Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) Pub77 Aronsson

Prepared by Dominick Parisi
IMS Inc.
3901 Calverton Blvd, Suite 200 Calverton, MD 20705
February 13, 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_77_caronsson_niddk_30may2014_1.sas7bdat", "m_77_caronsson_niddk_30may2014_2.sas7bdat" and "m_77_caronsson_niddk_30may2014_3.sas7bdat" datasets.

4 Statistical Methods

Analyses were performed to duplicate results for the data published by Carin Andrén Aronsson et al [1] in Clinical Gastroenterology and Hepatology 2016. To verify the integrity of the dataset, descriptive statistics were computed.

5 Results

For Table 1 in the publication [1], **Matching Factors and Age of tTGA Seroconversion in the TEDDY Swedish Birth Cohort and in Children With Celiac Disease**, 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 almost an exact match to the published results.

For Table 2 in the publication [1], **Infant Feeding Characteristics in Children With Celiac Disease and Matched Controls,** Table C lists the variables that were used in the replication and Table D compares the results calculated from the archived data files to the results published in Table 2. The results of the replication are almost an exact match to the published results.

For Table 3 in the publication [1], **Daily Gluten Intake in Children With Celiac Disease and Matched Controls,** Table E lists the variables that were used in the replication and Table F compares the results calculated from the archived data files to the results published in Table 3. The results of the replication are almost an exact match to the published results.

For Table 4 in the publication [1], **Daily Gluten Intake at Clinic Visit Before tTGA Seroconversion in Children With Celiac Disease and Matched Controls,** Table G lists the variables that were used in the replication and Table H compares the results calculated from the archived data files to the results published in Table 4. The results of the replication are almost an exact match to the published results.

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] Carin Andrén Aronsson, Hye-Seung Lee, Sibylle Koletzko, Ulla Uusitalo, Jimin Yang, Suvi M. Virtanen, Edwin Liu, Åke Lernmark, Jill M. Norris, Daniel Agardh, and the TEDDY Study Group. Effects of Gluten Intake on Risk of Celiac Disease: A Case-Control Study on a Swedish Birth Cohort. Clinical Gastroenterology and Hepatology 2012;14:403-409.

Table A: Variables used to replicate Table 1**:** Characteristics of the first 100 the Environmental Determinants of Diabetes in the Young (TEDDY) children diagnosed with T1D

Table Variable	**dataset.variable**
Sex	m_77_caronsson_niddk_30may2014_2.female
Birth year	m_77_caronsson_niddk_30may2014_2.birthyear
HLA genotype	m_77_caronsson_niddk_30may2014_2.hlarg
Sex	m_77_caronsson_niddk_30may2014_1.female
Birth year	m_77_caronsson_niddk_30may2014_1.birthyear
HLA genotype	m_77_caronsson_niddk_30may2014_1.hlarg
Age of tTGA seroconversion	m_77_caronsson_niddk_30may2014_1.mtime_postga

