Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) Pub35 JYang

Prepared by Michael Spriggs
IMS Inc.
3901 Calverton Blvd, Suite 200 Calverton MD 20705
September 18, 2015

Table of 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 “pub35_jyang_niddk” dataset.

4 Statistical Methods

Analyses were performed to duplicate results for the data published by J Yang et al [1]. To verify the integrity of the dataset, descriptive statistics were computed, by HLA DQ genotype.

5 Results

Table A lists the variables that were used in the replication and Table B compares the results calculated from the archived data file to the results published in Table 1. The results of the replication are identical with only percentage rounding differences?

6 Conclusions

The NIDDK repository is confident that the TEDDY Pub35 JYang data files to be distributed are a true copy of the study data.

7 References

[1]Yang, J., et al. "Prevalence of obesity was related to HLA-DQ in 2–4-year-old children at genetic risk for type 1 diabetes." International Journal of Obesity 38.12 (2014): 1491-1496.

Table A: Variables used to replicate Tables 1 in the publication.

Table Variable	Variables Used in Replication from the "Table 1"
Sex	sex
Age	current_age_mos
Country	country
Having FDR with type 1 diabetes	fdr
Birth weight (g)	babysweightgrams
Maternal Diabetes during pregnancy	diabetes
Maternal Prepregnancy BMI	BMI
Gestational age (weeks)	gestational_age
Gestational weight gain (kg)	weightgain

Table B: Comparison of values computed in integrity check to reference article Table 1 values: 1 Characteristics of the children by presence of diabetes in the family: The Environmental Determinants of Diabetes in the Young (TEDDY) birth cohort

	DQ2/8 [Manuscript]	**DQ2/8 [DSIC]**	**DQ2/8 [Difference]**	**DQ8/8 [Manuscript]**	**DQ8/8 [DSIC]**	**DQ8/8 [Difference]**
Sex (male/female)	1195/1146	1195/1146	0	593/587	593/587	0
Age (months)	55.8 (17.2)	55.8 (17.2)	0	55.7 (17.2)	55.7 (17.2)	0
United States (n=2314)	923 (40.0%)	923 (39.9%)	0(0.1)	473 (20.4%)	473 (20.4%)	0
Finland (n=1374)	465 (33.8%)	465 (33.8%)	0	222 (16.2%)	222 (16.2%)	0
Germany (n= 357)	141 (39.5%)	141 (39.5%)	0	64 (17.9%)	64 (17.9%)	0
Sweden (n=1924)	812 (42.2%)	812 (42.2%)	0	421 (21.9%)	421 (21.9%)	0
FDR with type 1 diabetes: Yes (n=675)	212 (31.4%)	212 (31.4%)	0	112 (16.6%)	112 (16.6%)	0
FDR with type 1 diabetes: No (n=5294)	2129 (40.2%)	2129 (40.2%)	0	1068 (20.2%)	1068 (20.2%)	0
Birth weight (g)	3516.2 (554.3)	3516.2 (554.3)	0	3510.4 (530.5)	3510.4 (530.5)	0
Pregnancy Diabetes Type 1(n=235)	71 (30.2%)	71 (30.2%)	0	41 (17.4%)	41 (17.4%)	0
Pregnancy Diabetes Type 2 (n= 16)	3 (18.7%)	3 (18.8%)	0(-0.1)	4 (25.0%)	4 (25.0%)	0
Pregnancy Diabetes GDM (n=330)	137 (41.5%)	137 (41.5%)	0	63 (19.1%)	63 (19.1%)	0
Pregnancy Diabetes None (n=5203)	2064 (39.6%)	2064 (39.7%)	0	1039 (20.0%)	1039 (20.0%)	0
Prepregnancy BMI	24.7 (5.1)	24.7 (5.1)	0	24.7 (5.1)	24.7 (5.1)	0
Gestational age (weeks)	39.6 (1.6)	39.6 (1.6)	0	39.5 (1.5)	39.5 (1.5)	0
Gestational weight gain (kg)	14.9 (6.0)	14.9 (6.0)	0	14.5 (6.5)	14.5 (6.5)	0

