Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) M176 RJohnson

Prepared by Sabrina Chen
IMS Inc.
3901 Calverton Blvd, Suite 200 Calverton, MD 20705
Aug 13, 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_176_RJohnson_NIDDK_Submission folder in the data package. For this replication, variables were taken from the “m_176_rjohnson_niddk_31may2012_2.sas7bdat”, “m_176_rjohnson_niddk_31may2012_3.sas7bdat ", and “m_176_rjohnson_niddk_31may2012_4.sas7bdat" datasets.

4 Statistical Methods

Analyses were performed to duplicate results for the data published by Randi K. Johnson et al [1] in Scientific Reports 2019. 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 an exact match to the published results.

7 References

[1] Johnson, R.K., Vanderlinden, L., DeFelice, B.C. *et al.* Metabolite-related dietary patterns and the development of islet autoimmunity. *Sci Rep* **9,** 14819 (2019). https://doi.org/10.1038/s41598-019-51251-4.

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

Table Variable	**dataset.variable**
Risk Set Matching Strata	m_176_rjohnson_niddk_31may2012_2.case_ind
IA = 1, else = 0	m_176_rjohnson_niddk_31may2012_2.outcome
IAA+ first and only at IA+ = 1, else = 0	m_176_rjohnson_niddk_31may2012_2.iaaonly
GADA+ first and only at IA+ = 1, else = 0	m_176_rjohnson_niddk_31may2012_2.gadaonly
mAb+ = 1, else = 0	m_176_rjohnson_niddk_31may2012_2.multab
Age(days) at Metabolomics Blood Draw	m_176_rjohnson_niddk_31may2012_2.drawdate_age_days
Risk Set Matching Strata	m_176_rjohnson_niddk_31may2012_3.case_ind
IA = 1, else = 0	m_176_rjohnson_niddk_31may2012_3.outcome
IAA+ first and only at IA+ = 1, else = 0	m_176_rjohnson_niddk_31may2012_3.iaaonly
GADA+ first and only at IA+ = 1, else = 0	m_176_rjohnson_niddk_31may2012_3.gadaonly
mAb+ = 1, else = 0	m_176_rjohnson_niddk_31may2012_3.multab
Age(days) at Metabolomics Blood Draw	m_176_rjohnson_niddk_31may2012_3.drawdate_age_days
Risk Set Matching Strata	m_176_rjohnson_niddk_31may2012_4.case_ind
IA = 1, else = 0	m_176_rjohnson_niddk_31may2012_4.outcome
IAA+ first and only at IA+ = 1, else = 0	m_176_rjohnson_niddk_31may2012_4.iaaonly
GADA+ first and only at IA+ = 1, else = 0	m_176_rjohnson_niddk_31may2012_4.gadaonly
mAb+ = 1, else = 0	m_176_rjohnson_niddk_31may2012_4.multab
Age(days) at Metabolomics Blood Draw	m_176_rjohnson_niddk_31may2012_4.drawdate_age_days

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

	IA					
	DSIC	Manuscript	Diff	DSIC	Manuscript	Diff
	n			mean(SD)		
Cross-Section	N_OUTCOME			AGE_OUTCOME		
Seroconversion	352	352	0	722(446)	722 (446)	0(0)
Pre-Seroconversion	366	366	0	625(412)	625 (412)	0(0)
Infancy 9-months	253	253	0	283(14)	283 (14)	0(0)

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

	IAA								
	DSIC	Manuscript	Diff	DSIC	Manuscript	Diff	DSIC	Manuscript	Diff
	n			% of IA cases			mean(SD)		
Cross-Section	N_IAAONLY			IAAONLY1_PCT			AGE_IAAONLY		
Seroconversion	171	171	0	48.6	48.6	0	586(370)	586 (370)	0(0)
Pre-Seroconversion	180	180	0	49.2	49.2	0	505(366)	505 (366)	0(0)
Infancy 9-months	114	114	0	45.1	45.1	0	283(14)	283 (14)	0(0)

Table B-3 GADA: Comparison of values computed in integrity check to reference article data values

	GADA								
	DSIC	Manuscript	Diff	DSIC	Manuscript	Diff	DSIC	Manuscript	Diff
	n			% of IA cases			mean(SD)		
Cross-Section	N_GADAONLY			GADAONLY1_PCT			AGE_GADAONLY		
Seroconversion	113	113	0	32.1	32.1	0	888(509)	888 (509)	0(0)
Pre-Seroconversion	116	116	0	31.7	31.7	0	786(445)	786 (445)	0(0)
Infancy 9-months	83	83	0	32.8	32.8	0	284(16)	284 (16)	0(0)

