Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) M74 Koketzko

Prepared by Sabrina Chen
IMS Inc.
3901 Calverton Blvd, Suite 200 Calverton, MD 20705
February 28, 2019

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_74_skoletzko_niddk_30apr2015.sas7bdat" dataset.

4 Statistical Methods

Analyses were performed to duplicate results for the data published by Koketzko et al [1] in *Journal of Pediatric Gastroenterology and Nutrition* in 2017. To verify the integrity of the dataset, descriptive statistics were computed.

5 Results

For Table 1 in the publication [1], Maternal and child characteristics in relation to mode of delivery. For continuous variables median (25th percentile, 75th percentile) is reported, 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 similar to the published results.

6 Conclusions

The NIDDK repository is confident that the TEDDY M74 data files to be distributed are a true copy of the study data.

7 References

[1] Sibylle Koletzko, MD, PhD, Hye-Seung Lee, Andreas Beyerlein, PhD, Carin A. Aronsson, MSc, Michael Hummel, MD, PhD, Edwin Liu, MD, Ville Simell, MSc, Kalle Kurppa, MD, PhD, Åke Lernmark, PhD, William Hagopian, MD, PhD, Marian Rewers, MD, PhD, Jin-Xiong She, PhD, Olli Simell, MD, PhD, Jorma Toppari, MD, PhD, Anette-G. Ziegler, MD, Jeffrey Krischer, PhD, Daniel Agardh, MD, PhD, for the TEDDY Study Group. "CAESAREAN SECTION ON THE RISK OF CELIAC DISEASE IN THE OFFSPRING: THE TEDDY STUDY". Journal of Pediatric Gastroenterology and Nutrition

Table A: Variables used to replicate Table 1: Maternal and child characteristics in relation to mode of delivery. For continuous variables, median (25th percentile, 75th percentile) is reported

Table Variable	**dataset.variable**
Age, years	m_74_skoletzko_niddk_30apr2015.maternal_age
Education	m_74_skoletzko_niddk_30apr2015.mom_education
Smoking during pregnancy	m_74_skoletzko_niddk_30apr2015.rsmoker
Antibiotic use during pregnancy	m_74_skoletzko_niddk_30apr2015.momantibiotic
Pre-pregnancy BMI	m_74_skoletzko_niddk_30apr2015.mdiab
Gestational weight gain, kg	m_74_skoletzko_niddk_30apr2015.mombmi
Any diabetes during pregnancy	m_74_skoletzko_niddk_30apr2015.prgwtgain
Gestational age (weeks)	m_74_skoletzko_niddk_30apr2015.childhosp
Birth weight (kg/40 weeks of gestational age)	m_74_skoletzko_niddk_30apr2015.gestational_age
Birth length (cm/40 weeks of gestational age)	m_74_skoletzko_niddk_30apr2015.bwkg
Hospitalized by 3 months of age	m_74_skoletzko_niddk_30apr2015.blcm
Diarrhea by 3 months of age	m_74_skoletzko_niddk_30apr2015.adiarrhea
Antibiotic use during 1st year of life	m_74_skoletzko_niddk_30apr2015.childantibiotic
Age at the first antibiotic use	m_74_skoletzko_niddk_30apr2015.ant_age
Breastfeeding initiation	m_74_skoletzko_niddk_30apr2015.ever_brstfed
Duration of exclusive breastfeeding (days)	m_74_skoletzko_niddk_30apr2015.hla_d
Duration of total breastfeeding (weeks)	m_74_skoletzko_niddk_30apr2015.female
Age at gluten introduction (weeks)	m_74_skoletzko_niddk_30apr2015.celiac_fdr
Known risk factors for CDA/CD	m_74_skoletzko_niddk_30apr2015.time_to_excl_stop
Sex	m_74_skoletzko_niddk_30apr2015.anybfw
First degree relatives with CD, n (%)	m_74_skoletzko_niddk_30apr2015.glutentime
Country	m_74_skoletzko_niddk_30apr2015.country

