Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) Pub31 SHummel

Prepared by Jane Wang
IMS Inc.
3901 Calverton Blvd, Suite 200 Calverton MD 20705
September 11, 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 “Pub31_SHummel_niddk_submission” dataset.

4 Statistical Methods

Analyses were performed to duplicate results for the data published by Sandra Hummel et al [1]. Public Health Nutrition, doi:10.1017/S1368980013003054. To verify the integrity of the dataset, descriptive statistics were computed, by presence of diabetes in the family.

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 the similar to the published results except birth weight. Education has differences as well which might be a typo.

6 Conclusions

The NIDDK repository is confident that the TEDDY Pub31 SHummel data files to be distributed are a true copy of the study data.

7 References

Sandra Hummel, Kendra Vehik, Ulla Uusitalo, Wendy McLeod, Carin Andre´n Aronsson, Nicole Frank, Patricia Gesualdo, Jimin Yang, Jill M Norris, Suvi M Virtanen, and the TEDDY Study Group. Infant feeding patterns in families with a diabetes history – observations from The Environmental Determinants of Diabetes in the Young (TEDDY) birth cohort study. Public Health Nutrition, doi:10.1017/S1368980013003054.

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

Table Variable	Variables Used in Replication from the "Table 1"
Females:males	sex
5 min Apgar score $9 (%)	apgar
Normal vaginal delivery (%)	delivery
first-born child (%)	firstborn
Gestational age (weeks)-	gestational_age
birth weight (g)-	babysweightgrams
Maternal BMI before pregnancy (kg/m2)-	bmi
Maternal weight gain (kg)-	weightgain
Maternal age (years)-	maternal_age
Maternal education (%)	education_mom_group1
Mother smoking during pregnancy (%)	smoker

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

	Mother with T1D N [Manuscript]	Mother with T1D N [DSIC]	Mother with T1D N [Difference]	Mother with T1D %/Median [Manuscript]	Mother with T1D %/Median [DSIC]	Mother with T1D %/Median [Difference]	Mother with T1D IQR [Manuscript]	Mother with T1D IQR [DSIC]	Mother with T1D IQR [Difference]
Females:males	292	292	0				151:141	151:141	0:0
5 min Apgar score $9 (%)	214	214	0	75	75	0			
Normal vaginal delivery (%)	281	281	0	33	33	0			
first-born child (%)	281	281	0	52	52	0			
Gestational age (weeks)-	291	291	0	38	38	0	37–39	37.0 - 39.0	0-0
birth weight (g)-	286	286	0	3716	3716	0	3300–4008	3300 - 4008	0-0
Maternal BMI before pregnancy (kg/m2)-	291	291	0	24.1	24.1	0	22.3–27.3	22.3 - 27.3	0-0
Maternal weight gain (kg)-	289	289	0	15.3	15.3	0	11.5–19.1	11.5 - 19.1	0-0
Maternal age (years)-	292	292	0	30	30	0	27.0–34.0	27.0 - 34.0	0-0
Maternal education (%)	278	278	0						
High school or less				15	15	0			
More than high school				85	85	0			
Mother smoking during pregnancy (%)	289	289	0	14	14	0			

	Mother with GDM N [Manuscript]	Mother with GDM N [DSIC]	Mother with GDM N [Difference]	Mother with GDM %/Median [Manuscript]	Mother with GDM %/Median [DSIC]	Mother with GDM %/Median [Difference]	Mother with GDM IQR [Manuscript]	Mother with GDM IQR [DSIC]	Mother with GDM IQR [Difference]
Females:males	404	404	0				196:208	196:208	0:0
5 min Apgar score $9 (%)	274	274	0	87	86	1			
Normal vaginal delivery (%)	383	383	0	68	68	0			
first-born child (%)	395	395	0	38	38	0			
Gestational age (weeks)-	403	403	0	40	40	0	38–40	38.0 - 40.0	0-0
birth weight (g)-	393	393	0	3560	3560	0	3220–3894	3220 - 3894	0-0
Maternal BMI before pregnancy (kg/m2)-	398	398	0	26.9	26.9	0	23.6–32.3	23.6 - 32.3	0-0
Maternal weight gain (kg)-	375	375	0	11	11	0	8.0–15.5	8.0 - 15.5	0-0
Maternal age (years)-	404	404	0	32	32		28–36	28.0 - 36.0	0-0
Maternal education (%)	394	394	0						
High school or less				5	15	-10			
More than high school				85	85	0			
Mother smoking during pregnancy (%)	399	399	0	13	13	0			

