Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) M106a Norris

Prepared by Sabrina Chen
IMS Inc.
3901 Calverton Blvd, Suite 200 Calverton, MD 20705
May 2, 2019

Contents

1 Standard Disclaimer 2

2 Study Background 2

3 Archived Datasets 2

4 Statistical Methods 2

5 Results 3

6 Conclusions 3

7 References 3

Table A: Variables used to replicate Table 1: Characteristics of the first 100 the Environmental Determinants of Diabetes in the Young (TEDDY) children diagnosed with T1D 4

Table B: Comparison of values computed in integrity check to reference article Table 1 values 5

Table C: Variables used to replicate Table 2: Symptoms and laboratory data at onset of T1D ..... **Error! Bookmark not defined.**

Table D: Comparison of values computed in integrity check to reference article Table 2 values..... **Error! Bookmark not defined.**

Table E: Variables used to replicate Figure 2: ..... **Error! Bookmark not defined.**

Table F: Comparison of values computed in integrity check to reference article Figure 2 **Error! Bookmark not defined.**

Attachment A: SAS Code 5

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_106a_jnorris_niddk_31may2012_1.sas7bdat" dataset.

4 Statistical Methods

Analyses were performed to duplicate results for the data published by Jill Norris et al [1] in *Diabetes.* 2018 Jan;67(1):146-154. To verify the integrity of the dataset, descriptive statistics were computed.

5 Results

For Table 1 in the publication [1], **Characteristics of subjects in the TEDDY nested case-control study**, 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] Norris JM, Lee HS, Frederiksen B, Erlund I, Uusitalo U, Yang J, Lernmark Å, Simell O, Toppari J, Rewers M, Ziegler AG, She JX, Onengut-Gumuscu S, Chen WM, Rich SS, Sundvall J, Akolkar B, Krischer J, Virtanen SM, Hagopian W; TEDDY Study Group. Plasma 25-Hydroxyvitamin D Concentration and Risk of Islet Autoimmunity. *Diabetes.*2018 Jan;67(1):146-154.

Table A: Variables used to replicate Table 1**:** Characteristics of subjects in the TEDDY nested case-control study**.**

Table Variable	**dataset.variable**
Clinical Center	m_106a_jnorris_niddk_31may2012_1.cc
Sex	m_106a_jnorris_niddk_31may2012_1.sex
FDR/GP Status	m_106a_jnorris_niddk_31may2012_1.fdr
Outcome	m_106a_jnorris_niddk_31may2012_1.outcome

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

Table 1	Manuscript	DSIC	Diff	Manuscript	DSIC	Diff
	Cases (n)			**Cases (percent)**		
Clinical Center						
Colorado	55	55	0	14.6	14.6	0
Georgia	23	23	0	6.1	6.1	0
Washington State	35	35	0	9.3	9.3	0
Finland	112	112	0	29.8	29.8	0
Germany	29	29	0	7.7	7.7	0
Sweden	122	122	0	32.5	32.4	-0.1
Sex						
Female	167	167	0	44.4	44.4	0
Male	209	209	0	55.6	55.6	0
FDR/GP Status						
FDR	91	91	0	24.2	24.2	0
GP	285	285	0	75.8	75.8	0

Attachment A: SAS Code

```
options nocenter validvarname=upcase;

title 'prj/niddk/ims_analysis/TEDDY/prog_initial_analysis/m_106a_dsic.sas';
run;

* DSIC for TEDDY M106a. Reproduce Table 1 of M_106a_JNorris_NIDDK_Manuscript.pdf ;

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

libname sasfile1 '/prj/niddk/ims_analysis/TEDDY/private_orig_data/M_106a_JNorris_NIDDK_Submission/';


************;
* MACROS    ;
************;
%macro readin(lib, ds);
  data &ds;
    set sasfile&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;
  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 yesno
  . = "Missing"
  0 = "No"
  1 = "Yes"
  ;

  value gender
  0 = "Male"
  1 = "Female"
  ;

  value mateduc
  1 = 'Basic Primary Education'
  2 = 'Graduated Trade School or some College/University'
  3 = 'Higher Education - graduated university/college or higher'
  ;

  value fdr
  1='FDR'
  0='GenPop'
  ;

  value site
  1 = 'Colorado'
  2 = 'Georgia'
  3 = 'Washington'
  4 = 'Finland'
  5 = 'Germany'
  6 = 'Sweden'
  ;


run;


%readin(1, m_106a_jnorris_niddk_31may2012_1);
%readin(1, m_106a_jnorris_niddk_31may2012_2);


proc freq data=m_106a_jnorris_niddk_31may2012_1;
  where outcome=1;
  tables CC CASE_IND sex first_vtd/missing;
run;

data analy;
```

```
  set m_106a_jnorris_niddk_31may2012_1;
  subset_all = 1;

  if sex = "Female" then sexnum=1;
  else if sex = "Male" then sexnum=0;

run;


** Table 1;

* n percent;
data prntcase;
  set _null_;
run;

%npercent(1   , cc                    , site    , outcome , case);
%npercent(2   , sexnum                        , gender  , outcome , case);
%npercent(3   , fdr                   , fdr     , outcome , case);

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

proc print data=prntcase;
  var rownum covar covarf count    percent;
run;
```