

Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) M124 Smith

Prepared by Sabrina Chen

IMS Inc.

3901 Calverton Blvd, Suite 200 Calverton, MD 20705

February 28, 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 2: Characteristics of HLA-eligible TEDDY children followed to age 6 years	4
Table B: Comparison of values computed in integrity check to reference article Table 2 values.....	5
Attachment A: SAS Code.....	6

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_124_lsmith_niddk_30jun2015_1.sas7bdat” dataset.

4 Statistical Methods

Analyses were performed to duplicate results for the data published by Smith et al [1] in *Pediatrics* in 2017. To verify the integrity of the dataset, descriptive statistics were computed.

5 Results

For Table 2 in the publication [1], CBCL Scores Reported by Mothers of No-CDA Children, Unaware-CDA Mothers, and Aware-CDA Mothers at 3.5 y of Age, 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 2. The results of the replication are similar to the published results.

6 Conclusions

The NIDDK repository is confident that the TEDDY M124 data files to be distributed are a true copy of the study data.

7 References

[1] Laura B. Smith, PhD, a Kristian F. Lynch, PhD, b Kalle Kurppa, MD, PhD, c Sibylle Koletzko, MD, PhD, d Jeffrey Krischer, PhD, a Edwin Liu, MD, e Suzanne Bennett Johnson, PhD, f Daniel Agardh, MD, PhD, g The TEDDY study group. "Psychological Manifestations of Celiac Disease Autoimmunity in Young Children". PEDIATRICS Volume 139, number 3, March 2017: e20162848.

Table A: Variables used to replicate Table 2: Characteristics of HLA-eligible TEDDY children followed to age 6 years

Table Variable	dataset.variable
Internalizing Composite	m_124_lsmith_niddk_30jun2015_1.INTERNALIZING_RAWSCORE_42MO
Anxious/Depressed	m_124_lsmith_niddk_30jun2015_1.ANXDEP_RAWSCORE_42MO
Emotionally Reactive	m_124_lsmith_niddk_30jun2015_1.EMOTIONAL_RAWSCORE_42MO
Somatic Complaints	m_124_lsmith_niddk_30jun2015_1.SOMATIC_RAWSCORE_42MO
Withdrawn	m_124_lsmith_niddk_30jun2015_1.WITHDRAWN_RAWSCORE_42MO
Externalizing Composite	m_124_lsmith_niddk_30jun2015_1.EXTERNALIZING_RAWSCORE_42MO
Attention Problems	m_124_lsmith_niddk_30jun2015_1.ATTENTION_RAWSCORE_42MO
Aggressive Behavior	m_124_lsmith_niddk_30jun2015_1.AGGRESSIVE_RAWSCORE_42MO
Sleep Problems	m_124_lsmith_niddk_30jun2015_1.SLEEPPROBS_RAWSCORE_42MO

Table B: Comparison of values computed in integrity check to reference article Table 2 values

CBCL Score	No-CDA (Mean)			No-CDA (SD)		
	Manuscript	DSIC	Diff	Manuscript	DSIC	Diff
Internalizing Composite	5.72	5.72	0	5.09	5.09	0
Anxious/Depressed	1.26	1.26	0	1.57	1.58	0.01
Emotionally Reactive	1.86	1.86	0	1.94	1.94	0
Somatic Complaints	1.63	1.63	0	1.78	1.78	0
Withdrawn	0.96	0.96	0	1.32	1.33	0.01
Externalizing Composite	9.54	9.55	0.01	6.91	6.91	0
Attention Problems	1.55	1.55	0	1.59	1.59	0
Aggressive Behavior	7.99	7.99	0	5.80	5.8	0
Sleep Problems	2.59	2.59	0	2.33	2.33	0

CBCL Score	Unaware-CDA (Mean)			Unaware-CDA (SD)		
	Manuscript	DSIC	Diff	Manuscript	DSIC	Diff
Internalizing Composite	6.80	6.8	0	6.75	6.75	0
Anxious/Depressed	1.82	1.82	0	2.58	2.58	0
Emotionally Reactive	2.05	2.05	0	2.41	2.41	0
Somatic Complaints	1.77	1.77	0	1.81	1.81	0
Withdrawn	1.17	1.17	0	1.40	1.4	0
Externalizing Composite	11.21	11.21	0	8.04	8.05	0.01
Attention Problems	1.65	1.65	0	1.70	1.7	0
Aggressive Behavior	9.56	9.56	0	6.67	6.67	0
Sleep Problems	3.32	3.32	0	2.64	2.64	0

CBCL Score	Aware-CDA (Mean)			Aware-CDA (SD)		
	Manuscript	DSIC	Diff	Manuscript	DSIC	Diff
Internalizing Composite	5.42	5.42	0	4.86	4.86	0
Anxious/Depressed	1.01	1.01	0	1.34	1.34	0
Emotionally Reactive	1.87	1.87	0	2.06	2.06	0
Somatic Complaints	1.71	1.71	0	1.77	1.77	0
Withdrawn	0.83	0.83	0	1.22	1.22	0
Externalizing Composite	8.89	8.89	0	6.64	6.65	0.01
Attention Problems	1.38	1.38	0	1.44	1.44	0
Aggressive Behavior	7.51	7.51	0	5.62	5.63	0.01
Sleep Problems	2.42	2.42	0	2.21	2.22	0.01

Attachment A: SAS Code

```
options nocenter validvarname=upcase;

title 'prj/niddk/ims_analysis/TEDDY/prog_initial_analysis/m_124_dsic.sas';
run;

* DSIC for TEDDY M124. Reproduce Table 2 of M_124_LSmith_NIDDK_Manuscript.pdf ;

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

libname sasfile1 '/prj/niddk/ims_analysis/TEDDY/private_orig_data/M_124_LSmith_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"
  ;