	father with T1D N [Manuscript]	father with T1D N [DSIC]	father with T1D N [Difference]	father with T1D %/Median [Manuscript]	father with T1D %/Median [DSIC]	father with T1D %/Median [Difference]	father with T1D IQR [Manuscript]	father with T1D IQR [DSIC]	father with T1D IQR [Difference]
Females:males	464	464	0				231:233	231:233	0:0
5 min Apgar score $9 (%)	291	291	0	88	88	0			
Normal vaginal delivery (%)	453	453	0	74	74	0			
first-born child (%)	454	454	0	35	35	0			
Gestational age (weeks)-	464	464	0	40	40	0	39–40	39.0 - 40.0	0-0
birth weight (g)-	449	449	0	3496	3496	0	3160–3850	3160.0 - 3825.0	0-25
Maternal BMI before pregnancy (kg/m2)-	459	459	0	23.1	23.1	0	21.2–26.2	21.1 - 26.2	0-0
Maternal weight gain (kg)-	451	451	0	14.5	14.5	0	11.0–18.0	11.0 - 18.0	0-0
Maternal age (years)-	464	464	0	32	32	0	28.0–35.0	28.0 - 35.0	0-0
Maternal education (%)	457	457	0						
High school or less				18	18	0			
More than high school				82	82	0			
Mother smoking during pregnancy (%)	459	459	0	9	9	0			

	No Diabetes family history N [Manuscript]	No Diabetes family history N [DSIC]	No Diabetes family history N [Difference]	No Diabetes family history %/Median [Manuscript]	No Diabetes family history %/Median [DSIC]	No Diabetes family history %/Median [Difference]	No Diabetes family history IQR [Manuscript]	No Diabetes family history IQR [DSIC]	No Diabetes family history IQR [Difference]
Females:males	5866	5866	0				2836:3030	2836:3030	0:0
5 min Apgar score $9 (%)	3592	3592	0	89	89	0			
Normal vaginal delivery (%)	5659	5659	0	73	73	0			
first-born child (%)	5722	5722	0	46	46	0			0-0
Gestational age (weeks)-	5854	5854	0	40	40	0	39–40	39.0 - 40.3	0-0
birth weight (g)-	4478	5715	-1227	3530	3525	5	3210–3825	3205 -3850	5 - (-25)
Maternal BMI before pregnancy (kg/m2)-	5798	5798	0	23.3	23.3	0	21.1–26.6	21.1 - 26.6	0-0
Maternal weight gain (kg)-	5704	5704	0	14.5	14.5	0	11.4–18.2	11.4 - 18.2	0-0
Maternal age (years)-	5866	5866	0	30	30	0	27–34	27.0 - 34.0	0-0
Maternal education (%)	5708	5708	0						
High school or less				20	20	0			
More than high school				80	80	0			
Mother smoking during pregnancy (%)	5804	5804	0	11	11	0			