	DQ2/2 [Manuscript]	DQ2/2 [DSIC]	DQ2/2 [Difference]	DQ8/X [Manuscript]	DQ8/X [DSIC]	DQ8/X [Difference]
Sex (male/female)	676/565	676/565	0	617/590	617/590	0
Age (months)	55.1 (16.9)	55.1 (16.9)	0	55.6 (17.0)	55.6 (17.0)	0
United States (n=2314)	556 (24.0%)	556 (24.0%)	0	362 (15.6%)	362 (15.6%)	0
Finland (n=1374)	205 (14.9%)	205 (14.9%)	0	482 (35.1%)	482 (35.1%)	0
Germany (n= 357)	71 (19.9%)	71 (19.9%)	0	81 (22.77%)	81 (22.7%)	0
Sweden (n=1924)	409 (21.3%)	409 (21.3%)	0	282 (14.6%)	282 (14.7%)	0(-0.1)
FDR with type 1 diabetes: Yes (n=675)	99 (14.7%)	99 (14.7%)	0	252 (37.3%)	252 (37.3%)	0
FDR with type 1 diabetes: No (n=5294)	1142 (21.6%)	1142 (21.6%)	0	955 (18.0%)	955 (18.0%)	0
Birth weight (g)	3500.6 (546.8)	3500.6 (546.8)	0	3506.5 (550.4)	3506.5 (550.4)	0
Pregnancy Diabetes Type 1(n=235)	26 (11.1%)	26 (11.1%)	0	97 (41.3%)	97 (41.3%)	0
Pregnancy Diabetes Type 2 (n= 16)	5 (31.3%)	5 (31.3%)	0	4 (25.0%)	4 (25.0%)	0
Pregnancy Diabetes GDM (n=330)	59 (17.9%)	59 (17.9%)	0	71 (21.5%)	71 (21.5%)	0
Pregnancy Diabetes None (n=5203)	1111(21.4%)	1111 (21.4%)	0	989 (19.0%)	989 (19.0%)	0
Prepregnancy BMI	25.2 (5.7)	25.2 (5.7)	0	24.8 (5.3)	24.8 (5.3)	0
Gestational age (weeks)	39.4 (1.7)	39.4 (1.7)	0	39.5 (1.7)	39.5 (1.7)	0
Gestational weight gain (kg)	14.9 (6.4)	14.9 (6.4)	0	14.1 (6.2)	14.1 (6.2)	0