Table B-4 mAB+: Comparison of values computed in integrity check to reference article data values

	mAb+								
	DSIC	Manuscript	Diff	DSIC	Manuscript	Diff	DSIC	Manuscript	Diff
	n			% of IA cases			mean(SD)		
Cross-Section	N_MULTAB			MULTAB1_PCT			AGE_MULTAB		
Seroconversion	211	211	0	59.9	59.9	0	655(365)	655 (365)	0(0)
Pre-Seroconversion	224	224	0	61.2	61.2	0	541(346)	541 (346)	0(0)
Infancy 9-months	153	153	0	60.5	60.5	0	282(14)	282 (14)	0(0)

Attachment A: SAS Code

```
options nocenter validvarname=upcase;

title '/prj/niddk/ims_analysis/TEDDY/prog_initial_analysis/m_176_dsic.sas';
run;

libname pcsas "/prj/niddk/ims_analysis/TEDDY/private_orig_data/M_176_RJohnson_NIDDK_Submission/";



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

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

run;



%macro match(dsnum);

        data analy&dsnum;
          set pcsas.m_176_rjohnson_niddk_31may2012_&dsnum;
        run;

        proc contents data=analy&dsnum;
        title3 "analy &dsnum";
        run;

        proc freq data=analy&dsnum;
          tables outcome IAAONLY GADAONLY MULTAB/list missing;
        run;


        * collapse by case set and hold the age of the case;
        proc sort data=analy&dsnum;
          by case_ind;
          run;

        data match&dsnum (keep=case_ind outcome1_age outcome_cntrl outcome_case
```

```
              iaaonly1_age  iaaonly_cntrl    iaaonly_case
              gadaonly1_age gadaonly_cntrl   gadaonly_case
              multab1_age   multab_cntrl     multab_case        );
set analy&dsnum;
by case_ind;

retain outcome1_age outcome_cntrl    outcome_case
       iaaonly1_age  iaaonly_cntrl    iaaonly_case
       gadaonly1_age gadaonly_cntrl   gadaonly_case
       multab1_age   multab_cntrl     multab_case
       ;

if first.case_ind then do;
  outcome1_age   = .;
  outcome_case   = .;
  outcome_cntrl  = 0;
  iaaonly1_age    = .;
  iaaonly_cntrl  = .;
  iaaonly_case   = 0;
  gadaonly1_age   = .;
  gadaonly_cntrl = .;
  gadaonly_case  = 0;
  multab1_age     = .;
  multab_cntrl   = .;
  multab_case    = 0;
end;

if outcome = 0 then outcome_cntrl = sum(outcome_cntrl, 1);
if outcome = 1 then do;
  outcome_case = outcome;
  outcome1_age = drawdate_age_days;
end;

if iaaonly = 0 then iaaonly_cntrl = sum(iaaonly_cntrl, 1);
if iaaonly = 1 then do;
  iaaonly_case = iaaonly;
  iaaonly1_age = drawdate_age_days;
end;

if gadaonly = 0 then gadaonly_cntrl = sum(gadaonly_cntrl, 1);
if gadaonly = 1 then do;
  gadaonly_case = gadaonly;
  gadaonly1_age = drawdate_age_days;
end;

if multab = 0 then multab_cntrl = sum(multab_cntrl, 1);
if multab = 1 then do;
  multab_case = multab;
  multab1_age = drawdate_age_days;
```

```
          end;

          if last.case_ind;

        run;

        data analy&dsnum;
          merge analy&dsnum (in=in1) match&dsnum;
          by case_ind;
          if in1;
        run;

        proc print data=analy&dsnum (obs=30);
          by case_ind;
          id case_ind;
          var mask_id outcome drawdate_age_days outcome1_age outcome_cntrl outcome_case
                      iaaonly  iaaonly1_age  iaaonly_cntrl    iaaonly_case
                      gadaonly gadaonly1_age gadaonly_cntrl   gadaonly_case
                      multab multab1_age   multab_cntrl     multab_case  ;
          title3 "analy &dsnum- check collapse";
        run;

        proc freq data=match&dsnum;
          tables  outcome_cntrl* outcome_case/list missing;
          tables  iaaonly_cntrl *   iaaonly_case  /list missing;
          tables  gadaonly_cntrl*   gadaonly_case /list missing;
