Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) Pub100 Beyerlein

Prepared by Allyson Mateja
IMS Inc.
3901 Calverton Blvd, Suite 200 Calverton, MD 20705
June 8, 2017

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 the "m_100_abeyerlein_niddk_1_new.sas7bdat" and "m_100_abeyerlein_niddk_2_new.sas7bdat" datasets.

4 Statistical Methods

Analyses were performed to duplicate results for the data published by Beyerlein et al [1] in The American Journal of Clinical Nutrition in 2015. To verify the integrity of the dataset, descriptive statistics were computed.

5 Results

For Table 1 in the publication [1], Characteristics of the data analyzed, 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 in Table 1. The results of the replication are an exact match to the published results.

For Figure 2 in the publication [1], Boxplots of soluble fiber intake, total energy intake from food items, and soluble fiber intake standardized to an energy intake of 1000 kcal by diet record visit (n=3358 subjects), Table C lists the variables that were used in the replication and Figure A compares the results calculated from the archived data files to the results published in Figure 2. The results of the replication are almost an exact match to the published results.

6 Conclusions

The NIDDK repository is confident that the TEDDY M100 data files to be distributed are a true copy of the study data.

7 References

[1] Beyerlein, A., Liu, X., Uusitalo, U.M., Harsunen, M., Norris, J.M., Foterek, K., Virtanen, S.M., Rewers, M.J., She, J., Simell, O., Lernmakr, A., Haopian, W., Akolkar, B., Ziegler, A., Krischer, J.P., Hummel, S., and the TEDDY study group. "Dietary intake of soluble fiber and risk of islet autoimmunity by 5 y of age: results from the TEDDY study". The American Journal of Clinical Nutrition (2015) 102:345-352.

Table A: Variables used to replicate Table 1**:** Characteristics of the data analyzed

Table Variable	**dataset.variable**
Duration of follow-up	m_100_abeyerlein_niddk_1_new.last_clinic_visit
Maternal prepregnancy BMI	m_100_abeyerlein_niddk_1_new.bmi
Developed any islet autoantibodies	m_100_abeyerlein_niddk_1_new.persist_conf_ab
Developed multiple islet autoantibodies	m_100_abeyerlein_niddk_1_new.mult_ab
Developed T1D	m_100_abeyerlein_niddk_1_new.t1d
Male child	m_100_abeyerlein_niddk_1_new.female
HLA-DR3/DR4 genotype	m_100_abeyerlein_niddk_1_new.dr34
Having a first-degree relative with T1D	m_100_abeyerlein_niddk_1_new.fdr
Maternal T1D	m_100_abeyerlein_niddk_1_new.mother_fdr
First child in the family	m_100_abeyerlein_niddk_1_new.mom_first_child
Born by cesarean delivery	m_100_abeyerlein_niddk_1_new.csection
Maternal smoking in pregnancy	m_100_abeyerlein_niddk_1_new.smoker
Maternal education less than high school	m_100_abeyerlein_niddk_1_new.mom_eduction
Child was never breastfed	m_100_abeyerlein_niddk_1_new.ever_brstfed
Country	m_100_abeyerlein_niddk_1_new.country