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

Variable	TEDDY Manuscript Birth Cohort (n=2062) N (%)	TEDDY DISC Birth Cohort (n=2062) N (%)	Diff. (n=0)	TEDDY Manuscript Celiac Disease (n=146) N (%)	TEDDY DSIC Celiac Disease (n=146) N (%)	Diff. (n=0)
Sex						
- Boys	1055 (51)	1055 (51)	0 (0)	49 (34)	49 (34)	0 (0)
- Girls	1007 (49)	1007 (49)	0 (0)	97 (66)	97 (66)	0 (0)
Birth year						
- 2004	91 (4)	91 (4)	0 (0)	8 (6)	8 (6)	0 (0)
- 2005	366 (18)	366 (18)	0 (0)	37 (25)	37 (25)	0 (0)
- 2006	377 (18)	377 (18)	0 (0)	22 (15)	22 (15)	0 (0)
- 2007	412 (20)	412 (20)	0 (0)	31 (21)	31 (21)	0 (0)
- 2008	363 (18)	363 (18)	0 (0)	25 (17)	25 (17)	0 (0)
- 2009	385 (19)	385 (19)	0 (0)	19 (13)	19 (13)	0 (0)
- 2010	68 (3)	68 (3)	0 (0)	4 (3)	4 (3)	0 (0)
HLA genotype						
- DR3-DQ2/DR3-DQ2	438 (21)	438 (21)	0 (0)	70 (48)	70 (48)	0 (0)
- DR3-DQ2/DR4-DQ8	868 (42)	868 (42)	0 (0)	48 (33)	48 (33)	0 (0)
- DR3-DQ2/DR9	1 (<1)	1 (<1)	0 (0)			
- DR4-DQ8/DR4-DQ8	455 (22)	455 (22)	0 (0)	26 (18)	26 (18)	0 (0)
- DR4-DQ8/DR8	268 (13)	268 (13)	0 (0)	2 (1)	2 (1)	0 (0)
- DR4-DQ8/DR1	18 (1)	18 (1)	0 (0)			
- DR4-DQ8/DR13	13 (<1)	13 (<1)	0 (0)			
- DR4-DQ8/DR9	1 (<1)	1 (<1)	0 (0)			

Variable	TEDDY Manuscript Age at tTGA seroconversion Median (Q1, Q3)	TEDDY DSIC Age at tTGA seroconversion Median (Q1, Q3)	Diff. (n=0)	TEDDY Manuscript Kruskal-Wallis p-value	TEDDY DSIC Kruskal-Wallis p-value	Diff. (n=0)
Sex				.066	.056	.010
- Boys	29 (21, 48)	29 (21, 48)	0 (0, 0)			
- Girls	24 (18, 36)	24 (18, 36)	0 (0, 0)			
Birth year				.066	.059	.007
- 2004						
- 2005	28 (18, 48)	28 (18, 48)	0 (0, 0)			
- 2006	30 (21, 59)	30 (21, 59)	0 (0, 0)			
- 2007	24 (20, 37)	24 (20, 37)	0 (0, 0)			
- 2008	30 (22, 36)	30 (22, 36)	0 (0, 0)			
- 2009	21 (17, 24)	21 (17, 24)	0 (0, 0)			
- 2010						
HLA genotype				<.0001	<.0001	0
- DR3-DQ2/DR3-DQ2	21.5 (17, 28)	21.5 (17, 28)	0 (0, 0)			
- DR3-DQ2/DR4-DQ8	36 (22.5, 48.5)	36 (22.5, 48.5)	0 (0, 0)			
- DR3-DQ2/DR9						
- DR4-DQ8/DR4-DQ8	35 (21.5, 37)	35 (21.5, 37)	0 (0, 0)			
- DR4-DQ8/DR8						
- DR4-DQ8/DR1						
- DR4-DQ8/DR13						
- DR4-DQ8/DR9						

Table C: Variables used to replicate Table 2: Symptoms and laboratory data at onset of T1D

Table Variable	**dataset.variable**
Breastfeeding duration, total	m_77_caronsson_niddk_30may2014_1.mbrst
Breastfeeding duration, exclusive	m_77_caronsson_niddk_30may2014_1.mexbrst
Age at first introduction, gluten-containing cereals	m_77_caronsson_niddk_30may2014_1.mgluten
Age at first introduction, wheat	m_77_caronsson_niddk_30may2014_1.mwheat
Energy intake, kcal	m_77_caronsson_niddk_30may2014_3.tot_ene

Table D: Comparison of values computed in integrity check to reference article Table 2 values

Variable	TEDDY Manuscript Celiac Disease (n=146)	TEDDY DSIC Celiac Disease (n=146)	Diff (n=0)
Breastfeeding duration, wk			
- Total	31 (20, 40)	31 (20, 40)	0 (0, 0)
- Exclusive	4 (1, 14)	4 (1, 14)	0 (0, 0)
Age at first introduction, wk			
- gluten-containing cereals	22 (18, 24)	22 (18, 24)	0 (0, 0)
- wheat	22 (20, 25)	22 (20, 25)	0 (0, 0)
Energy intake, kcal	1019 (840, 1164)	1019 (840, 1164)	0 (0, 0)

