Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) M177 Salami

Prepared by Sabrina Chen
IMS Inc.
3901 Calverton Blvd, Suite 200 Calverton, MD 20705
May 14, 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_177_fsalami_niddk_31jan2017_1.sas7bdat" dataset.

4 Statistical Methods

Analyses were performed to duplicate results for the data published by Falastin Salami et al [1] in *Diabetes.* Nov;67(11):2329-2336. To verify the integrity of the dataset, descriptive statistics were computed.

5 Results

For Table 1 in the publication [1], **Characteristics of TEDDY children (n=448) investigated for Complete Blood Counts (CBC) when negative or positive for one or several islet autoantibodies (IA)**, 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 almost 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] Falastin Salami1, Hye-Seung Lee2, Eva Freyhult3, Helena Elding Larsson1, Ake Lernmark1, Carina Torn1 and the TEDDY study group. Reduction in White Blood Cell, Neutrophil, and Red Blood Cell Counts Related to Sex, HLA, and Islet Autoantibodies in Swedish TEDDY Children at Increased Risk for Type 1 Diabetes. *Diabetes.* Nov;67(11):2329-2336.

Table A: **Variables used to replicate Table 1**: **Characteristics of TEDDY children (n=448) investigated for Complete Blood Counts (CBC) when negative or positive for one or several islet autoantibodies (IA).**

Table Variable	**dataset.variable**
Age at first CBC	m_177_fsalami_niddk_31jan2017_1.ageatfirstcbc
Months of Follow up	m_177_fsalami_niddk_31jan2017_1.mcbcfutime
Sex	m_177_fsalami_niddk_31jan2017_1.sex
IA positive/negative	m_177_fsalami_niddk_31jan2017_1.persist_conf_ab
Number of IA	m_177_fsalami_niddk_31jan2017_1.numia
HLA DR-DQ	m_177_fsalami_niddk_31jan2017_1.hlarg

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

Table 1	Manuscript	DSIC	Diff	Manuscript	DSIC	Diff
	IA Negative (n)	**IA Negative (percent)**				
Number of children	376			376		
Number of IA						
1	0	0	0			
2	0	0	0			
3	0	0	0			
HLA DR-DQ (n (%))						
DR3/4 -DQ2/8	151	151	0	40.2	40.2	0
DR4/4-DQ 8/8	87	87	0	23.1	23.1	0
DR4/8-DQ 8/4	40	40	0	10.6	10.6	0
DR3/3-DQ 2/2	91	91	0	24.2	24.2	0
DR4/1-DQ 8/5	4	4	0	1.1	1.1	0
DR4/13-DQ 8/6	2	2	0	0.5	0.5	0
HLA ineligible	1	1	0	0.3	0.3	0
Total IN PRESS	376	376	0	100	100.0	0

Table 1	IA Positive (n)			IA Positive (percent)		
Number of children	72	72	0	16	16	0
Number of IA						
1	25	25	0			
2	16	16	0			
3	31	31	0			
HLA DR-DQ (n (%))						
DR3/4 -DQ2/8	39	39	0	54.1	54.2	-0.1
DR4/4-DQ 8/8	12	12	0	16.7	16.7	0
DR4/8-DQ 8/4	12	12	0	16.7	16.7	0
DR3/3-DQ 2/2	9	9	0	12.5	12.5	0
DR4/1-DQ 8/5	0	0	0			
DR4/13-DQ 8/6	0	0	0			
HLA ineligible	0	0	0			
Total IN PRESS	72	72	0	100	100	0