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

Maternal Characteristics		**Vaginal delivery (N)**			**Vaginal delivery (%)**		
Variable	**Categories**	**Manuscript**	**DSIC**	**Diff**	**Manuscript**	**DSIC**	**Diff**
Age, years							
Education	<=High school	838	838	0	19	19	0
	>High school	3551	3551	0	79	79	0
Smoking during pregnancy		498	498	0	11	11	0
Antibiotic use during pregnancy		839	839	0	19	19	0
Pre-pregnancy BMI							
Gestational weight gain, kg							
Any diabetes during pregnancy		333	333	0	7	7	0
Child characteristics	Gestational age (weeks)						
	Birth weight (kg/40 weeks of gestational age)						
	Birth length (cm/40 weeks of gestational age)						
	Hospitalized by 3 months of age	590	590	0	13	13	0
	Diarrhea by 3 months of age	364	364	0	8	8	0
	Antibiotic use during 1st year of life	2025	2025	0	45	45	0
Age at the first antibiotic use	<=3 months	81	81	0	4	2	2
	3 to 12 months	1889	1889	0	93	42	51
	>12 months	54	54	0	3	1	2
Breastfeeding initiation		4396	4396	0	98	98	0
Duration of exclusive breastfeeding (days)							
Duration of total breastfeeding (weeks)							
Age at gluten introduction (weeks)							
Known risk factors for CDA/CD HLA, n (%)	DQ2/DQ2	889	889	0	20	20	0
	Others	3598	3598	0	80	80	0
Sex, n (%)	Girls	2231	2231	0	50	50	0
	Boys	2256	2256	0	50	50	0
First degree relatives with CD, n (%)	Yes	149	149	0	3	3	0

Maternal Characteristics		Vaginal delivery (N)			Vaginal delivery (%)		
Variable	**Categories**	**Manuscript**	**DSIC**	**Diff**	**Manuscript**	**DSIC**	**Diff**
	No	4338	4338	0	97	97	0
Country, n (%)	US	1561	1561	0	35	35	0
	Finland	1118	1118	0	25	25	0
	Germany	228	228	0	5	5	0
	Sweden	1580	1580	0	35	35	0

Maternal Characteristics		Vaginal delivery (Median)			Vaginal delivery (Q1)			Vaginal delivery (Q3)		
Variable	**Categories**	**Manuscript**	**DSIC**	**Diff**	**Manuscript**	**DSIC**	**Diff**	**Manuscript**	**DSIC**	**Diff**
Age, years		30	30	0	27	27	0	34	34	0
Pre-pregnancy BMI		23	23.2	-0.2	21	21	0	26	26.4	0.4
Gestational weight gain, kg		14	14	0	11	11	0	18	18	0
Child characteristics	Gestational age (weeks)	40	40	0	39	39	0	40.3	40.3	0
	Birth weight (kg/40 weeks of gestational age)	3.6	3.5	0.1	3.3	3.2	0.1	3.9	3.9	0
	Birth length (cm/40 weeks of gestational age)	51	51	0	50	49	1	53	53	0
Breastfeeding initiation										
Duration of exclusive breastfeeding (days)		28	28	0	0.5	0.5	0	112	112	0
Duration of total breastfeeding (weeks)		35	34.9	0.1	17	17.4	-0.4	52	52.1	0.1
Age at gluten introduction (weeks)		26	26.1	-0.1	22	21.7	0.3	30	30.4	0.4

Maternal Characteristics		Cesarean delivery (N)			Cesarean delivery (%)		
Variable	**Categories**	**Manuscript**	**DSIC**	**Diff**	**Manuscript**	**DSIC**	**Diff**
Age, years			.				
Education	<=High school	229	229	0	14	14	0
	>High school	1331	1331	0	83	83	0
Smoking during pregnancy		157	157	0	10	10	0
Antibiotic use during pregnancy		392	392	0	25	25	0
Pre-pregnancy BMI				0			0
Gestational weight gain, kg			.				
Any diabetes during pregnancy		276	276	0	17	17	0
Child characteristics	Gestational age (weeks)						
	Birth weight (kg/40 weeks of gestational age)		.				
	Birth length (cm/40 weeks of gestational age)		.				
	Hospitalized by 3 months of age	244	244	0	15	15	0
	Diarrhea by 3 months of age	164	164	0	10	10	0
	Antibiotic use during 1st year of life	806	806	0	50	50	0
Age at the first antibiotic use	<=3 months	38	38	0	5	2	3
	3 to 12 months	749	749	0	93	47	46
	>12 months	18	18	0	2	1	1
Breastfeeding initiation		1533	1533	0	96	96	0
Duration of exclusive breastfeeding (days)							
Duration of total breastfeeding (weeks)							
Age at gluten introduction (weeks)							
Known risk factors for CDA/CD HLA, n (%)	DQ2/DQ2	340	340	0	21	21	0
	Others	1260	1260	0	79	79	0
Sex, n (%)	Girls	761	761	0	48	48	0
	Boys	839	839	0	52	52	0
First degree relatives with CD, n (%)	Yes	43	43	0	3	3	0