          tables  multab_cntrl  *   multab_case   /list missing;
         run;

%mend;



%macro tb1means(cs, var, dsnum);

        proc means data=match&dsnum;
          where &var._cntrl > 0;
          var &var.1_age;
          output out=means_&var;
        run;

        proc transpose data=means_&var out=means_&var.2;
        run;

        data means_&var.2 (keep=cs  _NAME_ n_&var age_&var);
          length cs $20;
          set means_&var.2;
          age_&var = compress(round(col4) || "(" || round(col5) || ")");
          n_&var = col1;
```

```
          cs = "&cs";

          if substr(_name_,1,1) ne "_";

        run;

        proc print data=means_&var.2;
        run;
%mend;



%macro tb1pct(var, dsnum);

        proc freq data=match&dsnum noprint;
         where outcome_cntrl > 0;
         tables  outcome_case*&var._case /list missing out=pcnt_&var;
        run;

        data pcnt_&var (keep=_name_ &var.1_pct);
          set pcnt_&var;
          _name_ = upcase("&var.1_AGE");
          &var.1_pct = round(percent, 0.1);
          if &var._case = 1;
        run;

        proc print data=pcnt_&var;
        run;

%mend;


** Stats for Seroconversion row;
%match(2);

%tb1means(Seroconversion, outcome, 2);
%tb1means(Seroconversion, iaaonly, 2);
%tb1pct  (iaaonly, 2);
%tb1means(Seroconversion, gadaonly, 2);
%tb1pct  (gadaonly, 2);
%tb1means(Seroconversion, multab, 2);
%tb1pct  (multab, 2);

data row1;
  merge means_outcome2 means_iaaonly2 pcnt_iaaonly means_gadaonly2 pcnt_gadaonly means_multab2 pcnt_multab;
run;

proc print data=row1;
```

```
  var cs  n_outcome age_outcome    n_iaaonly    iaaonly1_pct age_iaaonly    n_gadaonly   gadaonly1_pct  age_gadaonly     n_multab    multab1_pct
age_multab;
run;


** Stats for pre-sero row;
%match(3);

%tb1means(Pre-Seroconversion, outcome, 3);
%tb1means(Pre-Seroconversion, iaaonly, 3);
%tb1pct  (iaaonly, 3);
%tb1means(Pre-Seroconversion, gadaonly, 3);
%tb1pct  (gadaonly, 3);
%tb1means(Pre-Seroconversion, multab, 3);
%tb1pct  (multab, 3);

data row2;
  merge means_outcome2 means_iaaonly2 pcnt_iaaonly means_gadaonly2 pcnt_gadaonly means_multab2 pcnt_multab;
run;

proc print data=row2;
  var cs  n_outcome age_outcome    n_iaaonly    iaaonly1_pct age_iaaonly    n_gadaonly   gadaonly1_pct  age_gadaonly     n_multab    multab1_pct
age_multab;
run;


** stats for Infancy 9-months;
%match(4)

%tb1means(Infancy 9-months, outcome, 4);
%tb1means(Infancy 9-months,iaaonly, 4);
%tb1pct  (iaaonly, 4);
%tb1means(Infancy 9-months,gadaonly, 4);
%tb1pct  (gadaonly, 4);
%tb1means(Infancy 9-months,multab, 4);
%tb1pct  (multab, 4);

data row3;
  merge means_outcome2 means_iaaonly2 pcnt_iaaonly means_gadaonly2 pcnt_gadaonly means_multab2 pcnt_multab;
run;

proc print data=row3;
  var cs  n_outcome age_outcome    n_iaaonly    iaaonly1_pct age_iaaonly    n_gadaonly   gadaonly1_pct  age_gadaonly     n_multab    multab1_pct
age_multab;
run;


* combine;
data table1;
```

```
  set row1 row2 row3;
run;

proc print data=table1;
  var cs  n_outcome age_outcome    n_iaaonly    iaaonly1_pct age_iaaonly    n_gadaonly   gadaonly1_pct  age_gadaonly     n_multab    multab1_pct
age_multab;
  title3 "Table 1";
run;


proc sql;
  create table tbl1_out as
  select cs,  n_outcome, age_outcome ,   n_iaaonly  ,  iaaonly1_pct ,age_iaaonly  ,  n_gadaonly ,  gadaonly1_pct , age_gadaonly ,    n_multab ,
multab1_pct,  age_multab from table1;
quit;


proc export outfile='/prj/niddk/ims_analysis/TEDDY/private_created_data/TEDDY.m176.Table1.xls'
  dbms=xls
  replace
  data=tbl1_out;
```