Table 1	Manuscript	DSIC	Diff	Manuscript	DSIC	Diff
	IA Negative			**IA Positive**		
Girls/Boys	182/194	182/194	0/0	30/42	30/42	0/0
Age at first CBC (months):median(min-max)	91 (52-145)	91 (52-145)	0(0-0)	101.5 (59-139)	101.5 (59-139)	0(0-0)
Girls/Boys	91 (53-144)/94 (52-145)	91 (53-144)/94 (52-145)		103 (59-139)/99 (59-137)	102.5 (59-139)/99 (59-137)	0.5(0-0)/0(0-0)
CBC measures per child (min-max)	1-6	1-6	0-0	1-9	1-9	0-0
Months of follow up (min-max)	1-30	0-31	1,-1	1-30	0-31	1,-1

Attac hment A: SAS Code

```
options nocenter validvarname=upcase;

title 'prj/niddk/ims_analysis/TEDDY/prog_initial_analysis/m_177_dsic.v2.sas';
run;

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

libname sasfile '/prj/niddk/ims_analysis/TEDDY/private_orig_data/M_177_FSalami_NIDDK_Submission(1)/';


************;
* MACROS    ;
************;
%macro readin(ds);
  data &ds;
    set sasfile.&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 hladrdqf
  . = "HLA ineligible"
  1 = "1 = DR3/4 -DQ2/8"
  2 = "2 = DR4/4-DQ 8/8"
  3 = "3 = DR4/8-DQ 8/4"
  4 = "4 = DR3/3-DQ 2/2"
  5 = "5 = DR4/1-DQ 8/5"
  6 = "6 = DR4/1-DQ 8/5"
  7 = "7 = DR4/13-DQ 8/6"
  ;

  value zerof
  . = "0"
  ;

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

run;

%readin(m_177_fsalami_niddk_31jan2017_3);
%readin(m_177_fsalami_niddk_31jan2017_2);


* this file contains the 448 subset;
%readin(m_177_fsalami_niddk_31jan2017_1);

proc freq data=m_177_fsalami_niddk_31jan2017_1;
  tables persist_conf_ab
         ageatfirstcbc
         hlarg
         numia
         sex
         mcbcfutime
         /missing;
title3 "m_177_fsalami_niddk_31jan2017_1";
run;


proc sort data=m_177_fsalami_niddk_31jan2017_1 out=analy;
  by persist_conf_ab sex;
run;
```

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

  if persist_conf_ab = 0 then persist_conf_ab_neg=1;
  else if persist_conf_ab = 1 then persist_conf_ab_pos=1;

  if persist_conf_ab=0 then do;
    if upcase(sex) = "FEMALE" then ia_neg_female=1;
    else if upcase(sex) = "MALE" then ia_neg_male=1;
    else if sex ne "" then abort;
  end;

  if persist_conf_ab=1 then do;
    if upcase(sex) = "FEMALE" then ia_pos_female=1;
    else if upcase(sex) = "MALE" then ia_pos_male=1;
    else if sex ne "" then abort;
  end;

  if upcase(sex) = "FEMALE" then sexn=2;
  else if upcase(sex) = "MALE" then sexn=1;

run;

proc freq data=analy;
  tables persist_conf_ab*persist_conf_ab_neg*persist_conf_ab_pos/list missing;
  tables persist_conf_ab*sex*ia_neg_female*ia_neg_male*ia_pos_female*ia_pos_male/list missing;
  tables sex*sexn/list missing;
  tables nr/missing;
run;


* med, min and max;
data prntunivneg;
  set _null_;
run;

%univ(3  , ageatfirstcbc, persist_conf_ab_neg       , neg);
%univ(4.1, ageatfirstcbc, ia_neg_female      , neg);
%univ(4.2, ageatfirstcbc, ia_neg_male        , neg);
%univ(5  , nr          , persist_conf_ab_neg  , neg);
%univ(6  , mcbcfutime  , persist_conf_ab_neg  , neg);

proc print data= prntunivneg noobs;
```

```
    var rownum _var_  covarf _nobs_ _median_ _min_ _max_ _std_;
run;


data prntunivpos;
  set _null_;
run;