Variable	TEDDY Manuscript Controls (n=436)	TEDDY DSIC Controls (n=436)	Diff (n=0)
Breastfeeding duration, wk			
- Total	33 (18, 43)	33 (18, 43)	0 (0, 0)
- Exclusive	6 (1, 16)	6 (1, 16)	0 (0, 0)
Age at first introduction, wk			
- gluten-containing cereals	22 (18, 24)	22 (18, 24)	0 (0, 0)
- wheat	22 (18, 25)	22 (18, 25)	0 (0, 0)
Energy intake, kcal	1009 (858, 1156)	1009 (858, 1156)	0 (0, 0)

Variable	TEDDY Manuscript OR (95% CI)	TEDDY DSIC OR (95% CI)	Diff (n=0)
Breastfeeding duration, wk			
- Total	0.99 (0.99-1.00)	1.00 (0.99-1.01)	0.01 (0-0.01)
- Exclusive	0.98 (0.96-1.00)	0.98 (0.96-1.01)	0 (0-0.01)
Age at first introduction, wk			
- gluten-containing cereals	0.99 (0.95-1.04)	1.00 (0.95-1.04)	0.01 (0-0)
- wheat	1.00 (0.96-1.05)	1.00 (0.96-1.05)	0 (0-0)
Energy intake, kcal	1.00 (1.00-1.00)	1.00 (1.00-1.00)	0 (0-0)

Variable	TEDDY Manuscript p-value	TEDDY DSIC p-value	Diff (n=0)
Breastfeeding duration, wk			
- Total	.361	.361	0
- Exclusive	.124	.124	0
Age at first introduction, wk			
- gluten-containing cereals	.866	.858	.008
- wheat	.888	.878	.010
Energy intake, kcal	.450	.450	0

Table E: Variables used to replicate Table 3: Daily Gluten Intake in Children With Celiac Disease and Matched Controls

Table Variable	**dataset.variable**
Total gluten (g) intake before tTGA seroconversion	m_77_caronsson_niddk_30may2014_1.sglut
Gluten (g) intake at the visit before tTGA seroconversion	m_77_caronsson_niddk_30may2014_1.last1glut

Table F: Comparison of values computed in integrity check to reference article Table 3 values

	TEDDY Manuscript Celiac Disease (N=146) Median (Q1, Q3)	TEDDY DSIC Celiac Disease (N=146) Median (Q1, Q3)	Diff (n)
Total gluten (g) intake before tTGA seroconversion	10.5 (7.6, 14.2)	10.5 (7.6, 14.2)	0 (0, 0)
Gluten (g) intake at the visit before tTGA seroconversion	4.9 (3.5, 5.9)	4.9 (3.5, 5.9)	0 (0, 0)

	TEDDY Manuscript Controls (N=146) Median (Q1, Q3)	TEDDY DSIC Controls (N=146) Median (Q1, Q3)	Diff (n)
Total gluten (g) intake before tTGA seroconversion	9.9 (5.9, 13.8)	9.9 (5.9, 13.7)	0 (0, 0.1)
Gluten (g) intake at the visit before tTGA seroconversion	3.9 (2.9, 5.2)	3.9 (2.9, 5.2)	0 (0, 0)

	TEDDY Manuscript OR (95% CI)	TEDDY DSIC OR (95% CI)	Diff (n)	TEDDY Manuscript p-value	TEDDY DSIC p-value	Diff (n)
Total gluten (g) intake before tTGA seroconversion	1.05 (1.01-1.10)	1.05 (1.00-1.10)	0 (0-0)	.030	.030	0
Gluten (g) intake at the visit before tTGA seroconversion	1.28 (1.13-1.46)	1.28 (1.13-1.46)	0 (0-0)	.0002	.0002	0

