1=1;
  if in2 then in_2=1;
run;

proc freq data=analy;
  tables in_1*in_2/list missing;
run;


* find cut points;
proc univariate data=analy outtable=cut1 noprint;
  where hla_category in(1,2);
  var  RS_MP162  ;
run;

proc print data=cut1;
  var  _VAR_   _LABEL_  _NOBS_ _NMISS_        _Q1_ _MEDIAN_   _Q3_;
  title3 "RS_MP162: cut points";
run;

proc univariate data=analy outtable=cut2 noprint;
  where hla_category in(1,2);
  var  RS_MP165;
run;

proc print data=cut2;
  var  _VAR_   _LABEL_  _NOBS_ _NMISS_        _Q1_ _MEDIAN_   _Q3_;
  title3 "RS_MP165: cut points";
run;

data analy (keep=mp237_maskid in_analy sex fdr T1D COUNTRY HLA_CATEGORY  RS_MP162 RS_MP162_gp MP162_Q1_ MP162_MEDIAN_ MP162_Q3_  RS_MP165 RS_MP165_gp
MP165_Q1_ MP165_MEDIAN_ MP165_Q3_ sexnum);
  if _n_ = 1 then do;
    set cut1 (keep= _Q1_ _MEDIAN_   _Q3_ rename=(_Q1_    =MP162_Q1_
                                          _MEDIAN_=MP162_MEDIAN_
                                        _Q3_    =MP162_Q3_    ));
    set cut2 (keep= _Q1_ _MEDIAN_   _Q3_ rename=(_Q1_    =MP165_Q1_
                                          _MEDIAN_=MP165_MEDIAN_
                                        _Q3_    =MP165_Q3_    ));
  end;
  set analy;

  in_analy=1;

  if hla_category in(1,2) then do;

          if MP162_Q3_ < RS_MP162 then RS_MP162_gp = 1;
```

```
          else if MP162_Q1_ <= RS_MP162 <= MP162_Q3_ then RS_MP162_gp = 2;
          else if . < RS_MP162 < MP162_Q1_ then RS_MP162_gp = 3;

          if MP165_Q3_ < RS_MP165 then RS_MP165_gp = 1;
          else if MP165_Q1_ <= RS_MP165 <= MP165_Q3_ then RS_MP165_gp = 2;
          else if . < RS_MP165 < MP165_Q1_ then RS_MP165_gp = 3;
  end;

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

run;

proc freq data=analy;
  tables sex*sexnum/list missing;
*  tables RS_MP162_gp*RS_MP162/list missing;
*  tables RS_MP165_gp*RS_MP165/list missing;
  tables RS_MP162_gp RS_MP165_gp/missing;
title3 "checking";
run;

proc freq data=analy;
  tables sex fdr COUNTRY HLA_CATEGORY/missing;
  format hla_category hlagenof. country sitef. fdr fdr.;
title3 "Table 1. Characteristics of the study population (n = 8,556)";
run;



* Generate Table 1;
data prntall;
 * length _VAR_ $100;
  set _null_;
run;

%npercent(1, Sexnum      , SEXnumF  , in_analy, all);
%npercent(2, fdr         , fdr      , in_analy, all);
%npercent(3, COUNTRY     , sitef    , in_analy, all);
%npercent(4, HLA_CATEGORY, hlagenof , in_analy, all);

data prntall;
  set prntall;
  percent = round(percent,0.1);
run;

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

```
proc print data=prntall;
  var rownum covar covarf count percent;
  title3 "Table 1. Characteristics of the study population (n = 8,556)";
run;

proc freq data=analy;
  where hla_category in(1,2);
  tables RS_MP165_gp/missprint;
  format RS_MP165_gp quartf.;
title3 "Table 1. Characteristics of the study population (n = 8,556)";
run;
```