Maternal Characteristics		Cesarean delivery (N)			Cesarean delivery (%)		
Variable	**Categories**	**Manuscript**	**DSIC**	**Diff**	**Manuscript**	**DSIC**	**Diff**
	No	1557	1557	0	97	97	0
Country, n (%)	US	918	918	0	57	57	0
	Finland	249	249	0	16	16	0
	Germany	139	139	0	9	9	0
	Sweden	294	294	0	18	18	0

Maternal Characteristics		Cesarean delivery (Median)			Cesarean delivery (Q1)			Cesarean delivery (Q3)		
Variable	**Categories**	**Manuscript**	**DSIC**	**Diff**	**Manuscript**	**DSIC**	**Diff**	**Manuscript**	**DSIC**	**Diff**
Age, years		32	32	0	28	28	0	36	36	0
Pre-pregnancy BMI		25	24.6	0.4	22	21.9	0.1	29	29	0
Gestational weight gain, kg		15	14.5	0.5	11	10.9	0.1	19	18.6	0.4
Child characteristics	Gestational age (weeks)	40	40	0	38	38	0	40	40	0
	Birth weight (kg/40 weeks of gestational age)	3.6	3.5	0.1	3.2	3.1	0.1	3.9	3.9	0
	Birth length (cm/40 weeks of gestational age)	52	50.8	1.2	50	49	1	54	53	1
Duration of exclusive breastfeeding (days)		7	7	0	0.5	0.5	0	42	42	0
Duration of total breastfeeding (weeks)		30	30.4	-0.4	8	8	0	50	50	0
Age at gluten introduction (weeks)		26	26.1	-0.1	22	21.7	0.3	35	34.9	0.1

Attachment A: SAS Code

```
options nocenter validvarname=upcase;

title 'prj/niddk/ims_analysis/TEDDY/prog_initial_analysis/m_74_dsic.sas';
run;

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

libname sasfile2 '/prj/niddk/ims_analysis/TEDDY/private_orig_data/M_74_SKoletzko_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 smokef
  0="Did not smoke at all"
  1="Always smoked or smoked some and then quit"
  ;

  value abage
  1="<=3 months"
  2="3 to 12 months"
  3=">12 months"
  ;

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

run;


%readin(2, m_74_skoletzko_niddk_30apr2015);

proc freq data=m_74_skoletzko_niddk_30apr2015;
  tables CSEC MOM_EDUCATION/missing;
run;

data analy;
  set m_74_skoletzko_niddk_30apr2015;
  * create subset flag for each row to use in macro call;
  all = 1;
```

```
  if CSEC=0 then vaginal=1;
  else if CSEC=1 then cesarean=1;
run;

proc freq data=analy;
  tables csec*vaginal*cesarean/list missing;
run;



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

%univ(1   , maternal_age      , vaginal , vag);
%univ(6   , mombmi            , vaginal , vag);
%univ(7   , prgwtgain         , vaginal , vag);
%univ(9   , gestational_age   , vaginal , vag);
%univ(10  , bwkg              , vaginal , vag);
%univ(11  , blcm              , vaginal , vag);
%univ(19  , time_to_excl_stop , vaginal , vag);
%univ(20  , anybfw            , vaginal , vag);
%univ(21  , glutentime        , vaginal , vag);

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


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

%univ(1  , maternal_age              , cesarean , ces);
%univ(6  , mombmi             , cesarean , ces);
%univ(7  , prgwtgain          , cesarean , ces);
%univ(9  , gestational_age    , cesarean , ces);
%univ(10  , bwkg              , cesarean , ces);
%univ(11  , blcm              , cesarean , ces);
%univ(19  , time_to_excl_stop , cesarean , ces);
%univ(20  , anybfw            , cesarean , ces);
%univ(21  , glutentime        , cesarean , ces);
```