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

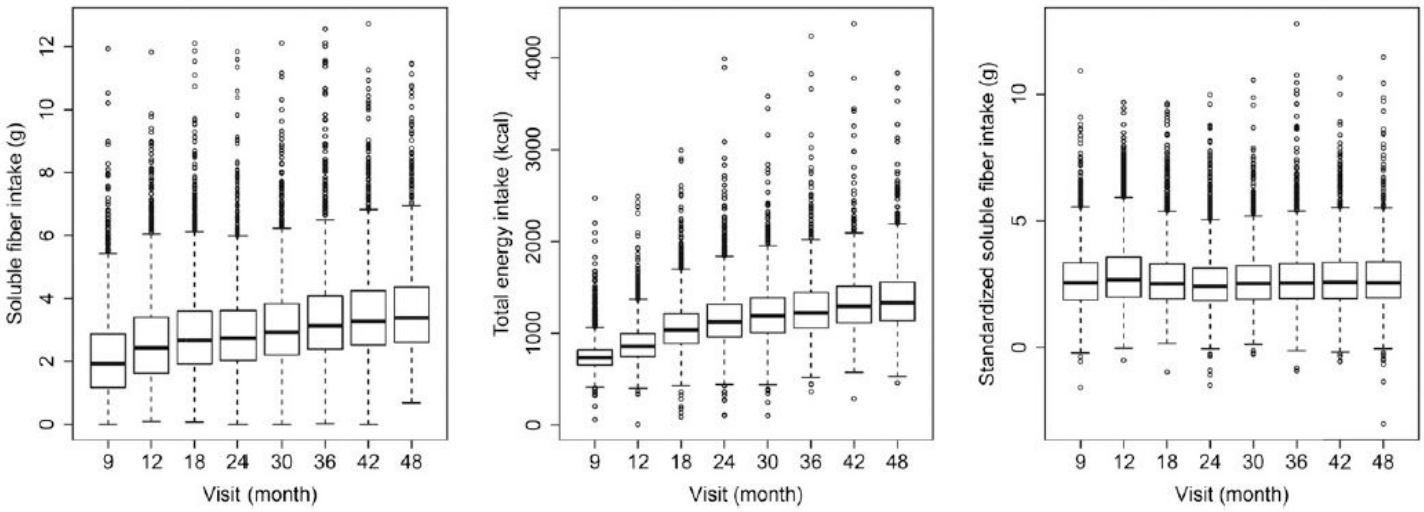
Variable	United States Manuscript (n=2912)	United States DSIC (n=2912)	Diff. (n=0)	Germany Manuscript (n=446)	Germany DSIC (n=446)	Diff. (n=0)
Duration of follow-up, y	5.0 (4.0-6.5)	5.0 (4.0-6.5)	0 (0-0)	5.0 (3.5-7.0)	5.0 (3.5-7.0)	0 (0-0)
Maternal prepregnancy BMI, kg/m²	24.0 (21.4-28.3)	24.0 (21.4-28.3)	0 (0-0)	23.1 (20.8-26.4)	23.1 (20.8-26.4)	0 (0-0)
Developed any islet autoantibodies, n (%)	198 (6.8)	198 (6.8)	0 (0)	44 (9.9)	44 (9.9)	0 (0)
Developed multiple islet autoantibodies, n (%)	118 (4.1)	118 (4.1)	0 (0)	33 (7.4)	33 (7.4)	0 (0)
Developed T1D, n (%)	52 (1.8)	52 (1.8)	0 (0)	19 (4.3)	19 (4.3)	0 (0)
Male child, n (%)	1418 (48.7)	1418 (48.7)	0 (0)	214 (48.0)	214 (48.0)	0 (0)
HLA-DR3/DR4 genotype, n (%)	1184 (40.7)	1184 (40.7)	0 (0)	169 (37.9)	169 (37.9)	0 (0)
Having a first-degree relative with T1D, n (%)	326 (11.2)	326 (11.2)	0 (0)	171 (38.3)	171 (38.3)	0 (0)
Maternal T1D, n (%)	101 (3.5)	101 (3.5)	0 (0)	81 (18.2)	81 (18.2)	0 (0)
First child in the family, n (%)	1202 (42.0)	1202 (42.0)	0 (0)	219 (50.9)	219 (50.9)	0 (0)
Born by cesarean delivery, n (%)	1072 (36.8)	1072 (36.8)	0 (0)	158 (35.4)	158 (35.4)	0 (0)
Maternal smoking in pregnancy, n (%)	256 (8.9)	256 (8.9)	0 (0)	69 (15.5)	69 (15.5)	0 (0)
Maternal education less than high school, n (%)	391 (13.6)	391 (13.6)	0 (0)	45 (10.5)	45 (10.5)	0 (0)
Child was never breastfed, n (%)	155 (5.3)	155 (5.3)	0 (0)	12 (2.7)	12 (2.7)	0 (0)

Table C: Variables used to replicate Figure 2**:** Boxplots of soluble fiber intake, total energy intake from food items, and soluble fiber intake standardized to an energy intake of 1000 kcal by diet record visit (n=3358 subjects)

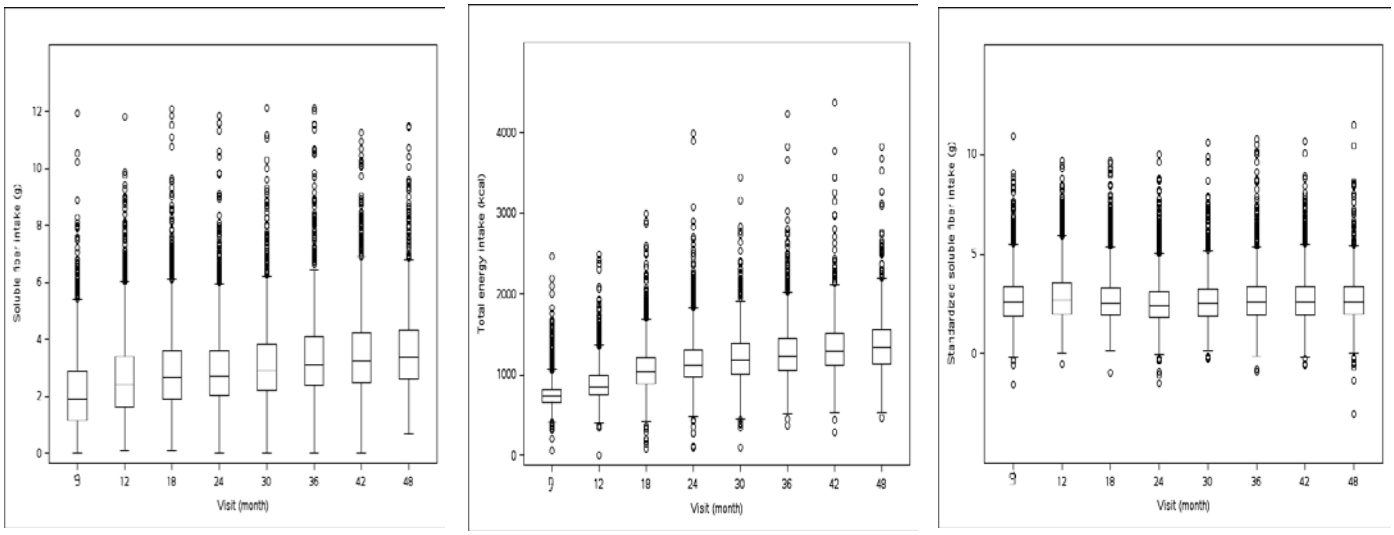
Table Variable	**dataset.variable**
Soluble fiber intake	m_100_abeyerlein_niddk_2_new.sol_fiber
Total energy intake	m_100_abeyerlein_niddk_2_new.tot_ene
Standardized soluble fiber intake	m_100_abeyerlein_niddk_2_new.res_solf_timedep
Visit	m_100_abeyerlein_niddk_2_new.intake_age

