Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) Study: Johnson Data File

Prepared by

RTI International
3040 Cornwallis Road
Research Triangle Park, NC 27709-2194
October, 2012

Revision History

Version	Author / Title	Date	Comments
1.0	Norma Pugh	October 8, 2012	Original

Table of 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 on a first (or second) exercise 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 and other items. Experience suggests that most discrepancies can ordinarily be resolved by consulting with the study data coordinating center (DCC); however, this process is labor-intensive for both DCC and Repository staff. Therefore, it is 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 is an international, longitudinal, observational study that identifies young infants at increased genetic risk for type 1 diabetes (T1DM). The study, conducted in the countries of Finland, Germany, Sweden and the United States, aims to identify environmental triggers of T1DM in genetically at-risk children through observation and data collection over a 15 year time period [2].

The TEDDY Study is a particularly demanding protocol, including blood draws, stool sample collections, diet records, interviews, and questionnaires. The visit schedule is, likewise, demanding. Johnson et al describe family characteristics that predicted withdrawal during the first year of the TEDDY study, among families with no immediate family history of type 1 diabetes. Such research is useful in the design of similar studies and may influence future efforts to retain families in such studies [1].

3 Archived Datasets

All SAS data files, as provided by the Data Coordinating Center (DCC), are located in the TEDDY Data folder in the Official Archive. For this replication, all variables were taken from the SAS data file,

M_10_SJOHNSON_NIDDK_FINAL, located in the \\TEDDY\TEDDY Official Archive_v2\Teddy_DATA sub-folder.

4 Statistical Methods

To verify the integrity of the M_10_JOHNSON_NIDDK_FINAL data file housed at the repository, analyses were performed to duplicate results for the data published by Johnson et al [1] in Pediatric Diabetes in May 2010. Results are presented in tables A-C and the SAS code for our analysis is included in Attachment 1.

5 Results

Table 1 in the publication [1], Logistic regression results for the sample with no missing data and the total sample with missing data imputed, reports on the final logistic regression models. Our Table A lists the variables we used in our replication. Table 1 in the publication [1] is extensive, including beta point estimates, standard errors, p-values, odds ratios, and 95% confidence intervals for the sample with no missing data; and beta point estimates, standard errors, and p-values for the sample with missing data imputed, using multiple imputation. Results are included for the following predictor variables: country, child sex, maternal age, smoking, alcohol consumption in the last trimester, employment, dad participation, risk perception, the State Anxiety Inventory score, the State Anxiety Inventory score – by-risk perception interaction, and missing data points. For simplicity of presentation, this extensive table is not included in this Integrity Check. The Integrity Check replication for the Table 1compares the results calculated from the archived data file to the published results, and includes only the sample with no missing data. Sample sizes, odds ratios and 95% confidence intervals were compared. Sample size matches exactly. Odds ratios and 95% confidence intervals are similar to published results.

Table 2 in the publication [1], Characteristics of TEDDY general population families associated with study withdrawal in the first year after enrollment, reports on the descriptive information for each of the significant predictors for withdrawal. Our Table B lists the variables we used in our replication and Table C compares the results calculated from the archived data file to the results published in Table 2. Again, the results of the replication are similar to published results.

6 Conclusions

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

7 References

1. Johnson SB, Lee H-S, Baxter J, Lernmark B, Roth R, Simell T. The Environmental Determinants of Diabetes in the Young (TEDDY) Study: predictors of early study withdrawal among participants with no family history of type 1 diabetes. Pediatr Diabetes. 2011 May;12(3 Pt 1):165-71. doi: 10.1111/j.1399-5448.2010.00686.x. Epub 2010 Oct 28.
2. Barbro Lernmark, et al. Enrollment experiences in a pediatric longitudinal observational study: The Environmental Determinants of Diabetes in the Young (TEDDY) study; Contemporary Clinical Trials 32(2011) 517-523.

