

Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) Pub29 CAronsson

**Prepared by Jane Wang
IMS Inc.**

3901 Calverton Blvd, Suite 200 Calverton MD 20705

September 08, 2015

Table of Contents

1 Standard Disclaimer.....	1
2 Study Background.....	1
3 Archived Datasets.....	2
4 Statistical Methods.....	2
5 Results.....	2
6 Conclusion.....	2
7 References.....	2
Attachment A: SAS Code.....	8
Table A: Variables used to replicate Table 1	3
Table B: Comparison of values computed in integrity check to reference article Table 1 <u>Socio demographic characteristics of mothers and their infants participating in The Environmental Determinants of Diabetes in the Young Study for 2 years or more</u>	4
Table C: Variables used to replicate Table 2	6
Table D: Comparison of values computed in integrity check to reference article Table 2 <u>values Breastfeeding duration (weeks) and age at first introduction to complementary foods (weeks), in children (0–24 months of age) in the multinational The Environmental Determinants of Diabetes in the Young study</u>	6

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 “Pub29_CARosson_niddk_submission” dataset.

4 Statistical Methods

Analyses were performed to duplicate results for the data published by Carin Andrén Aronsson et al [1]. *Maternal and Child Nutrition* (2013), DOI: 10.1111/mcn.12084. To verify the integrity of the dataset, descriptive statistics of Socio demographic variables (table 1), and Breastfeeding duration and age at first introduction to complementary foods variables (table 2) were computed, by country.

5 Results

Table A and C lists the variables that were used in the replication and Table B and D compares the results calculated from the archived data file to the results published in Table 1 and Table 2. The results of the replication are the similar to the published results.

6 Conclusions

The NIDDK repository is confident that the TEDDY Pub29 CARonsson data files to be distributed are a true copy of the study data.

7 References

Carin Andrén Aronsson, Ulla Uusitalo, Kendra Vehik, Jimin Yang, Katherine Silvis, Sandra Hummel, Suvi M. Virtanen, Jill M. Norris, and the TEDDY Study Group. Age at first introduction to complementary foods is associated with sociodemographic factors in children with increased genetic risk of developing type 1 diabetes. *Maternal and Child Nutrition* (2013), DOI: 10.1111/mcn.12084

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

Table Variable	Variables Used in Replication from the "Table 1"
Maternal age,	matage
Mothers first child	mom_first_child
Marital status	marital
Maternal education	education_mom_group1
Smoked during pregnancy	smoker
Alcohol during pregnancy	drinker
Crowding	crowd
sex	gender
Birthweight (g)	babys
Gestational age, weeks	ge
Postmature (>42 weeks) Child with T1D FDR	fdr

Table B: Comparison of values computed in integrity check to reference article Table 1 values: Action to prevent type 1 diabetes by category

