Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) M111 Beyerlein

Prepared by Sabrina Chen
IMS Inc.
3901 Calverton Blvd, Suite 200 Calverton, MD 20705
March 10, 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_111_ABeyerlein_NIDDK_Submission folder in the data package. For this replication, variables were taken from the “m_111_abeyerlein_niddk_18oct2017.sas7bdat” dataset.

4 Statistical Methods

Analyses were performed to duplicate results for the data published by Andreas Beyerlein et al [1] in Pediatric Obesity in 2017. To verify the integrity of the dataset, descriptive statistics were computed.

5 Results

For Extended Data Table 1 in the publication [1], **Description of the study population**, 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.

6 Conclusions

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

7 References

[1] Andreas Beyerlein, Ulla M. Uusitalo, Suvi M. Virtanen, Kendra Vehik, Jimin Yang, Christiane Winkler, Mathilde Kersting, Sibylle Koletzko, Desmond Schatz, Carin Andren Aronsson, Helena Elding Larsson, Jeffrey P. Krischer, Anette-G. Ziegler, Jill M. Norris, and Sandra Hummel. Intake of Energy and Protein is Associated with Overweight Risk at Age 5.5 Years: Results from the Prospective TEDDY Study. Obesity (Silver Spring). 2017 Aug; 25(8): 1435–1441.

Table A: Variables used to replicate Extended Data Table 1: Description of the study population

Table Variable	**dataset.variable**
BMI SDS at the age of 5.5 y	m_111_abeyerlein_niddk_18oct2017.z_bmi
Has overweight at the age of 5.5 y	m_111_abeyerlein_niddk_18oct2017.Babysweightgrams
Has obesity at the age of 5.5 y	m_111_abeyerlein_niddk_18oct2017.maternal_age
Males	m_111_abeyerlein_niddk_18oct2017.bmi_mom
Country	m_111_abeyerlein_niddk_18oct2017.overweight
Birth weight (g)	m_111_abeyerlein_niddk_18oct2017.obese
Maternal age at birth of child (y)	m_111_abeyerlein_niddk_18oct2017.sex
Maternal prepregnancy BMI (kg/m2)	m_111_abeyerlein_niddk_18oct2017.country
High maternal education (high school)	m_111_abeyerlein_niddk_18oct2017.mom_education
Breastfeeding 6 months	m_111_abeyerlein_niddk_18oct2017.brstfed_6mo
Excessive total gestational weight gain (according to Institute of Medicine guidelines) (21)	m_111_abeyerlein_niddk_18oct2017.gwg_iom
Maternal diabetes (yes)	m_111_abeyerlein_niddk_18oct2017.diabetes
Maternal smoking during pregnancy (yes)	m_111_abeyerlein_niddk_18oct2017.smoker
Maternal alcohol intake during pregnancy (yes)	m_111_abeyerlein_niddk_18oct2017.drinker

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

Variable	Manu script	DSIC	Diff.	Manuscri pt	DSIC	Diff.	Manuscri pt	DSIC	Diff.
Median/Q1/Q3									
BMI SDS at the age of 5.5 y	0.28	0.28	0	-0.31	-0.31	0	0.92	0.92	0
Birth weight (g)	3,525	3524.68	0.32	3175	3175	0	3860	3860	0
Maternal age at birth of child (y)	31	31	0	28	28	0	34	34	0
Maternal prepregnancy BMI (kg/m2)	23.5	23.51	-0.01	21.3	21.3	0	27	26.99	0.01
	Manu script	DSIC	Diff.	Manuscri pt	DSIC	Diff.	Manuscri pt	DSIC	Diff.
n/percent									
Has overweight at the age of 5.5 y	1,253	1253	0	22.5	22.5	0			
Has obesity at the age of 5.5 y	337	337	0	6.1	6.1	0			
Males	2,862	2862	0	51.5	51.4	0.1			
Country									
United States	2,104	2104	0	37.8	37.8	0			
Finland	1,301	1301	0	23.4	23.4	0			
Germany	307	307	0	5.5	5.5	0			
Sweden	1,851	1851	0	33.3	33.3	0			

Variable	Manu script	DSIC	Diff.	Manuscri pt	DSIC	Diff.	Manuscri pt	DSIC	Diff.
High maternal education (high school)	4,532	4532	0	82.8	81.5	1.3			
Breastfeeding >= 6 months	3,603	3603	0	64.8	64.8	0			
Excessive total gestational weight gain (according to Institute of Medicine guidelines) (21)	2,519	2518	1	46.3	45.3	1			
Maternal diabetes (yes)	571	571	0	10.6	10.3	0.3			
Maternal smoking during pregnancy (yes)	541	541	0	9.8	9.7	0.1			
Maternal alcohol intake during pregnancy (yes)	1,911	1911	0	34.6	34.4	0.2			

