Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) M64 SJohnson

Prepared by Sabrina Chen
IMS Inc.
3901 Calverton Blvd, Suite 200 Calverton, MD 20705
May 28, 2020

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_64_SJohnson_NIDDK_Submission folder in the data package. For this replication, variables were taken from the “m_64_sjohnson_niddk_31oct2015_1.sas7bdat” and “m_64_sjohnson_niddk_31oct2015_3.sas7bdat" datasets.

4 Statistical Methods

Analyses were performed to duplicate results for the data published by Suzanne Johnson et al [1] in Diabetes Care 2017. 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 almost an exact match to the published results.

7 References

[1] Suzanne Bennett Johnson, Kristian F. Lynch, Roswith Roth, Desmond Schatz, and the TEDDY Study Group. My Child Is Islet Autoantibody Positive: Impact on Parental Anxiety. Diabetes Care 2017 Sep; 40(9): 1167-1172.

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

Table Variable	**dataset.variable**
Country	m_64_sjohnson_niddk_31oct2015_1.country
Child has FDR relative with type 1 diabetes	m_64_sjohnson_niddk_31oct2015_1.fdr
SAI score before IA+ test result	m_64_sjohnson_niddk_31oct2015_1.told_ia_pos
Risk perception (mother)	m_64_sjohnson_niddk_31oct2015_3.anydev_before
Risk perception (father)	m_64_sjohnson_niddk_31oct2015_3.anydev_dad_before
Fathers 6-item State Anxiety Inventory (SAI) score as reported immediately before first IA result	m_64_sjohnson_niddk_31oct2015_3.stai_dad_before
Fathers 6-item State Anxiety Inventory (SAI) score as reported immediately after first IA result	m_64_sjohnson_niddk_31oct2015_3.stai_dad_after

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

		Mother			Father		
		Manuscript	**DSIC**	**Diff**	**Manuscript**	**DSIC**	**Diff**
Country	USA	331	331	0	264	264	0
	Finland	160	161	-1	155	155	0
	Germany	30	30	0	28	28	0
	Sweden	292	292	0	276	276	0
Child has FDR relative with type 1 diabetes	no	694	695	-1	618	618	0
	yes	119	119	0	105	105	0
SAI score before IA+ test result		813	814	-1	723	723	0
Risk perception	Underestimate	299	299	0	365	365	0
	Accurate	514	514	0	358	358	0

Attachment A: SAS Code

```
options nocenter validvarname=upcase;

title '/prj/niddk/ims_analysis/TEDDY/prog_initial_analysis/m_64_dsic.sas';
run;

libname pcsas "/prj/niddk/ims_analysis/TEDDY/private_orig_data/M_64_SJohnson_NIDDK_Submission/";



proc format;
 value val
 .     = "no value"
 other = "   value"
 ;

 value countryf
 1 = 'US'
 2 = 'Finland'
 3 = 'Germany'
 4 = 'Sweden'
 ;

 value yesnof
 0 = "No"
 1 = "Yes"
 ;

run;


* produce n and %;
%macro npercent(ds, rownum, var, varf, subset,  subsetname);
  proc freq data=&ds 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;


data analy1;
  set pcsas.m_64_sjohnson_niddk_31oct2015_1;
run;

proc contents data=analy1;
run;

proc freq data=analy1;
  tables TOLD_IA_POS*FOLLOWED_AFTER_IA /list missing;
format maskid val.;
run;


data analy2;
  set pcsas.m_64_sjohnson_niddk_31oct2015_2;
run;

proc contents data=analy2;
run;

proc freq data=analy2;
format maskid val.;
run;


data analy3;
  set pcsas.m_64_sjohnson_niddk_31oct2015_3;
run;

proc contents data=analy3;
run;

proc freq data=analy3;
*  tables IA_VISIT/missing;
```

```
format maskid val.;
  run;



data analy4;
  set pcsas.m_64_sjohnson_niddk_31oct2015_4;
run;

proc contents data=analy4;
run;

proc freq data=analy4 ;
 * tables AB_TOLD_GRPS MULTI_TOLD persist_told/missing;
 * tables AB_TOLD_GRPS*MULTI_TOLD*persist_told/list missing;
 format maskid val.;
  run;



proc sort data=analy1;
  by maskid;
  run;

proc sort data=analy3;
  by maskid;
  run;

data combine;
  merge analy1    (in=in1 keep=maskid country fdr told_ia_pos)
        analy3    (in=in2 keep=maskid anydev_before anydev_dad_before stai_dad_before stai_dad_after);
  by maskid;
  if in2 then in_analy3=1;  * subset for mother;

  * per Mike Toth email (4/15/2020)- subset for father;
  if (anydev_dad_before ne .) and (stai_dad_before ne .) then in_father=1;

run;

proc freq data=combine;
  tables in_father*anydev_dad_before* stai_dad_before*stai_dad_after/list missing;
  format anydev_dad_before  stai_dad_before stai_dad_after val.;
  title3 "checking";
  run;
```

```
* Table 1;
proc freq data=combine;
  where in_analy3=1;
  tables country fdr TOLD_IA_POS /*stai_long*/ anydev_before /missing;
  format country countryf. fdr yesnof.;
title3 "subset to mother";
  run;


proc freq data=combine;
  where in_father=1;
  tables country fdr TOLD_IA_POS /*stai_long*/ anydev_dad_before /missing;
  format country countryf. fdr yesnof.;
title3 "subset to father";
  run;
```