n (%)	All n = 6404 [Manuscript]	All n = 6404 [DSIC]	All n = 6404 [Difference]	US n = 2540 [Manuscript]	US n = 2540 [DSIC]	US n = 2540 [Difference]
Maternal age,<25.0	687 (10.7)	687 (10.7)	0(0)	327 (12.9)	327 (12.9)	0(0)
Maternal age,25.0–29.9	1867 (29.1)	1867 (29.2)	0(-0.1)	647 (25.5)	647 (25.5)	0(0)
Maternal age,30.0–34.9	2328 (36.4)	2328 (36.4)	0(0)	878 (34.5)	878 (34.6)	0(-0.1)
Maternal age,≥35.0	1522 (23.8)	1522 (23.8)	0(0)	688 (27.1)	688 (27.1)	0(0)
Mothers first child, n (%) Yes	2879 (45.0)	2879 (45.0)	0(0)	1066 (42.0)	1066 (42.0)	0(0)
Marital status,Married	4451 (69.5)	4451 (69.5)	0(0)	2168 (85.4)	2168 (85.4)	0(0)
Marital status,Cohabit	1709 (26.7)	1709 (26.7)	0(0)	232 (9.1)	232 (9.1)	0(0)
Marital status,Single, divorced, widow	244 (3.8)	244 (3.8)	0(0)	140 (5.5)	140 (5.5)	0(0)
Maternal education, Basic primary	1190 (18.6)	1190 (18.6)	0(0)	350 (13.8)	350 (13.8)	0(0)
Maternal education, High school or more	5214 (81.4)	5214 (81.4)	0(0)	2190 (86.2)	2190 (86.2)	0(0)
Smoked during pregnancy,Yes	686 (10.7)	686 (10.7)	0(0)	209 (8.2)	209 (8.2)	0(0)
Alcohol during pregnancy,Yes	2242 (35.0)	2242 (35.0)	0(0)	1027 (40.4)	1027 (40.4)	0(0)
Crowding, ≤1 people per room	5850 (91.4)	5850 (91.3)	0(0.1)	2342 (92.2)	2342 (92.2)	0(0)
Crowding, >1 people per room	554 (8.6)	554 (8.7)	0(-0.1)	198 (7.8)	198 (7.8)	0(0)
Sex, n (%)Girls	3114 (48.6)	3290 (51.4)	0(0)	1223 (48.1)	1223 (48.1)	0(0)
Birthweight (g),<2500	365 (5.7)	365 (5.7)	0(0)	252 (9.9)	252 (9.9)	0(0)
Birthweight (g),2500–3999	4991 (77.9)	4991 (77.9)	0(0)	1995 (78.6)	1995 (78.6)	0(0)
Birthweight (g),≥4000	1048 (16.4)	1048 (16.4)	0(0)	293 (11.5)	293 (11.5)	0(0)
Gestational age, weeks, Premature (<37 weeks)	341 (5.3)	341 (5.3)	0(0)	168 (7.1)	168 (6.6)	0(0.5)
Gestational age, weeks, Normal (37–42 weeks)	5932 (92.8)	5932 (92.8)	0(0)	2360 (93.1)	2360 (93.1)	0(0)
Gestational age, weeks, Postmature (>42 weeks)	112 (1.9)	122 (1.9)	0(0)	6 (0.2)	6 (0.2)	0(0)
Child with T1D FDR, n (%)Yes	742 (11.6)	742 (11.6)	0(0)	296 (11.6)	296 (11.7)	0(-0.1)

n (%)	Sweden n= 2005 [Manuscript]	Sweden n= 2005 [DSIC]	Sweden n= 2005 [Diff]	Finland n= 1445 [Manuscript]	Finland n= 1445 [DSIC]	Finland n= 1445 [Diff]	Germany n= 414 [Manuscript]	Germany n= 414 [DSIC]	Germany n= 414 [Diff]
Maternal age,<25.0	154 (7.7)	154	0(0)	177 (12.2)	177	0(0)	29 (7.0)	29 (7.0)	0(0)
Maternal age,25.0–29.9	607 (30.3)	607	0(0)	515 (35.6)	515	0(0)	98 (23.7)	98 (23.7)	0(0)
Maternal age,30.0–34.9	805 (40.1)	805	0(0)	481 (33.3)	481	0(0)	164 (39.6)	164	0(0)
Maternal age,≥35.0	439 (21.9)	439	0(0)	272 (18.8)	272	0(0)	123 (29.7)	123	0(0)
Mothers first child, n (%)	955 (47.6)	955	0(0)	653 (45.2)	653	0(0)	205 (49.5)	205	0(0)
Marital status,Married	966 (48.2)	966	0(0)	980 (67.8)	980	0(0)	337 (81.4)	337	0(0)
Marital status,Cohabit	998 (49.8)	998	0(0)	416 (28.8)	416	0(0)	63 (15.2)	63 (15.2)	0(0)
Marital status,Single, divorced, widow	41 (2.0)	41 (2.0)	0(0)	49 (3.4)	49 (3.4)	0(0)	14 (3.4)	14 (3.4)	0(0)
Maternal education, Basic	659 (32.9)	659	0(0)	132 (9.1)	132	0(0)	49 (11.8)	49 (11.8)	0(0)
Maternal education, High school or more	1346 (67.1)	1346 (67.1)	0(0)	1313 (90.9)	1313 (90.9)	0(0)	365 (88.2)	365 (88.2)	0(0)
Smoked during	229 (11.4)	229	0(0)	181 (12.5)	181	0(0)	67 (16.2)	67 (16.2)	0(0)
Alcohol during	570 (28.4)	570	0(0)	444 (30.7)	444	0(0)	201 (48.6)	201	0(0)
Crowding, ≤1 people per	1864 (93.0)	1864	0(0)	1284 (88.8)	1284	0(0)	360 (87.0)	360	0(0)
Crowding, >1 people per	141 (7.0)	141	0(0)	161 (11.2)	161	0(0)	54 (13.0)	54 (13.0)	0(0)
Sex, n (%)Girls	981 (48.9)	981	0(0)	709 (49.1)	709	0(0)	201 (48.6)	201	0(0)
Birthweight (g),<2500	50 (2.5)	50 (2.5)	0(0)	47 (3.2)	47	0(-0.1)	16 (3.9)	16 (3.9)	0(0)
Birthweight (g),2500–3999	1522 (75.9)	1522	0(0)	1125 (77.9)	1125	0(0)	349 (84.3)	349	0(0)
Birthweight (g),≥4000	433 (21.6)	433	0(0)	273 (18.9)	273	0(0)	49 (11.8)	49 (11.8)	0(0)
Gestational age, weeks, Premature (<37 weeks)	88 (4.4)	88 (4.4)	0(0)	59 (4.1)	59 (4.1)	0(0)	26 (6.3)	26 (6.3)	0(0)
Gestational age, weeks, Normal (37–42 weeks)	1823 (91.0)	1823 (91.0)	0(0)	1362 (94.3)	1362 (94.3)	0(0)	387 (93.7)	387 (93.7)	0(0)
Gestational age, weeks, Postmature (>42 weeks)	93 (4.6)	93 (4.6)	0(0)	23 (1.6)	23 (1.6)	0(0)	0 (0.0)	0 (0.0)	0(0)
Child with T1D FDR, n	137 (6.8)	137	0(0)	136 (9.4)	136	0(0)	173 (41.8)	173	0(0)