Attachment A: SAS Code

```
*************************************************************************************************************************
***Program:
***Programmer: Michael Spriggs
***Date Created: 09/10/2015
***Purpose:
*************************************************************************************************************************;

title1 "%sysfunc(getoption(sysin))";
title2 " ";

libname sas_data "/prj/niddk/ims_analysis/TEDDY/private_orig_data/Pub35_JYang_niddk_submission/";
data pub35_jyang_niddk; set sas_data.pub35_jyang_niddk;

data pubshort(keep=maskid sex country fdr babysweightgrams diabetes BMI gestational_age weightgain dq current_age_mos);
       set pub35_jyang_niddk;

proc sort data=pubshort noduprecs;
       by maskid;

proc freq data=pubshort;
       tables (sex country fdr diabetes)*dq/missing;
       title3 'Categorical check';

proc means data=pubshort;
       var current_age_mos babysweightgrams BMI gestational_age weightgain;
       class DQ;
       title3 'Continuous check';

       title1 "%sysfunc(getoption(sysin))";
title2 " ";

%global caser;


*** Frequency Macro, N and % ***;
%macro freqdata2(order=, invar=, level=, roundvar=, digit=);

data data0 data1;
  set _null_;

  proc freq data=table1 noprint;
       tables &invar*&caser/out=data0 outpct;
       format _all_;
       run;

data data1;
  set data0;
  length LEVEL $100;
```

```
  LEVEL=strip(&invar);

  data data1(keep=LEVEL &caser name CHARALL ORDERER);
    set data1;
    length name $100 CHARALL $100;
    name=upcase("&invar");
    PCT_DISP=round(PCT_ROW,&roundvar.);
    CHARALL=compress(put(COUNT,8.))||" ("||compress(put(PCT_DISP,8.&digit))||'%)';
    ORDERER=&order;
    if level in &level then output data1;

data accumfreq1;
  set accumfreq1 data1;

%mend freqdata2;

%macro meandata1(order=, invar=, roundvar=, digit=);
proc means data=table1 mean stddev noprint;
       var &invar;
       class &caser;
       output out=data1 mean=mean stddev=stddev;
       run;

data data1(drop=_TYPE_ _FREQ_ mean stddev);
  set data1;
  length name CHARALL $100;
  name=upcase("&invar");
  mean=round(mean,&roundvar);
  stddev=round(stddev,&roundvar);
  /*CHARALL=compress(put(mean,8.&digit))||" ± "||compress(put(stddev,8.&digit));*/
  CHARALL=compress(put(mean,8.&digit))||" ("||compress(put(stddev,8.&digit))||")";
  ORDERER=&order;

data accummean1;
  set accummean1 data1;

%mend meandata1;

%macro inertdata1(order=);

data inert1;
  length orderer 8. &caser $13.;
  orderer=&order.;
  output;

data accuminert1;
  set accuminert1 inert1;

%mend inertdata1;
```

```
%let caser=dq;

*** Column processing: Baseline;
data accumfreq1 accummean1 accummedian1 accuminert1;
  set _null_;

data table1;
  set pubshort;
  if diabetes=1 then diab_char='gest';
  if diabetes=2 then diab_char='typ1';
  if diabetes=3 then diab_char='typ2';
  if diabetes=4 then diab_char='none';

%freqdata2(order=1 , invar=SEX               , level=("Male") , roundvar=1, digit=0);
%freqdata2(order=1.5 , invar=SEX             , level=("Female") , roundvar=1, digit=0);
%meandata1(order=2 , invar=current_age_mos   , roundvar=.1, digit=1);
%freqdata2(order=3 , invar=Country           , level=("1") , roundvar=.1, digit=1);
%freqdata2(order=4 , invar=Country           , level=("2") , roundvar=.1, digit=1);
%freqdata2(order=5 , invar=Country           , level=("3") , roundvar=.1, digit=1);
%freqdata2(order=6 , invar=Country           , level=("4") , roundvar=.1, digit=1);
%freqdata2(order=7 , invar=FDR               , level=("1") , roundvar=.1, digit=1);
%freqdata2(order=8 , invar=FDR               , level=("0") , roundvar=.1, digit=1);
%meandata1(order=9 , invar=babysweightgrams  , roundvar=.1, digit=1);
%freqdata2(order=10 , invar=diab_char        , level=("typ1") , roundvar=.1, digit=1);
%freqdata2(order=11 , invar=diab_char        , level=("typ2") , roundvar=.1, digit=1);
%freqdata2(order=12 , invar=diab_char        , level=("gest") , roundvar=.1, digit=1);
%freqdata2(order=13 , invar=diab_char        , level=("none") , roundvar=.1, digit=1);
%meandata1(order=14 , invar=BMI              , roundvar=.1, digit=1);
%meandata1(order=15 , invar=gestational_age  , roundvar=.1, digit=1);
%meandata1(order=16 , invar=weightgain       , roundvar=.1, digit=1);

data accumtab1;
  set accumfreq1 accummean1 accummedian1 accuminert1;
  if &caser=" " then delete;

*** Display processing ***;

proc sort data=accumtab1;
  by dq orderer;

proc print data=accumtab1 noobs;
  var name charall orderer;
  by dq;
  pageby dq;
  title3 'Table 1';
```