Table A: Variables Used to Replicate Table 1, Logistic regression results for the sample with no missing data and the total sample with missing data imputed

Table Variable	Variables Used in Replication
Country	country
Child sex	sex
Maternal age	maternal_age
Smoking	smoker
Alcohol consumption at third trimester	nalc3
Employment status	worked
Dad participation in TEDDY	dadactive
Risk perception	anydev
State Anxiety Inventory Score	stai
Data missing	smoker, nalc3, worked, anydev, stai

*All variables taken from dataset M_10_JOHNSON_NIDDK_FINAL.

Table B: Variables Used to Replicate Table 2, Characteristics of TEDDY general population families associated with study withdrawal in the first year after enrollment

Table Variable	Variables Used in Replication
Study withdrawal status	earlydrop
Country	country
Child sex	sex
Maternal age	maternal_age
Smoking	smoker
Alcohol consumption at third trimester	nalc3
Employment status	worked
Dad participation in TEDDY	dadactive
Risk perception	anydev
State Anxiety Inventory Score	stai
Data missing	smoker, nalc3, worked, anydev, stai

*All variables taken from dataset M_10_JOHNSON_NIDDK_FINAL.

Table C: Comparison of Values Computed in Integrity Check to Reference Article Table 2 Values

Characteristic	Actives			Withdrawals			Total sample		
	Johnson	Integrity Check	Diff	Johnson	Integrity Check	Diff	Johnson	Integrity Check	Diff
Sample size, n	2994	2994	0	763	763	0	3757	3757	0
Country, n (%)									
Finland	747 (84)	747 (84)	0	140 (16)	140 (16)	0	887	887	0
Germany	106 (75)	106 (75)	0	36 (25)	36 (25)	0	142	142	0
Sweden	1052 (82)	1052 (82)	0	231 (18)	231 (18)	0	1283	1283	0
USA	1089 (75)	1089 (75)	0	356 (25)	356 (25)	0	1445	1445	0
Child sex, n (%)									
Male	1538 (81)	1538 (81)	0	352 (19)	352 (19)	0	1890	1890	0
Female	1456 (78)	1456 (78)	0	411 (22)	411 (22)	0	1867	1867	0
Maternal age (yr), mean (sd)	30.8 (5.0)	30.8 (5.0)	0	28.5 (5.7)	28.5 (5.7)	0	30.4 (5.2)	30.4 (5.2)	0
Maternal lifestyle behaviors during pregnancy									
Smoking, n (%)									
Smoked	296 (63)	296 (63)	0	171 (37)	171 (37)	0	467	467	0
Did not smoke	2602 (84)	2602 (84)	0	510 (16)	510 (16)	0	3112	3112	0
Data missing	96 (54)	96 (54)	0	82 (46)	82 (46)	0	178	178	0
Alcohol consumption at third trimester, n (%)									
Alcohol 1-2 times per month	474 (87)	474 (87)	0	72 (13)	72 (13)	0	546	546	0
Alcohol ≥3 times per month	105 (89)	105 (89)	0	13 (11)	13 (11)	0	118	118	0
No alcohol	2359 (79)	2359 (79)	0	609 (21)	609 (21)	0	2968	2968	0
Data missing	56 (45)	56 (45)	0	69 (55)	69 (55)	0	125	125	0
Employment status, n (%)									
Worked all three trimesters	1418 (85)	1418 (85)	0	251 (15)	251 (15)	0	1669	1669	0
Reduced work/quit/no work	1426 (77)	1426 (77)	0	417 (23)	417 (23)	0	1843	1843	0
Data missing	150 (61)	150 (61)	0	95 (39)	95 (39)	0	245	245	0