Table C: Variables used to replicate Tables 2 in the publication.

Table Variable	Variables Used in Replication from the "Table 2"
Breastfeeding duration, weeks Exclusive	exclbrst_weeks
Breastfeeding duration, weeks Any	anybrst_weeks
Cow's milk	cowmilk_weeks
Potatoes and roots	roots_weeks
Fruits and berries	fruits_weeks
Cereals	cereals_weeks
Gluten-containing cereals	gluten_weeks

Table D: Comparison of values computed in integrity check to reference article Table 2 values: Socio demographic variables associated with actions to prevent diabetes

Median (Q1,Q3)	All n = 6404 [Manuscript]	All n = 6404 [DSIC]	All n = 6404 [Difference]	US n = 2540 [Manuscript]	US n = 2540 [DSIC]	US n = 2540 [Difference]
Breastfeeding duration, weeks Exclusive	2.0 (0.07, 13.0)	2 (0.07, 13)	0(0,0)	1.0 (0.07, 8.0)	1.0 (0.07, 8.0)	0(0,0)
Breastfeeding duration, weeks Any	33.0 (12.0, 52.1)	33 (12, 52.1)	0(0,0)	30.4 (8.0, 53.7)	30.4 (8.00, 53.7)	0(0,0)
Age at first introduction, weeks Cow's milk	3.0 (0.07, 17.4)	3 (0.07, 17.4)	0(0,0)	1.0 (0.07, 12.0)	1.0 (0.07, 12.0)	0(0,0)
Age at first introduction, weeks Potatoes and roots	19.6 (17.4, 23.8)	19.6 (17.4, 23.9)	0(0,-0.1)	23.8 (19.6, 26.1)	23.9 (19.6, 26.1)	0(0,0)
Age at first introduction, weeks Fruits and berries	21.7 (17.4, 26.1)	21.7 (17.4, 26.1)	0(0,0)	21.7 (17.4, 26.1)	21.7 (17.46, 26.1)	0(0,0)
Age at first introduction, weeks Cereals	21.7 (17.4, 23.8)	21.7 (17.4, 23.9)	0(0,-0.1)	19.6 (16.0, 23.8)	19.6 (16.0, 23.9)	0(0,0)
Age at first introduction, weeks Gluten-containing cereals	26.1 (21.7, 30.4)	26.1 (21.7, 30.4)	0(0,0)	30.4 (26.1, 34.8)	30.4 (26.1, 34.9)	0(0,0)

