Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) M112 Yang

Prepared by Sabrina Chen
IMS Inc.
3901 Calverton Blvd, Suite 200 Calverton, MD 20705
September 11, 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 data in the publication. 4

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

Attachment A: SAS Code 8

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_112_JYang_NIDDK_Submission folder in the data package. For this replication, variables were taken from the “m_112_jyang_niddk_29mar2016.sas7bdat” dataset.

4 Statistical Methods

Analyses were performed to duplicate results for the data published by Jimin Yang et al [1] in the British Journal of Nutrition in 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] Jimin Yang, Roy N. Tamura, Carin A. Aronsson, Ulla M. Uusitalo, Åke Lernmark, Marian Rewers, William A. Hagopian, Jin-Xiong She, Jorma Toppari, Anette G. Ziegler, Beena Akolkar, Jeffrey P. Krischer, Jill M. Norris, Suvi M. Virtanen, Daniel Agardh and The Environmental Determinants of Diabetes in The Young study group. Maternal use of dietary supplements during pregnancy is not associated with coeliac disease in the offspring: The Environmental Determinants of Diabetes in the Young (TEDDY) study. British Journal of Nutrition (2017), 117, 466–472.

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

Table Variable	**dataset.variable**
Country	m_112_jyang_niddk_29mar2016.country
HLA Category	m_112_jyang_niddk_29mar2016.hla_category
Vitamin D	m_112_jyang_niddk_29mar2016.sum_tot_vit_d_ug
n-3 FA	m_112_jyang_niddk_29mar2016.sum_omega3_g
Fe	m_112_jyang_niddk_29mar2016.sum_tot_iron_content_mg

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

Variable	Country	Manuscript	DSIC	Diff.	Manuscript	DSIC	Diff.	Manuscript	DSIC	Diff.
		N			**Percent**					
Vitamin D	**All**	4369	4312	57	66	65	1			
	USA	2178	2163	15	82	82	0			
	Finland	1104	1086	18	73	72	1			
	Germany	87	83	4	22	21	1			
	Sweden	1000	980	20	48	47	1			
n-3 FA	**All**	1107	1055	52	17	16	1			
	USA	677	651	26	26	25	1			
	Finland	138	127	11	9	8	1			
	Germany	138	130	8	34	32	2			
	Sweden	154	147	7	7	7	0			
Fe	**All**	6216	5910	306	94	89	5			
	USA	2579	2542	37	97	96	1			

Variable	Country	Manuscript	DSIC	Diff.	Manuscript	DSIC	Diff.	Manuscript	DSIC	Diff.
	Finland	1274	1234	40	85	82	3			
	Germany	316	356	-40	78	88	-10			
	Sweden	2047	1778	269	99	86	13			
		Median			**Q1**			**Q3**		
Vitamin D	**All**	1540	1470	70	1000	980	20	2800	2800	0
	USA	2800	2800	0	2030	2000	30	2800	2800	0
	Finland	1200	1260	-60	680	665	15	1400	1400	0
	Germany	1050	1050	0	600	600	0	1400	1400	0
	Sweden	1050	1050	0	600	600	0	1120	1155	-35
n-3 FA	**All**	62	62	0	35	35	0	92	92	0
	USA	67	67	0	42	42	0	84	84	0
	Finland	75	75	0	26	26	0	153	153	0
	Germany	38	38	0	24	24	0	56	56	0
	Sweden	73	73	0	29	29	0	130	130	0

Variable	Country	Manuscript	DSIC	Diff.	Manuscript	DSIC	Diff.	Manuscript	DSIC	Diff.
Fe	**All**	7728	7728	0	4200	4200	0	11410	11410	0
	USA	7840	7840	0	5040	5040	0	7840	7840	0
	Finland	3186	3200	-14	2030	2030	0	9800	9800	0
	Germany	9600	9600	0	3500	3500	0	21420	21420	0
	Sweden	9200	9200	0	4200	4200	0	16000	16000	0