Attachment A: SAS Code

```
*************************************************************************************************************************
***Program:
***Programmer: Jane Wang
***Date Created: 08/19/2015
***Purpose:
*************************************************************************************************************************;

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

options nofmterr;
options nofmterr;
libname sas_data "/prj/niddk/ims_analysis/TEDDY/private_orig_data/Pub31_SHummel_niddk_submission/";
data Pub31_shummel_niddk              ; set sas_data.Pub31_shummel_niddk     ;

%macro baseline_freq1(dataset_name,var_name,);

       *** Creating a frequency table in the format of Table 1 in the primary diabexp paper;

  proc freq data = &dataset_name ;
       table (&var_name.)*diabexp ;
       title3 "Frequency table of the &var_name. variable in the analysis dataset";

       *** Outputting the frequency data to work.&var_name._cross using the ODS output;
  ods output CrossTabFreqs = work.&var_name._cross;
  proc print data = &var_name._cross;

  data &var_name._cross1(keep = diabexp     Frequency);
    set &var_name._cross;
    if diabexp ne . and &var_name = .;


  data &var_name._cross2(keep = diabexp     colPercent);
    set &var_name._cross;
    if "&var_name" = 'delivery' then do;
      if diabexp ne . and &var_name = 3;
    end;
    else if diabexp ne . and &var_name = 1;


data &var_name._cross;
  merge &var_name._cross1 &var_name._cross2;
  by diabexp;

  proc print data = &var_name._cross;
```

```
  data &var_name._cross;
    set &var_name._cross;
    length table_name $30.;
    table_name  = "&var_name" ;

    proc sort data = &var_name._cross;
     by  table_name  diabexp;


  data &var_name._cross_1(drop = diabexp Colpercent   Frequency i);
    set &var_name._cross;
    by table_name diabexp ;
    array temp1(4) countno countmt1d countgdm countft1d ;
    array temp2(4) pertno pertmt1d pertgdm pertft1d ;
    retain countno countmt1d countgdm countft1d  pertno pertmt1d pertgdm pertft1d ;
    if first.table_name then do i = 1 to 4;
      temp1(i) = .;
      temp2(i) = .;
    end;
      temp1(_n_) = Frequency;
      temp2(_n_) = round(Colpercent,1);
    if last.table_name;


%mend;

%macro baseline_means(dataset_name,var_name);

proc sort data = Pub31_shummel_niddk;
  by diabexp;
       *** Creating a frequency table in the format of Table 1 in the primary outcome paper;
  proc means data = &dataset_name n median p25 p75;
       var &var_name.;
       by diabexp;
       title3 "Frequency table of the &var_name. variable in the analysis dataset";

       *** Outputting the frequency data to work.&var_name._cross using the ODS output;
  ods output Summary = work.&var_name._means1;
  run;

  data &var_name._means1;
    set &var_name._means1;
    length table_name $30.;
    table_name  ="&var_name";
    if diabexp ne .;


   proc sort data = &var_name._means1;
```

```
     by table_name diabexp;
  data &var_name._means_1(drop = diabexp &var_name._Median &var_name._P25  &var_name._P75 i &var_name._n) ;
    set &var_name._means1;
    by table_name;
*    array temp1(4) countno countmt1d countgdm countft1d ;

    array temp1(4)  n_no n_mt1d n_gdm n_ft1d ;
    array temp2(4)  median_no median_mt1d median_gdm median_ft1d ;
    array temp3(4)  p25_no p25_mt1d p25_gdm p25_ft1d  ;
    array temp4(4)  p75_no p75_mt1d p75_gdm p75_ft1d  ;
    retain n_no n_mt1d n_gdm n_ft1d median_no median_mt1d median_gdm median_ft1d p25_no p25_mt1d p25_gdm p25_ft1d p75_no p75_mt1d
p75_gdm p75_ft1d ;
    if first.table_name then do i = 1 to 4;
      temp1(i) = .;
      temp2(i) = .;
      temp3(i) = .;
      temp4(i) = .;
    end;
      temp1(_n_) = round(&var_name._n,1);
      temp2(_n_) = round(&var_name._Median,0.1);
      temp3(_n_) = round(&var_name._P25,0.1);
      temp4(_n_) = round(&var_name._P75,0.1);
    if last.table_name;

%mend;

*******************************************************************************;
*************** Check Table 1 ************************************************;
*******************************************************************************;

*** Running the baseline_freq on the  categorical variables in the Table 1 manuscript file;
%baseline_means(Pub31_shummel_niddk, gestational_age      );
%baseline_means(Pub31_shummel_niddk, babysweightgrams    );
%baseline_means(Pub31_shummel_niddk, bmi                   );
%baseline_means(Pub31_shummel_niddk, weightgain           );
%baseline_means(Pub31_shummel_niddk, maternal_age           );


data data table1_mean;
  set
gestational_age_means_1
babysweightgrams_means_1
bmi_means_1
weightgain_means_1
maternal_age_means_1
;
```

```
%baseline_freq1(Pub31_shummel_niddk,apgar               );
%baseline_freq1(Pub31_shummel_niddk,delivery            );
%baseline_freq1(Pub31_shummel_niddk,firstborn            );
%baseline_freq1(Pub31_shummel_niddk,education_mom_group1 );
%baseline_freq1(Pub31_shummel_niddk,smoker                );

data table1_freq;
  set
apgar_cross_1
delivery_cross_1
firstborn_cross_1
education_mom_group1_cross_1
smoker_cross_1
;


proc print data = table1_freq;
proc print data = table1_mean;

proc freq data = Pub31_shummel_niddk;
  tables diabexp *  sex;
```