Median (Q1,Q3)	Sweden n = 2005 [Manuscript]	Sweden n = 2005 [DSIC]	Sweden n = 2005 [Difference]	Finland n = 1445 [Manuscript]	Finland n = 1445 [DSIC]	Finland n = 1445 [Difference]
Breastfeeding duration, weeks Exclusive	4.0 (0.07, 16.0)	4.0 (0.07, 16.0)	0(0,0)	3.0 (0.07, 13.0)	3.0(0.07,13.0)	0(0,0)
Breastfeeding duration, weeks Any	31.6 (17.4, 43.4)	31.6 (17.4, 43.4)	0(0,0)	38.7 (21.0, 54.0)	38.7(21.0,54.0)	0(0,0)
Age at first introduction, weeks Cow's milk	4.0 (0.07, 17.4)	4.0 (0.07, 17.4)	0(0,0)	4.3 (1.0, 21.7)	4.3(1.0,21.7)	0(0,0)
Age at first introduction, weeks Potatoes and roots	17.4 (17.4, 19.6)	17.4 (17.4, 19.6)	0(0,0)	17.4 (16.0, 21.7)	17.4(16.0,21.7)	0(0,0)
Age at first introduction, weeks Fruits and berries	19.6 (17.4, 21.7)	19.6 (17.4, 21.7)	0(0,0)	19.6 (17.4, 23.8)	19.6(17.4,23.9)	0(0,-0.1)
Age at first introduction, weeks Cereals	19.6 (17.4, 21.7)	19.6 (17.4, 21.7)	0(0,0)	21.7 (21.7, 26.1)	21.7(21.7,26.1)	0(0,0)
Age at first introduction, weeks Gluten-containing cereals	21.7 (17.4, 23.8)	21.7 (17.4, 23.9)	0(0,-0.1)	26.1 (23.8, 30.4)	26.1(23.9,30.4)	0(-0.1,0)

Median (Q1,Q3)	Germany n = 414 [Manuscript]	Germany n = 414 [DSIC]	Germany n = 414 [Difference]
Breastfeeding duration, weeks Exclusive	3.5 (0.07, 21.7)	3.5(0.07,21.7)	0(0,0)
Breastfeeding duration, weeks Any	32.1 (10.0, 48.1)	32.1(10.0,48.1)	0(0,0)
Age at first introduction, weeks Cow's milk	6.0 (0.07, 26.1)	6.0(0.07,26.1)	0(0,0)
Age at first introduction, weeks Potatoes and roots	23.8 (19.6, 26.1)	23.9(19.6,26.1)	0(0,0)
Age at first introduction, weeks Fruits and berries	26.1 (21.7, 29.3)	26.1(21.7,29.3)	0(0,0)
Age at first introduction, weeks Cereals	26.1 (21.7, 30.4)	26.1(21.7,30.4)	0(0,0)
Age at first introduction, weeks Gluten-containing cereals	30.4 (26.1, 37.0)	30.4(26.1,37.0)	0(0,0)

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/Pub29_CARosson_niddk_submission/";
data Pub29_caronsson_niddk ; set sas_data.Pub29_caronsson_niddk ;

%macro baseline_freq1(dataset_name,var_name,);

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

proc freq data = &dataset_name ;
    table (&var_name.)*ctry ;
    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._cross(keep = ctry Frequency Colpercent table_name &var_name);
    set &var_name._cross;
    if ctry ne . and Colpercent ne .;
    length table_name $30.;
    table_name = "&var_name" ;

proc sort data = &var_name._cross;
    by &var_name ctry;
proc print data = &var_name._cross;

data &var_name._cross;
    set &var_name._cross;
    by &var_name;
    retain count 0;
    if first.&var_name then count = 0;
    count = count + 1;