%univ(3  , ageatfirstcbc, persist_conf_ab_pos       , pos);
%univ(4.1, ageatfirstcbc, ia_pos_female      , pos);
%univ(4.2, ageatfirstcbc, ia_pos_male        , pos);
%univ(5  , nr          , persist_conf_ab_pos  , pos);
%univ(6  , mcbcfutime  , persist_conf_ab_pos  , pos);

proc print data= prntunivpos noobs;
    var rownum _var_  covarf _nobs_ _median_ _min_ _max_ _std_;
run;


* combine rows;
proc sort data=prntunivneg;
  by rownum covarf;
run;

proc sort data=prntunivpos (rename=(_median_ = pos_median_
                                    _min_    = pos_min_
                                  _max_    = pos_max_ ))
                                  ;
  by rownum covarf;
run;

data alluniv;
  merge prntunivneg (in=in1 keep = rownum _var_ covarf _median_ _min_ _max_)
        prntunivpos (in=in2 keep = rownum _var_ covarf pos_median_ pos_min_ pos_max_);
  by rownum;
  if in1 or in2;
run;


* n and percent;
data prntneg;
  set _null_;
run;

%npercent(1, persist_conf_ab  , negpos       , all              , neg);
```

```
%npercent(2, sexn             , sexf         , persist_conf_ab_neg, neg);
%npercent(7, numia            , zerof        , persist_conf_ab_neg, neg);
%npercent(8, hlarg            , hladrdqf     , persist_conf_ab_neg, neg);
%npercent(9, persist_conf_ab  , negpos        , persist_conf_ab_neg, neg);

proc print data=prntneg;
  var rownum covarf COUNT    PERCENT;
title3 "Negative";
run;


data prntpos;
  set _null_;
run;

%npercent(1, persist_conf_ab  , negpos       , all               , pos);
%npercent(2, sexn             , sexf         , persist_conf_ab_pos, pos);
%npercent(7, numia            , zerof        , persist_conf_ab_pos, pos);
%npercent(8, hlarg            , hladrdqf     , persist_conf_ab_pos, pos);
%npercent(9, persist_conf_ab  , negpos        , persist_conf_ab_pos, pos);

proc print data=prntpos;
  var rownum covarf COUNT    PERCENT;
title3 "Positive";
run;


* combine rows;
proc sort data=prntneg;
  where covarf ne "Positive";
  by rownum covarf;
run;

proc sort data=prntpos (rename=(count   = pos_count
                                percent = pos_percent));
  where covarf ne "Negative";
  by rownum covarf;
run;

data allnprcnt;
  merge prntneg (in=in1 keep=rownum covarf count percent)
        prntpos (in=in2 keep=rownum covarf pos_count pos_percent);
  by rownum covarf;
  if in1 or in2;
run;
```

```
* set all rows together and format;
data table;
  set allnprcnt (rename=(PERCENT    =xPERCENT
                         pos_PERCENT=xpos_PERCENT ))
      alluniv   (rename=(_MEDIAN_   =x_MEDIAN_
                         pos_MEDIAN_=xpos_MEDIAN_
                         _MAX_      =x_MAX_
                         POS_MAX_   =xPOS_MAX_));

  * round to nearest integer;
  _MEDIAN_     = round(x_MEDIAN_);
  pos_MEDIAN_  = round(xpos_MEDIAN_);
  _MAX_        = round(x_MAX_     );
  POS_MAX_     = round(xPOS_MAX_  );

  * round to the nearest tenth;
  if rownum in(8,9) then do;
    PERCENT    = round(xPERCENT, 0.1);
    pos_PERCENT= round(xpos_PERCENT, 0.1);
  end;
  else do;
    PERCENT    = round(xPERCENT);
    pos_PERCENT= round(xpos_PERCENT);
  end;

run;

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

proc print data=table;
  var rownum _var_ covarf count percent _median_ _min_  _max_ pos_count pos_percent pos_median_ pos_min_ pos_max_;
run;
```