```

```

run;

* where child is 42 months;
%readin(1, m_124_lsmith_niddk_30jun2015_1);
proc freq data=m_124_lsmith_niddk_30jun2015_1;
  tables CDA_42MO_GROUPS CHILDS_AGE/missing;
run;

%readin(1, m_124_lsmith_niddk_30jun2015_2);
proc freq data=m_124_lsmith_niddk_30jun2015_2;
  tables CDA_54MO_GROUPS CHILDS_AGE/missing;
run;

data analy;
  set m_124_lsmith_niddk_30jun2015_1;

  * create subsets;
  if CDA_42MO_GROUPS = 0 then cda_no = 1;
  if CDA_42MO_GROUPS = 1 then cda_unaware = 1;
  if CDA_42MO_GROUPS in(2,3,4) then cda_aware = 1;
run;

proc freq data = analy;
  tables CDA_42MO_GROUPS* cda_no *cda_unaware * cda_aware/list missing;
run;

** Table 2;

* No CDA;
data prntunivno;
  length _VAR_ $100;
  set _null_;
run;

%univ(1 , INTERNALIZING_RAWSCORE_42MO , cda_no , no);
%univ(2 , ANXDEP_RAWSCORE_42MO , cda_no , no);
%univ(3 , EMOTIONAL_RAWSCORE_42MO , cda_no , no);
%univ(4 , SOMATIC_RAWSCORE_42MO , cda_no , no);
%univ(5 , WITHDRAWN_RAWSCORE_42MO , cda_no , no);
%univ(6 , EXTERNALIZING_RAWSCORE_42MO , cda_no , no);
%univ(7 , ATTENTION_RAWSCORE_42MO , cda_no , no);
%univ(8 , AGGRESSIVE_RAWSCORE_42MO , cda_no , no);

```

```

%univ(9 , SLEEPPROBS_RAWSCORE_42MO , cda_no , no);

proc print data= prntunivno noobs;
  var rownum _var_ covarf /*_nobs_ _median_ _q1_ _q3_ _min_ _max_ */ _mean_ _std_;
run;

* Unaware CDA;
data prntunivunaware;
  length _VAR_ $100;
  set _null_;
run;

%univ(1 , INTERNALIZING_RAWSCORE_42MO , cda_unaware , unaware);
%univ(2 , ANXDEP_RAWSCORE_42MO , cda_unaware , unaware);
%univ(3 , EMOTIONAL_RAWSCORE_42MO , cda_unaware , unaware);
%univ(4 , SOMATIC_RAWSCORE_42MO , cda_unaware , unaware);
%univ(5 , WITHDRAWN_RAWSCORE_42MO , cda_unaware , unaware);
%univ(6 , EXTERNALIZING_RAWSCORE_42MO , cda_unaware , unaware);
%univ(7 , ATTENTION_RAWSCORE_42MO , cda_unaware , unaware);
%univ(8 , AGGRESSIVE_RAWSCORE_42MO , cda_unaware , unaware);
%univ(9 , SLEEPPROBS_RAWSCORE_42MO , cda_unaware , unaware);

proc print data= prntunivunaware noobs;
  var rownum _var_ covarf /*_nobs_ _median_ _q1_ _q3_ _min_ _max_ */ _mean_ _std_;
run;

* Aware CDA;
data prntunivaware;
  length _VAR_ $100;
  set _null_;
run;

%univ(1 , INTERNALIZING_RAWSCORE_42MO , cda_aware , aware);
%univ(2 , ANXDEP_RAWSCORE_42MO , cda_aware , aware);
%univ(3 , EMOTIONAL_RAWSCORE_42MO , cda_aware , aware);
%univ(4 , SOMATIC_RAWSCORE_42MO , cda_aware , aware);
%univ(5 , WITHDRAWN_RAWSCORE_42MO , cda_aware , aware);
%univ(6 , EXTERNALIZING_RAWSCORE_42MO , cda_aware , aware);
%univ(7 , ATTENTION_RAWSCORE_42MO , cda_aware , aware);
%univ(8 , AGGRESSIVE_RAWSCORE_42MO , cda_aware , aware);
%univ(9 , SLEEPPROBS_RAWSCORE_42MO , cda_aware , aware);

proc print data= prntunivaware noobs;

```

```

var rownum _var_ covarf /*_nobs_ _median_ _q1_ _q3_ _min_ _max_ */ _mean_ _std_;
run;

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

proc sort data=prntunivunaware (rename=( _mean_ = unaw_mean_
                                         _std_ = unaw_std_ ))
;
  by rownum covarf;
run;

proc sort data=prntunivaware (rename=( _mean_ = aw_mean_
                                       _std_ = aw_std_ ))
;
  by rownum covarf;
run;

data alluniv;
  merge prntunivno      (in=in1 keep = rownum _var_ covarf _mean_ _std_)
        prntunivunaware (in=in2 keep = rownum _var_ covarf unaw_mean_ unaw_std_)
        prntunivaware  (in=in3 keep = rownum _var_ covarf aw_mean_ aw_std_);
  by rownum;
  if in1 or in2 or in3;

  * round to hundredths;
  f_mean_   = round(_mean_      , 0.01);
  f_std_    = round(_std_      , 0.01);
  funaw_mean_ = round(unaw_mean_ , 0.01);
  funaw_std_ = round(unaw_std_  , 0.01);
  faw_mean_  = round(aw_mean_   , 0.01);
  faw_std_   = round(aw_std_    , 0.01);
run;

proc print data=alluniv;
  var rownum  _var_ f_mean_ f_std_
          funaw_mean_ funaw_std_
          faw_mean_ faw_std_;
title3 "Table 2";
run;

```