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

```

data &var_name._cross_1(drop = Frequency Colpercent i);
set &var_name._cross;
by &var_name ;
array temp1(4) countus countfinland countgermany countsweden ;
array temp2(4) pertus pertfinland pertgermany pertsweden ;
retain countus countfinland countgermany countsweden pertus pertfinland pertgermany pertsweden ;
if first.&var_name then do i = 1 to 4;
    temp1(i) = .;
    temp2(i) = .;
end;
temp1(count) = Frequency;
temp2(count) = round(Colpercent,0.1);
if last.&var_name;

proc print data = &var_name._cross_1;

proc freq data = &dataset_name ;
    table &var_name/ out = &var_name._freq;
    title3 "Frequency table of the &var_name. variable in the analysis dataset";
data &var_name._freq;
set &var_name._freq;
length table_name $30.;
table_name = "&var_name" ;
PERCENT = round(PERCENT,0.1);
rename count = countall PERCENT =PERTall;

proc print data = &var_name._freq;

data &var_name._c;
merge &var_name._freq &var_name._cross_1;
by table_name &var_name;
rename &var_name = var_level;
table_name = compress(table_name || put(&var_name,1.));
proc print data = &var_name._c;

%mend;

%macro baseline_means(dataset_name,var_name);

proc sort data = Pub29_caronsson_niddk;
by ctry;
    *** Creating a frequency table in the format of Table 1 in the primary outcome paper;
proc means data = &dataset_name median p25 p75;
var &var_name.;
by ctry;
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 ctry ne .;

proc print data = &var_name._means1;

proc sort data = &var_name._means1;
  by table_name ctry;
data &var_name._means_1(drop = ctry &var_name._Median &var_name._P25 &var_name._P75 i);
  set &var_name._means1;
  by table_name;
  array temp1(4) median_us median_finland median_germany median_sweden ;
  array temp2(4) p25_us p25_finland p25_germany p25_sweden ;
  array temp3(4) p75_us p75_finland p75_germany p75_sweden ;
  retain median_us median_finland median_germany median_sweden p25_us p25_finland p25_germany p25_sweden p75_us p75_finland
p75_germany p75_sweden ;
  if first.table_name then do i = 1 to 4;
    temp1(i) = .;
    temp2(i) = .;
    temp3(i) = .;
  end;
  temp1(_n_) = round(&var_name._Median,0.1);
  temp2(_n_) = round(&var_name._P25,0.01);
  temp3(_n_) = round(&var_name._P75,0.1);
  if last.table_name;

proc print data = &var_name._means_1;

proc means data = &dataset_name median p25 p75;
  var &var_name.;
  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._means2;
run;

data &var_name._means2;
  set &var_name._means2;
  length table_name $30.;
  table_name = "&var_name";
  &var_name._Median = round(&var_name._Median,0.1);
  &var_name._P25 = round(&var_name._P25,0.01);

```

```

        &var_name._P75      = round(&var_name._P75,0.1);
    rename &var_name._Median = all_Median
        &var_name._P25      = all_P25
        &var_name._P75      = all_P75
;
    proc print data = &var_name._means2;

data &var_name._m;
    merge &var_name._means_1 &var_name._means2;
    by table_name;

proc print data = &var_name._m;

%mend;

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

*** Running the baseline_freq on the categorical variables in the Table 1 manuscript file;
%baseline_freq1(Pub29_caronsson_niddk,matage );
%baseline_freq1(pub29_caronsson_niddk,mom_first_child );
%baseline_freq1(pub29_caronsson_niddk,marital );
%baseline_freq1(pub29_caronsson_niddk,education_mom_group1 );
%baseline_freq1(pub29_caronsson_niddk,smoker );
%baseline_freq1(pub29_caronsson_niddk,drinker );
%baseline_freq1(pub29_caronsson_niddk,crowd );
%baseline_freq1(pub29_caronsson_niddk,gender );
%baseline_freq1(pub29_caronsson_niddk,babys );
%baseline_freq1(pub29_caronsson_niddk,ge );
%baseline_freq1(pub29_caronsson_niddk,fdr );

data table1_compare(drop = ctry COUNT var_level);
    set
matage_c
mom_first_child_c (where = (var_level = 1))
marital_c
education_mom_group1_c
smoker_c (where = (var_level = 1))
drinker_c (where = (var_level = 1))
crowd_c
gender_c (where = (var_level = 0))
babys_c
ge_c (where = (var_level ne .))
fdr_c (where = (var_level = 1))
;

```

```

*****;
***** Check Table 2 *****;
*****;

*** Running the baseline_freq on the categorical variables in the Table 2 manuscript file;

proc means data = Pub29_caronsson_niddk median p25 p75;
  var exclbrst_weeks anybrst_weeks cowmilk_weeks roots_weeks fruits_weeks cereals_weeks gluten_weeks;

%baseline_means(Pub29_caronsson_niddk, exclbrst_weeks );
%baseline_means(Pub29_caronsson_niddk, anybrst_weeks );
%baseline_means(Pub29_caronsson_niddk, cowmilk_weeks );
%baseline_means(Pub29_caronsson_niddk, roots_weeks );
%baseline_means(Pub29_caronsson_niddk, fruits_weeks );
%baseline_means(Pub29_caronsson_niddk, cereals_weeks );
%baseline_means(Pub29_caronsson_niddk, gluten_weeks );

data table2_compare;
  set
  exclbrst_weeks_m
  anybrst_weeks_m
  cowmilk_weeks_m
  roots_weeks_m
  fruits_weeks_m
  cereals_weeks_m
  gluten_weeks_m
  ;

proc print data = table1_compare;
  title3 'DSIC for table 1';
proc print data = table2_compare;
  title3 'DSIC for table 2';

```