Figure A: Comparison of values computed in integrity check to reference article Figure 2 values

Manuscript



DSIC



Attachment A: SAS Code

```
*** TEDDY M100 DSIC;
*** Programmer: Allyson Mateja;
*** Date: 4/3/17;

proc format;
       value $countryf 'US' = 'United States'
                       'Ge' = 'Germany';
       value yesnof 0 = 'No'
                    1 = 'Yes';

libname sas_data '/prj/niddk/ims_analysis/TEDDY/private_orig_data/M_100_ABeyerlein_NIDDK_Submission/';

data table1;
       set sas_data.m_100_abeyerlein_niddk_1_new;

data table1;
       set table1;
       followup_time = last_clinic_visit/12;

data table2;
       set sas_data.m_100_abeyerlein_niddk_2_new;

proc contents data = table1;
proc contents data = table2;

proc freq data = table1;
       tables country;
       format country $countryf.;
       title 'Table 1 - Country';

proc sort data = table1;
       by country;

proc means data = table1 n median p25 p75;
       var followup_time;
       class country;
       format country $countryf.;
       title 'Table 1 - Duration of follow-up';

proc means data = table1 n median p25 p75;
       var bmi;
       class country;
       format country $countryf.;
       title 'Table 1 - Maternal prepregnancy BMI';

proc freq data = table1;
       tables persist_conf_ab;
```

```
        by country;
        format country $countryf. persist_conf_ab yesnof.;
        title 'Table 1 - Developed any islet autoantibodies';

proc freq data = table1;
        tables mult_ab;
        by country;
        format country $countryf. mult_ab yesnof.;
        title 'Table 1 - Developed multiple islet autoantibodies';

proc freq data = table1;
        tables t1d;
        by country;
        format country $countryf. t1d yesnof.;
        title 'Table 1 - Developed T1D';

proc freq data = table1;
        tables female;
        by country;
        format country $countryf.;
        title 'Table 1 - Male child';

proc freq data = table1;
        tables dr34;
        by country;
        format country $countryf. dr34 yesnof.;
        title 'Table 1 - HLA-DR3/DR4 genotype';

proc freq data = table1;
        tables fdr;
        by country;
        format country $countryf.;
        title 'Table 1 - Having a first degree relative with T1D';

proc freq data = table1;
        tables mother_fdr;
        by country;
        format country $countryf. mother_fdr yesnof.;
        title 'Table 1 - Maternal T1D';

proc freq data = table1;
        tables mom_first_child;
        by country;
        format country $countryf. mom_first_child yesnof.;
        title 'Table 1 - First child in family';

proc freq data = table1;
        tables csection;
        by country;
        format country $countryf.;
        title 'Table 1 - Born by Cesarean delivery';
```

```
proc freq data = table1;
       tables smoker;
       by country;
       format country $countryf.;
       title 'Table 1 - Maternal smoking in pregnancy';

proc freq data = table1;
       tables mom_education;
       by country;
       format country $countryf.;
       title 'Table 1 - Maternal education less than high school';

proc freq data = table1;
       tables ever_brstfed;
       by country;
       format country $countryf. ever_brstfed yesnof.;
       title 'Table 1 - Child was never breastfed';

data table2;
       set table2;
       age_mos = round((intake_age/365.25)*12, 3);

proc sort data = table2;
       by age_mos;

data table2;
       set table2;
       if age_mos = 9 then age_mos = 6;
       if age_mos in (6,12,18,24,30,36,42,48);

ods graphics on;
title ' ';

proc boxplot data=table2;
       plot sol_fiber*age_mos /boxstyle=schematic vaxis = (0,2,4,6,8,10,12,14) ;
       label age_mos = 'Visit (month)'
             sol_fiber = 'Soluble fiber intake (g)';

proc boxplot data=table2;
       plot tot_ene*age_mos /boxstyle=schematic vaxis = (0,1000,2000,3000,4000,5000) ;
       label age_mos = 'Visit (month)'
             tot_ene = 'Total energy intake (kcal)';

proc boxplot data=table2;
       plot res_solf_timedep*age_mos /boxstyle=schematic vaxis = (-5,0,5,10,15) ;
       label age_mos = 'Visit (month)'
             res_solf_timedep = 'Standardized soluble fiber intake (g)';
```