Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) Pub67 Swartling

Prepared by Allyson Mateja
IMS Inc.
3901 Calverton Blvd, Suite 200 Calverton, MD 20705
April 24, 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_67_uswartling_niddk_30apr2013.sas7bdat" dataset.

4 Statistical Methods

Analyses were performed to duplicate results for the data published by Swartling et al [1] in The Journal of Empirical Research on Human Research Ethics in 2016. To verify the integrity of the dataset, descriptive statistics were computed.

5 Results

For Table 1 in the publication [1], Mothers and Fathers Risk Perception Accuracy at 3 (Mothers Only), 6, 15, and 27 Months by GP/FDR Status, Table A lists the variables that were used in the replication and Table B compares the results calculated from the archived data file 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 M67 data files to be distributed are a true copy of the study data.

7 References

[1] Swartling, U., Lynch, K., Smith, L., and Johnson, S.B.. "Parental Estimation of Their Child's Increased Type 1 Diabetes Risk During the First 2 Years of Participation in an International Observational Study: Results from the TEDDY study". The Journal of Empirical Research on Human Research Ethics (2016) 1-9.

Table A: Variables used to replicate Table 1**:** Mothers and Fathers Risk Perception Accuracy at 3 (Mothers Only), 6, 15, and 27 Months by GP/FDR Status

Table Variable	**Dataset Variable**
FDR/GP	fdr
Mother's risk at 3 months	develop
Mother's risk at 6 months	devlop_six
Mother's risk at 15 months	develop_fifteen
Mother's risk at 27 months	develop_twoseven
Father's risk at 6 months	develop_dad_six
Father's risk at 15 months	develop_dad_fifteen
Father's risk at 27 months	develop_dad_twoseven

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

Parent	FDR/GP	Age/Study visit month	n Manuscript	n DSIC	Diff.	Much lower % Manuscript	Much lower % DSIC	Diff.
Mothers	GP	3*	5,248	5,248	0	3.0	3.4	0.4
		6	5,130	5,130	0	3.0	3.0	0.0
		15	4,896	4,896	0	3.1	3.1	0.0
		27*	4,852	4,652	200	2.5	2.5	0.0
	FDR	3	655	655	0	0.0	0.0	0.0
		6	639	639	0	0.5	0.5	0.0
		15	601	601	0	0.8	0.8	0.0
		27	549	549	0	0.5	0.6	0.1
Fathers	GP	6	4,659	4,659	0	4.4	4.4	0.0
		15	4,414	4,414	0	3.8	3.8	0.0
		27	4,134	4,134	0	3.3	3.3	0.0
	FDR	6	588	588	0	1.0	1.0	0.0
		15	553	553	0	1.3	1.3	0.0
		27	489	489	0	0.6	0.6	0.0

Parent	FDR/GP	Age/Study visit month	Somewhat lower % Manuscript	Somewhat lower % DSIC	Diff.	About the same % Manuscript	About the same % DSIC	Diff.
Mothers	GP	3*	5.5	6.0	0.5	31.2	29.1	2.1
		6	5.5	5.5	0.0	31.2	31.2	0.0
		15	4.8	4.8	0.0	32.7	32.7	0.0
		27	4.6	4.6	0.0	32.8	32.8	0.0
	FDR	3	0.9	0.9	0.0	10.4	10.4	0.0
		6	0.6	0.6	0.0	11.7	11.7	0.0
		15	0.5	0.5	0.0	14.3	14.3	0.0
		27	0.5	0.6	0.1	12.8	12.8	0.0
Fathers	GP	6	7.0	7.0	0.0	41.9	41.9	0.0
		15	6.2	6.2	0.0	42.8	42.8	0.0
		27	6.2	6.2	0.0	42.3	42.3	0.0
	FDR	6	2.2	2.2	0.0	21.8	21.8	0.0
		15	2.4	2.4	0.0	23.9	23.9	0.0
		27	2.2	2.3	0.1	23.1	23.1	0.0

Parent	FDR/GP	Age/Study visit month	Inaccurate estimation % Manuscript	Inaccurate estimation % DSIC	Diff.
Mothers	GP	3*	38.7	38.5	0.2
		6	39.7	39.7	0.0
		15	40.5	40.5	0.0
		27	40.0	40.0	0.0
	FDR	3	11.3	11.3	0.0
		6	12.8	12.8	0.0
		15	15.6	15.6	0.0
		27	13.8	13.8	0.0
Fathers	GP	6	53.2	53.2	0.0
		15	52.8	52.8	0.0
		27	51.8	51.8	0.0
	FDR	6	25.0	25.0	0.0
		15	27.5	27.5	0.0
		27	26.0	26.0	0.0