Table G: Variables used to replicate Table 4: Daily Gluten Intake at Clinic Visit Before tTGA Seroconversion in Children With Celiac Disease and Matched Controls

Table Variable	**dataset.variable**
Age at 3-day food record, mo	m_77_caronsson_niddk_30may2014_1.lastmonth
Gluten (g) intake at the visit before tTGA seroconversion	m_77_caronsson_niddk_30may2014_1.last1glut

Table H: Comparison of values computed in integrity check to reference article Table 4 values

Variable	TEDDY Manuscript Celiac Disease (n=132) (N)	TEDDY DSIC Celiac Disease (n=132) (N)	Diff (n=0)	TEDDY Manuscript Controls (n=385) (N)	TEDDY DSIC Controls (n=385) (N)	Diff (n=0)
Age at 3-day food record						
- 9 months	6	6	0	17	17	0
- 12 months	32	32	0	89	89	0
- 18 months	37	37	0	103	103	0
- 24 months	57	57	0	176	176	0

Variable	TEDDY Manuscript Celiac Disease (n=132) Med (Q1, Q3)	TEDDY DSIC Celiac Disease (n=132) Med (Q1, Q3)	Diff (n=0)	TEDDY Manuscript Controls (n=385) Med (Q1, Q3)	TEDDY DSIC Controls (n=385) Med (Q1, Q3)	Diff (n=0)
Age at 3-day food record						
- 9 months	1.6 (1.4, 1.8)	1.6 (1.4, 1.8)	0 (0, 0)	1.9 (1.1, 2.4)	1.9 (1.1, 2.3)	0 (0, 0.1)
- 12 months	4.9 (3.5, 5.6)	4.9 (3.5, 5.6)	0 (0, 0)	3.2 (2.5, 4.5)	3.2 (2.5, 4.5)	0 (0, 0)
- 18 months	4.9 (3.9, 5.9)	4.9 (3.9, 5.9)	0 (0, 0)	3.9 (3.2, 5.2)	3.9 (3.1, 5.2)	0 (0, 0)
- 24 months	5.1 (3.7, 6.2)	5.1 (3.7, 6.2)	0 (0, 0)	4.3 (3.3, 5.7)	4.3 (3.3, 5.7)	0 (0, 0)

Variable	TEDDY Manuscript OR (95% CI)	TEDDY DSIC OR (95%CI)	Diff	TEDDY Manuscript p-value	TEDDY DSIC p-value	Diff
Age at 3-day food record						
- 9 months	0.63 (0.19-2.05)	0.63 (0.19-2.05)	0 (0-0)	.444	.444	0
- 12 months	1.58 (1.17-2.13)	1.58 (1.17-2.13)	0 (0-0)	.003	.003	0
- 18 months	1.22 (0.99-1.51)	1.22 (0.99-1.51)	0 (0-0)	.077	.067	.01
- 24 months	1.23 (1.01-1.49)	1.23 (1.01-1.49)	0 (0-0)	.043	.043	0