Attachment A: SAS Code

```
title 'prj/niddk/ims_analysis/TEDDY/prog_initial_analysis/m_111_dsic.sas';
run;

***********;
* INPUT    ;
***********;

libname sasfile2 '/prj/niddk/ims_analysis/TEDDY/private_orig_data/M_111_ABeyerlein_NIDDK_Submission/';


************;
* MACROS    ;
************;
%macro readin(lib, ds);
  data &ds;
    set sasfile&lib..&ds;
  run;

  proc contents data=&ds;
  title3 "&ds";
  run;
%mend;


* 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;
  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, subsetname);

  proc univariate data=analy outtable= univ&subsetname noprint;
    where &subset=1;
    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;



*************;
* FORMATS    ;
*************;
proc format;
  value novalue
      . = "No Value"
  other = "   Value"
  ;

  value negpos
  0 = "Negative"
  1 = "Positive"
  ;

  value yesno
  0 = "No"
```

```
  1 = "Yes"
  ;

  value sexf
  0 = "Male"
  1 = "Female"
  ;

  value educf
  0="High school or less"
  1="More than high school"
  ;

  value diabf
  - 1 = "Gestational diabetes only"
    2 = "Type 1 diabetes only"
    3 = "Type 2 diabetes only"
    4 = "Did not have diabetes of any kind"
    ;

  value gwgf
  -1 = "Inadequate"
   0 = "Adequate"
   1 = "Excessive"
   ;

  value country
  1="US"
  2="FIN"
  3="GER"
  4="SWE"
  ;

run;


%readin(2, m_111_abeyerlein_niddk_18oct2017);


proc freq data=m_111_abeyerlein_niddk_18oct2017;
  tables dropout_mo66*bmi_mo66_available /list missing;
run;

proc freq data=m_111_abeyerlein_niddk_18oct2017;
  where bmi_mo66_available=1;
```

```
  tables overweight obese sex country mom_education brstfed_6mo gwg_iom diabetes smoker drinker/missing;
run;

proc means data=m_111_abeyerlein_niddk_18oct2017;
  where bmi_mo66_available=1;
  var z_bmi wtkg maternal_age bmi_mom ;
run;

data analy;
  set m_111_abeyerlein_niddk_18oct2017;

  if bmi_mo66_available=1;

  * create subset flag for each row to use in macro call;
  all = 1;

  if SEX = "Male" then sexn = 0;
  else if sex = "Female" then sexn=1;

  if diabetes in(1,2,3) then diabetes_yes = 1;
  else if diabetes = 4 then diabetes_yes = 0;

run;

proc freq data=analy;
  tables sex*sexn/list missing;
  tables diabetes*diabetes_yes/list missing;
run;

* med, min and max;
data prntunivall;
  length _VAR_ $100;
  set _null_;
run;

%univ(1   , z_bmi            , all , all);
%univ(6   , Babysweightgrams , all , all);
%univ(7   , maternal_age     , all , all);
%univ(8   , bmi_mom          , all , all);

data alluniv;
  set prntunivall (in=in1 keep = rownum _var_  _median_ _q1_ _q3_);
  _median_   = round(_median_  , 0.1);
  _q1_       = round(_q1_      , 0.1);
  _q3_       = round(_q3_      , 0.1);
```

```
run;

proc print data= prntunivall noobs;
    var rownum _var_  covarf /*_nobs_*/ _median_ _q1_ _q3_ /*_min_ _max_ _std_*/;
run;


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

%npercent(2, overweight      , yesno  , all , all);
%npercent(3, obese           , yesno  , all , all);
%npercent(4, sexn            , sexf   , all , all);
%npercent(5, country         , country       , all , all);
%npercent(9, mom_education   , yesno  , all , all);
%npercent(10,brstfed_6mo     , yesno  , all , all);
%npercent(11,gwg_iom         , gwgf   , all , all);
%npercent(12,diabetes_yes    , yesno  , all , all);
%npercent(13,smoker          , yesno  , all , all);
%npercent(14,drinker         , yesno  , all , all);

proc print data=prntall;
  var rownum covar covarf count    percent;
title3 "all";
run;


* Table 1;
data table1mod;
  set prntunivall prntall;
  percent = round(percent, .1);
run;

proc sort data=table1mod;
  by rownum covarf;
run;

proc print data=table1mod;
  var rownum covar covarf _var_ count percent  _median_ _q1_ _q3_;
  title3 "Table 1";
run;
```