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

* combine rows;
proc sort data=prntunivvag;
  by rownum ;
run;

proc sort data=prntunivces (rename=(_median_ = ces_median_
                                    _q1_    = ces_q1_
                                   _q3_    = ces_q3_ ))
                                   ;
  by rownum ;
run;

data alluniv;
  merge prntunivvag (in=in1 keep = rownum _var_  _median_ _q1_ _q3_)
        prntunivces (in=in2 keep = rownum _var_  ces_median_ ces_q1_ ces_q3_);
  by rownum;
  if in1 or in2;
  _median_    = round(_median_   , 0.1);
  _q1_       = round(_q1_      , 0.1);
  _q3_      = round(_q3_      , 0.1);
  ces_median_ = round(ces_median_, 0.1);
  ces_q1_    = round(ces_q1_   , 0.1);
  ces_q3_    = round(ces_q3_   , 0.1);
run;



* n and percent;

*vaginal;
data prntvag;
  set _null_;
run;

%npercent(2, mom_education    , educf        , vaginal    , vag);
%npercent(3, rsmoker          , smokef       , vaginal    , vag);
%npercent(4, momantibiotic    , yesno        , vaginal    , vag);
%npercent(5, mdiab            , yesno        , vaginal    , vag);
%npercent(8, childhosp        , yesno        , vaginal    , vag);
%npercent(12, adiarrhea              , yesno        , vaginal    , vag);
```

```
%npercent(13, childantibiotic , yesno        , vaginal    , vag);
%npercent(14, ant_age         , abage              , vaginal    , vag);
%npercent(15, ever_brstfed      , yesno            , vaginal    , vag);
%npercent(16, hla_d           , yesno              , vaginal    , vag);
%npercent(17, female          , sexf         , vaginal    , vag);
%npercent(18, celiac_fdr        , yesno            , vaginal    , vag);
%npercent(22, country           , country    , vaginal    , vag);

proc print data=prntvag;
  var rownum covar covarf count    percent;
title3 "vaginal";
run;


*cesarean;
data prntces;
  set _null_;
run;

%npercent(2, mom_education    , educf        , cesarean    , ces);
%npercent(3, rsmoker          , smokef       , cesarean    , ces);
%npercent(4, momantibiotic    , yesno        , cesarean    , ces);
%npercent(5, mdiab            , yesno        , cesarean    , ces);
%npercent(8, childhosp        , yesno        , cesarean    , ces);
%npercent(12, adiarrhea               , yesno        , cesarean    , ces);
%npercent(13, childantibiotic , yesno        , cesarean    , ces);
%npercent(14, ant_age         , abage              , cesarean    , ces);
%npercent(15, ever_brstfed      , yesno            , cesarean    , ces);
%npercent(16, hla_d           , yesno              , cesarean    , ces);
%npercent(17, female          , sexf         , cesarean    , ces);
%npercent(18, celiac_fdr        , yesno            , cesarean    , ces);
%npercent(22, country           , country    , cesarean    , ces);

proc print data=prntces;
  var rownum covar covarf count    percent;
title3 "cesarean";
run;


* combine rows;
proc sort data=prntvag;
  by rownum covar covarf;
run;

proc sort data=prntces (rename=(count    = count_ces
```

```
                              percent  = percent_ces))
                              ;
  by rownum covar covarf;
run;

data allnpercent;
  merge prntvag (in=in1 keep = rownum covar covarf  count   percent   )
        prntces (in=in2 keep = rownum covar covarf  count_ces  percent_ces);
  by rownum covar covarf;
  if in1 or in2;
  * round percentages;
  percent     = round(percent);
  percent_ces = round(percent_ces);
run;



* Table 1;
data table1mod;
  set alluniv allnpercent;
run;

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

proc print data=table1mod;
  var rownum covar covarf _var_ count percent  _median_ _q1_ _q3_ count_ces  percent_ces ces_median_ ces_q1_ ces_q3_;
  title3 "Table 1";
run;
```