Attachment A: SAS Code

```
options nocenter validvarname=upcase;

title '/prj/niddk/ims_analysis/TEDDY/prog_initial_analysis/m_112_dsic_20190709.sas';
run;

libname dat '/prj/niddk/ims_analysis/TEDDY/private_orig_data/M_112_JYang_NIDDK_Submission';


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

 value oneplus
 . = "no value"
 0 = "0"
 1-high = "1+"
 ;

 value zerohi
 . = "no value"
 0-high = "0-high"
 ;
run;



* 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;
    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;


%macro univ(rownum, var, subset, supplement, subsetname);

  proc univariate data=analy outtable= univ&subsetname noprint;
    where &subset=1 and &supplement=1 and &var not in(.,0);
    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;



data analy;
   length hla_category_desc_v2 $28;
   set dat.m_112_jyang_niddk_29mar2016;

   * convert to months;
   time_to_brstfed_stop_mon = time_to_brstfed_stop/30;

   * create other grouping;
   if hla_category in(3,5,6,7,8,10) then hla_category_desc_v2 = "Other";
   else hla_category_desc_v2 = hla_category_desc;
```

```
   * subsets;
   if HLA_CATEGORY not in(-1, 0) then do;
     * create subset flag for each row to use in macro call;
     all = 1;
     * country;
     if country=1 then usa=1;
     else if country=2 then Finland=1;
     else if country=3 then Germany=1;
     else if country=4 then Sweden=1;
   end;


   *Vitamin D:  If MOTHER_A1 or MOTHER_B2  or MOTHER_B5 or MOTHER_B6 or MOTHER_B7  =  1 then   mother took Vitamin-D
         ;
   if max(MOTHER_A1, MOTHER_B2, MOTHER_B5, MOTHER_B6, MOTHER_B7)  =  1 then vitamin_d_stat = 1;
   else vitamin_d_stat = 0;

   *Fatty Acid: IF MOTHER_A4 or MOTHER_B4  or MOTHER_B5 or MOTHER_B6 or MOTHER_B8 or MOTHER_B10 = 1 then mother took Omega3 Fatty
Acid.   ;
   if max(MOTHER_A4, MOTHER_B4, MOTHER_B5, MOTHER_B6, MOTHER_B8, MOTHER_B10) = 1 then Omega3_stat = 1;
   else omega3_stat = 0;

   *Iron:       IF MOTHER_A11  or MOTHER_B1 or MOTHER_B2 or MOTHER_B3 or MOTHER_B4 or MOTHER_B5 or MOTHER_B6 or MOTHER_B7 or
MOTYHER_B8 or MOTHER_B9 or MOTHER_B10  = 1 then mother took Iron ;
   if max(MOTHER_A11, MOTHER_B1, MOTHER_B2, MOTHER_B3, MOTHER_B4, MOTHER_B5, MOTHER_B6, MOTHER_B7, MOTHER_B8,  MOTHER_B10) = 1 then
iron_stat = 1;
   else iron_stat = 0;

run;

proc contents data = analy;
title3 'm112';
run;

* check for dups;
proc sort data=analy;
  by MASKID;
run;

data dups;
  set analy;
  by MASKID;
  if not (first.MASKID and last.MASKID);
run;
```

```
data _null_;
  set dups;
  abort;
run;

proc freq data=analy;
  tables hla_category_desc_v2*hla_category_desc/list missing;
  tables NOMOMSUPP/list missing;
  tables NOMOMSUPP*hla_category/list missing;
  tables NOMOMSUPP*SUM_TIMES_IRON * SUM_TIMES_OMEGA3 *  SUM_TIMES_VITD    /list missing;
  tables SUM_OMEGA3_G SUM_TOT_IRON_CONTENT_MG SUM_TOT_VIT_D_UG/missing;
  tables vitamin_d_stat* MOTHER_A1 * MOTHER_B2  * MOTHER_B5 * MOTHER_B6 * MOTHER_B7/list missing;
  tables omega3_stat*MOTHER_A4* MOTHER_B4* MOTHER_B5* MOTHER_B6* MOTHER_B8* MOTHER_B10/list missing;
  tables iron_stat* MOTHER_A11 *MOTHER_B1* MOTHER_B2* MOTHER_B3* MOTHER_B4* MOTHER_B5 * MOTHER_B6 * MOTHER_B7 * MOTHER_B8 *
MOTHER_B10/list missing;
  tables vitamin_d_stat*omega3_stat*iron_stat/list missing;
  tables vitamin_d_stat omega3_stat iron_stat/list missing;
  format SUM_TIMES_IRON  SUM_TIMES_OMEGA3   SUM_TIMES_VITD SUM_OMEGA3_G SUM_TOT_IRON_CONTENT_MG SUM_TOT_VIT_D_UG oneplus.;
  title3 'check new groupings';
run;


proc freq data=analy;
  where HLA_CATEGORY not in(-1, 0);
  tables COUNTRY/missing;
  tables all * country * usa*Finland*Germany*Sweden/list missing;
  tables SUM_TIMES_VITD SUM_TIMES_OMEGA3 SUM_TIMES_IRON    /list missing;
  tables SUM_TOT_VIT_D_UG SUM_OMEGA3_G SUM_TOT_IRON_CONTENT_MG /missing;
  format SUM_TIMES_IRON  SUM_TIMES_OMEGA3   SUM_TIMES_VITD oneplus.
         SUM_OMEGA3_G SUM_TOT_IRON_CONTENT_MG SUM_TOT_VIT_D_UG zerohi.;
run;


* med, q1, q3;
data prntunivall;
  length _VAR_ $100;
  set _null_;
run;