Parent	FDR/GP	Age/Study visit month	Somewhat higher % Manuscript	Somewhat higher % DSIC	Diff.	Much higher % Manuscript	Much higher % DSIC	Diff.
Mothers	GP	3*	57.3	55.5	1.8	3.0	6.1	3.1
		6	57.3	57.3	0.0	3.0	3.0	0.0
		15	56.9	56.9	0.0	2.6	2.6	0.0
		27	57.1	57.1	0.0	3.0	3.0	0.0
	FDR	3*	67.1	62.0	5.1	20.0	26.7	6.7
		6	67.1	67.1	0.0	20.0	20.0	0.0
		15	70.4	70.4	0.0	14.0	14.0	0.0
		27	71.8	71.8	0.0	14.4	14.4	0.0
Fathers	GP	6	44.2	44.2	0.0	2.6	2.6	0.0
		15	45.0	45.0	0.0	2.3	2.3	0.0
		27	46.2	46.2	0.0	2.0	2.0	0.0
	FDR	6	58.8	58.8	0.0	16.2	16.2	0.0
		15	59.7	59.7	0.0	12.8	12.8	0.0
		27	63.2	63.2	0.0	10.8	10.8	0.0

Parent	FDR/GP	Age/Study visit month	Accurate estimation % Manuscript	Accurate estimation % DSIC	Diff.
Mothers	GP	3*	61.3	61.5	0.2
		6	60.3	60.3	0.0
		15	59.5	59.5	0.0
		27	60.0	60.0	0.0
	FDR	3	88.7	88.7	0.0
		6	87.2	87.2	0.0
		15	84.4	84.4	0.0
		27	86.2	86.2	0.0
Fathers	GP	6	46.8	46.8	0.0
		15	47.2	47.2	0.0
		27	48.2	48.2	0.0
	FDR	6	75.0	75.0	0.0
		15	72.5	72.5	0.0
		27	74.0	74.0	0.0

*Note that the values published in the manuscript are typos, and those calculated in the DSIC are correct.

Attachment A: SAS Code

```
*** TEDDY M67 DSIC;
*** Programmer: Allyson Mateja;
*** Date: 3/2/2017;

libname sas_data '/prj/niddk/ims_analysis/TEDDY/private_orig_data/m_67_uswartling_niddk_submission';

data teddym67;
	set sas_data.m_67_uswartling_niddk_30apr2013;

proc format;
	value fdrf 0 = 'GP'
		   1 = 'FDR';
	value riskf 1 = 'Much lower'
		    2 = 'Somewhat lower'
		    3 = 'About the same'
		    4 = 'Somewhat higher'
		    5 = 'Much higher';

proc contents data = teddym67;

data table1;
	length estimation_mom_3 estimation_mom_6 estimation_mom_15 estimation_mom_27
		estimation_dad_6 estimation_dad_15 estimation_dad_27 $15.;
	set teddym67;
	if develop in (1,2,3) then estimation_mom_3 = 'Inaccurate';
	else if develop in (4,5) then estimation_mom_3 = 'Accurate';
	if develop_six in (1,2,3) then estimation_mom_6 = 'Inaccurate';
	else if develop_six in (4,5) then estimation_mom_6 = 'Accurate';
	if develop_fifteen in (1,2,3) then estimation_mom_15 = 'Inaccurate';
	else if develop_fifteen in (4,5) then estimation_mom_15 = 'Accurate';
	if develop_twoseven in (1,2,3) then estimation_mom_27 = 'Inaccurate';
	else if develop_twoseven in (4,5) then estimation_mom_27 = 'Accurate';
	if develop_dad_six in (1,2,3) then estimation_dad_6 = 'Inaccurate';
	else if develop_dad_six in (4,5) then estimation_dad_6 = 'Accurate';
	if develop_dad_fifteen in (1,2,3) then estimation_dad_15 = 'Inaccurate';
	else if develop_dad_fifteen in (4,5) then estimation_dad_15 = 'Accurate';
	if develop_dad_twoseven in (1,2,3) then estimation_dad_27 = 'Inaccurate';
	else if develop_dad_twoseven in (4,5) then estimation_dad_27 = 'Accurate';
	if exclude=0;

proc freq data = table1;
	tables fdr*develop fdr*estimation_mom_3 fdr*develop_six fdr*estimation_mom_6 fdr*develop_fifteen fdr*estimation_mom_15
fdr*develop_twoseven fdr*estimation_mom_27
		fdr*develop_dad_six fdr*estimation_dad_6 fdr*develop_dad_fifteen fdr*estimation_dad_15 fdr*develop_dad_twoseven
fdr*estimation_dad_27 /nocol nopercent;
	format fdr fdrf.
		develop develop_sex develop_dad_six develop_fifteen develop_dad_fifteen develop_twoseven develop_dad_twoseven riskf.;
```

```
title 'Table 1';
```