Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) M53 JYang

Prepared by Sabrina Chen
IMS Inc.
3901 Calverton Blvd, Suite 200 Calverton, MD 20705
February 28, 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_53_jyang_niddk_31jan2013.sas7bdat" dataset.

4 Statistical Methods

Analyses were performed to duplicate results for the data published by JYang et al [1] in *Public Health Nutrition* in 2015. To verify the integrity of the dataset, descriptive statistics were computed.

5 Results

For Table 1 in the publication [1], Characteristics of 8096 TEDDY participants who completed one or more clinic visits between 6 and 36 months 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 1. The results of the replication are similar to the published results.

6 Conclusions

The NIDDK repository is confident that the TEDDY M53 data files to be distributed are a true copy of the study data.

7 References

Yang, J., Lynch, K. F., Uusitalo, U. M., Foterek, K., Hummel, S., Silvis, K., Aronsson, C. A., Riikonen, A., Rewers, M., She, J. X., Ziegler, A. G., Simell, O. G., Toppari, J., Hagopian, W. A., Lernmark, Å., Akolkar, B., Krischer, J. P., Norris, J. M., Virtanen, S. M., Johnson, S. B., TEDDY Study Group (2015). Factors associated with longitudinal food record compliance in a paediatric cohort study. *Public health nutrition*, *19*(5), 804-13.

Table A: Variables used to replicate Table 1: Characteristics of subjects by the status of islet autoimmunity (IA) and type 1 diabetes (T1D) in The Environmental Determinants of Diabetes in the Young (TEDDY) study.

Table Variable	**dataset.variable**
Study center	m_53_jyang_niddk_31jan2013.site
First-degree relative(s) with type 1 diabetes	m_53_jyang_niddk_31jan2013.fdr
Child's sex	m_53_jyang_niddk_31jan2013.gender
Ethnic minority	m_53_jyang_niddk_31jan2013.ethnic_minority
Being the only child	m_53_jyang_niddk_31jan2013.single_child
Maternal age at child's birth (years)	m_53_jyang_niddk_31jan2013.education_mom_group3
Maternal education	m_53_jyang_niddk_31jan2013.maternal_age

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

		N or Mean			% or SD		
		Manuscript	**DSIC**	**Diff**	**Manuscript**	**DSIC**	**Diff**
Study Center	Colorado	1277	1277	0	15.8	15.8	0
	Georgia/Florida	867	858	9	10.7	10.6	0.1
	Washington	1245	1245	0	15.4	15.4	0
	Finland	1760	1760	0	21.7	21.7	0
	Germany	565	565	0	7	7	0
	Sweden	2382	2382	0	29.4	29.4	0
First-degree relative(s) with type 1 diabetes	No	7209	7209	0	89	89	0
	Yes	887	887	0	11	11	0
Child's sex	Male	4117	4117	0	50.9	50.9	0
	Female	3979	3979	0	49.1	49.1	0
Ethnic minority*	No	6332	6332	0	78.2	78.2	0
	Yes	1255	1255	0	15.5	15.5	0
	Missing	509	509	0	6.3	6.3	0
Being the only child	No	4412	4412	0	54.5	54.5	0
	Yes	3150	3150	0	38.9	38.9	0
	Missing	534	534	0	6.6	6.6	0
	Maternal age at child's birth (years)	30.5	30.5	0	5.2	5.2	0
Maternal education	Higher school or less	1516	1516	0	18.7	18.7	0
	Some college or trade school	1911	1911	0	23.6	23.6	0
	Graduated from college	4128	4128	0	51	51	0
	Missing	541	541	0	6.7	6.7	0
Household crowding	Normalized score	2	2	0	1.2	1.2	0

Attachment A: SAS Code

```
options nocenter validvarname=upcase;

title 'prj/niddk/ims_analysis/TEDDY/prog_initial_analysis/m_53_dsic.sas';
run;

* DSIC for TEDDY M53. Reproduce Table 2 of M_53_JYang_NIDDK_Manuscript.pdf ;

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

libname sasfile1 '/prj/niddk/ims_analysis/TEDDY/private_orig_data/M_53_JYang_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/Florida'
  3 = 'Washington'
  4 = 'Finland'
  5 = 'Germany'
  6 = 'Sweden'
  ;


run;


%readin(1, m_53_jyang_niddk_31jan2013);

proc freq data=m_53_jyang_niddk_31jan2013;
  tables gender site fdr ethnic_minority single_child maternal_age education_mom_group3 crowding_norm/missing;
run;


data analy;
  set m_53_jyang_niddk_31jan2013;
```

```
  subset_all = 1;
run;


** Table 1;

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

%npercent(1   , site                 , site    , subset_all , all);
%npercent(2   , fdr                  , fdr     , subset_all , all);
%npercent(3   , gender               , gender  , subset_all , all);
%npercent(4   , ethnic_minority              , yesno   , subset_all , all);
%npercent(5   , single_child         , yesno   , subset_all , all);
%npercent(7   , education_mom_group3 , mateduc , subset_all , all);

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


* mean std;
data prntunivall;
  length _VAR_ $100;
  set _null_;
run;

%univ(6   , maternal_age , subset_all , all);
%univ(8   , crowding_norm     , subset_all , all);

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


data table1;
  set prntall         (keep = rownum  covar covarf count    percent)
      prntunivall      (keep = rownum _var_ covarf _mean_ _std_) ;

  * round to hundredths;
  _mean_ = round(_mean_ , 0.1);
  _std_  = round(_std_  , 0.1);
  percent = round(percent, 0.1);
```

```
run;

proc sort data=table1;
  by rownum;
run;

proc print data=table1;
  var rownum  _var_ covar covarf count percent _mean_ _std_      ;
title3 "Table 1";
run;
```