%univ(1   , SUM_TOT_VIT_D_UG        ,  all ,vitamin_d_stat      , all);
%univ(2   , SUM_OMEGA3_G            ,  all ,omega3_stat         , all);
%univ(3   , SUM_TOT_IRON_CONTENT_MG ,  all ,iron_stat           , all);

data prntunivusa;
  length _VAR_ $100;
```

```
  set _null_;
run;

%univ(1   , SUM_TOT_VIT_D_UG        , usa , vitamin_d_stat      ,usa);
%univ(2   , SUM_OMEGA3_G            , usa , omega3_stat         ,usa);
%univ(3   , SUM_TOT_IRON_CONTENT_MG , usa , iron_stat           ,usa);


data prntunivfinland;
  length _VAR_ $100;
  set _null_;
run;

%univ(1   , SUM_TOT_VIT_D_UG        , finland ,vitamin_d_stat      , finland);
%univ(2   , SUM_OMEGA3_G            , finland ,omega3_stat         , finland);
%univ(3   , SUM_TOT_IRON_CONTENT_MG , finland ,iron_stat           , finland);



data prntunivgermany;
  length _VAR_ $100;
  set _null_;
run;

%univ(1   , SUM_TOT_VIT_D_UG        , germany ,vitamin_d_stat      , germany);
%univ(2   , SUM_OMEGA3_G            , germany ,omega3_stat         , germany);
%univ(3   , SUM_TOT_IRON_CONTENT_MG , germany ,iron_stat           , germany);


data prntunivsweden;
  length _VAR_ $100;
  set _null_;
run;

%univ(1   , SUM_TOT_VIT_D_UG        , sweden , vitamin_d_stat      ,sweden);
%univ(2   , SUM_OMEGA3_G            , sweden , omega3_stat         ,sweden);
%univ(3   , SUM_TOT_IRON_CONTENT_MG , sweden , iron_stat           ,sweden);



data alluniv;
  set prntunivall     (in=in1 keep = rownum _var_ covarf _nobs_ _median_ _q1_ _q3_)
      prntunivusa     (in=in1 keep = rownum _var_ covarf _nobs_ _median_ _q1_ _q3_)
      prntunivfinland (in=in1 keep = rownum _var_ covarf _nobs_ _median_ _q1_ _q3_)
      prntunivgermany (in=in1 keep = rownum _var_ covarf _nobs_ _median_ _q1_ _q3_)
```

```
      prntunivsweden  (in=in1 keep = rownum _var_ covarf _nobs_ _median_ _q1_ _q3_)
  ;
  _median_   = round(_median_ );
  _q1_       = round(_q1_     );
  _q3_      = round(_q3_     );
run;

proc sort data=alluniv;
  by rownum;
  run;

proc print data= alluniv noobs;
    var rownum _var_  covarf _nobs_ _median_ _q1_ _q3_ /*_min_ _max_ _std_*/;
    title3 "Table 1 - median, q1, q3 for each subset";
run;


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

%npercent(1, SUM_TOT_VIT_D_UG         , zerohi      , all       ,all);
%npercent(2, SUM_OMEGA3_G             , zerohi      , all       ,all);
%npercent(3, SUM_TOT_IRON_CONTENT_MG  , zerohi      , all       ,all);



data prntusa;
  set _null_;
run;

%npercent(1, SUM_TOT_VIT_D_UG         , zerohi      , usa       , usa);
%npercent(2, SUM_OMEGA3_G             , zerohi      , usa       , usa);
%npercent(3, SUM_TOT_IRON_CONTENT_MG  , zerohi      , usa       , usa);



data prntfinland;
  set _null_;
run;

%npercent(1, SUM_TOT_VIT_D_UG         , zerohi      , finland       , finland);
%npercent(2, SUM_OMEGA3_G             , zerohi      , finland       , finland);
%npercent(3, SUM_TOT_IRON_CONTENT_MG  , zerohi      , finland       , finland);
```

```
data prntgermany;
  set _null_;
run;

%npercent(1, SUM_TOT_VIT_D_UG         , zerohi      , germany      , germany);
%npercent(2, SUM_OMEGA3_G             , zerohi      , germany      , germany);
%npercent(3, SUM_TOT_IRON_CONTENT_MG  , zerohi      , germany      , germany);



data prntsweden;
  set _null_;
run;

%npercent(1, SUM_TOT_VIT_D_UG         , zerohi      , sweden       , sweden);
%npercent(2, SUM_OMEGA3_G             , zerohi      , sweden       , sweden);
%npercent(3, SUM_TOT_IRON_CONTENT_MG  , zerohi      , sweden       , sweden);



* Table 1;
data npercent;
  length subgroup $10;
  set prntall (in=in1) prntusa (in=in2) prntfinland (in=in3) prntgermany (in=in4) prntsweden (in=in5);
  if in1 then subgroup = "all";
  if in2 then subgroup = "usa";
  if in3 then subgroup = "finland";
  if in4 then subgroup = "germany";
  if in5 then subgroup = "sweden";

  percent = round(percent);
run;

proc sort data=npercent;
  by rownum covarf;
run;

proc print data=npercent;
    where covarf = "0-high";
  var rownum subgroup covar covarf count percent;
  title3 "Table 1 - n, percent";
run;
```