Attachment A: SAS Code

```
LIBNAME SASDATA '/prj/niddk/ims_analysis/TEDDY/private_orig_data/M_77_CAronsson_NIDDK_Submission/';

/*******************/
/* Import datasets */
/*******************/
DATA ANALYSIS1;
  SET SASDATA.m_77_caronsson_niddk_30may2014_1;
RUN;

DATA ANALYSIS2;
  SET SASDATA.m_77_caronsson_niddk_30may2014_2;
RUN;

DATA ANALYSIS3;
  SET SASDATA.m_77_caronsson_niddk_30may2014_3;
RUN;

/***********/
/* Table 1 */
/***********/
TITLE2 'Table 1';
PROC FREQ DATA=ANALYSIS2;
  TABLE FEMALE BIRTHYEAR HLARG;
RUN;

PROC FREQ DATA=ANALYSIS1;
  TABLE female birthyear hlarg;
  WHERE OUTCOME=1;
RUN;

PROC MEANS DATA=ANALYSIS1 MEDIAN P25 P75;
  VAR mtime_postga;
  CLASS FEMALE;
  WHERE OUTCOME=1;
RUN;

PROC MEANS DATA=ANALYSIS1 MEDIAN P25 P75;
  VAR mtime_postga;
  CLASS BIRTHYEAR;
  WHERE OUTCOME=1;
RUN;

PROC MEANS DATA=ANALYSIS1 MEDIAN P25 P75;
  VAR mtime_postga;
  CLASS HLARG;
  WHERE OUTCOME=1;
  FORMAT HLARG HLA_CLPSD.;
```

```
RUN;

PROC NPAR1WAY DATA=ANALYSIS1 WILCOXON;
  VAR MTIME_POSTGA;
  CLASS FEMALE;
  WHERE OUTCOME=1;
RUN;

PROC NPAR1WAY DATA=ANALYSIS1 WILCOXON;
  VAR MTIME_POSTGA;
  CLASS BIRTHYEAR;
  WHERE OUTCOME=1 AND BIRTHYEAR ^IN(2004,2010);
RUN;

PROC NPAR1WAY DATA=ANALYSIS1 WILCOXON;
  VAR MTIME_POSTGA;
  CLASS HLARG;
  WHERE OUTCOME=1;
  FORMAT HLARG HLA_CLPSD.;
RUN;

/***********/
/* Table 2 */
/***********/
TITLE2 'Table 2';
PROC MEANS DATA=ANALYSIS1 NMISS MEDIAN P25 P75;
  VAR mbrst mexbrst mgluten mwheat;
  CLASS OUTCOME;
RUN;

PROC PHREG DATA=ANALYSIS1;
  MODEL TIME*OUTCOME(0) = mbrst /rl;
  STRATA NSUBJ;
RUN;

PROC PHREG DATA=ANALYSIS1;
  MODEL TIME*OUTCOME(0) = mexbrst /rl;
  STRATA NSUBJ;
RUN;

PROC PHREG DATA=ANALYSIS1;
  MODEL TIME*OUTCOME(0) = mgluten /rl;
  STRATA NSUBJ;
RUN;

PROC PHREG DATA=ANALYSIS1;
  MODEL TIME*OUTCOME(0) = mwheat /rl;
  STRATA NSUBJ;
RUN;
```

```
PROC MEANS DATA=ANALYSIS3 MEDIAN P25 P75;
  VAR TOT_ENE;
  CLASS OUTCOME;
RUN;

PROC PHREG DATA=ANALYSIS3;
  MODEL TIME*OUTCOME(0) = tot_ene /rl;
  STRATA NSUBJ;
RUN;

/***********/
/* Table 3 */
/***********/
TITLE2 'Table 3';
PROC MEANS DATA=ANALYSIS1 MEDIAN P25 P75;
  VAR sglut last1glut;
  CLASS OUTCOME;
RUN;

PROC PHREG DATA=ANALYSIS1;
  MODEL TIME*OUTCOME(0) = sglut /rl;
  STRATA NSUBJ;
RUN;

PROC PHREG DATA=ANALYSIS1;
  MODEL TIME*OUTCOME(0) = last1glut /rl;
  STRATA NSUBJ;
RUN;

/***********/
/* Table 4 */
/***********/
TITLE2 'Table 4';
PROC MEANS DATA=ANALYSIS1 MEDIAN P25 P75;
  VAR LAST1GLUT;
  CLASS LASTMONTH OUTCOME;
  WHERE LAST1GLUT^=.;
RUN;

PROC SORT DATA=ANALYSIS1;
  BY LASTMONTH;
RUN;

PROC PHREG DATA=ANALYSIS1;
  MODEL TIME*OUTCOME(0) = LAST1GLUT /rl;
  STRATA NSUBJ;
  BY LASTMONTH;
RUN;
```