Characteristic	Actives			Withdrawals			Total sample		
	Johnson	Integrity Check	Diff	Johnson	Integrity Check	Diff	Johnson	Integrity Check	Diff
Dad participation in TEDDY, n (%)									
Participated	2813 (82)	2813 (82)	0	624 (18)	624 (18)	0	3437	3437	0
Did not participate	181 (57)	181 (57)	0	139 (43)	139 (43)	0	320	320	0
Maternal reactions to child's increased T1DM risk									
Risk perception, n (%)									
Accurate	1809 (84)	1809 (84)	0	355 (16)	355 (16)	0	2164	2164	0
Underestimate	1132 (77)	1132 (77)	0	343 (23)	343 (23)	0	1475	1475	0
Data missing	53 (45)	53 (45)	0	65 (55)	65 (55)	0	118	118	0
State Anxiety Inventory Score, mean (sd)									
Total sample	38.7 (9.7)	38.7 (9.7)	0	40.8 (10.6)	40.8 (10.6)	0	39.1 (9.9)	39.1 (9.9)	0
Risk perception: accurate	38.8 (10.2)	38.8 (9.3)	0 (-0.9)	41.7 (10.4)	41.7 (10.4)	0	39.3 (9.6)	39.3 (9.6)	0
Risk perception: underestimate	38.4 (10.2)	38.4 (10.2)	0	39.9 (10.8)	39.9 (10.8)	0	38.8 (10.4)	38.8 (10.4)	0
Data missing, n (%)	46 (42)	46 (42)	0	63 (58)	63 (58)	0	109	109	0
Missing data, n (%)									
≤1 missing data points	2944 (81)	2944 (81)	0	695 (19)	695 (19)	0	3639	3639	0
>1 missing data points	50 (42)	50 (42)	0	68 (58)	68 (58)	0	118	118	0

Attachment A: SAS Code

```
options errorabend nofmterr mprint;
/*******************************************************************************/
/*
/* Program: R:\05_Users\Norma\TEDDY\JohnsonPaper\table2.sas
/* Author:  Norma Pugh
/* Date:    September 2012
/* Purpose: Replicate table 2 results.
/*
/*******************************************************************************/
/* DATA SOURCE */
libname data
'\\samba1.rtp.rti.org\NIDDK\03_Data_And_Tools\Studies\TEDDY\Delivery_from_DCC\20120517_from_Steven_Fiske';

/**********************/
/* ADDITIONAL FORMATS */
/**********************/
proc format;

 value country  1 = "1 = U.S."
                2 = "2 = Finland"
                3 = "3 = Germany"
                4 = "4 = Sweden";

 value alc     0 = "0 = No alcohol"
               1 = "1 = 1-2/month"
               2 = "2 = >2 month"
               . = "Data missing";

 value employ  0 = "0 = Reduced/quit/didn't work"
               1 = "1 = Worked all 3 trimesters"
               . = "Data missing";

 value dad    0 = "0 = Did not participate"
              1 = "1 = Participated";

 value risk   0 = "0 = Underestimate"
              1 = "1 = Accurate"
              . = "Data missing";

 value yn     0 = "0 = No"
              1 = "1 = Yes"
              . = "Data missing";

 value mdata  1 = "1: <=1 missing data points"
              2 = "2: >1 missing data point";
run;


/************/
/* GET DATA */
/************/
data logit; set data.m_10_sjohnson_niddk_final;
 /* FDR participants excluded b/c study w/d is rare in this popn */
 if fdr=0 & earlydrop in(0,1);

 /* Create var: "Missing STAI score" */
 if stai=. then missstai=1; else missstai=0;

 /* Create var: "Missing data points" */
 nummiss = nmiss(smoker,nalc3,worked,anydev,stai);

 if nummiss<=1 then mdata=1; else if nummiss>1 then mdata=2;
run;
```

```
/*****************************/
/* REPLICATE ANALYSIS RESULTS */
/*****************************/
proc logistic data=logit descending;
 class country (ref=first) sex (ref=first) smoker (ref=first) nalc3 (ref=first) worked (ref=first)
dadactive (ref=first) anydev (ref=first);
 model earlydrop = country sex maternal_age smoker nalc3 worked dadactive anydev stai stai*anydev;
run;
```

```
options errorabend nofmterr mprint;
/*****************************************************************************************/
/*
/* Program: R:\05_Users\Norma\TEDDY\JohnsonPaper\table2.sas
/* Author:  Norma Pugh
/* Date:    September 2012
/* Purpose: Replicate table 2 results.
/*
/*****************************************************************************************/
/* DATA SOURCE */
libname data
'\\samba1.rtp.rti.org\NIDDK\03_Data_And_Tools\Studies\TEDDY\Delivery_from_DCC\20120517_from_Steven_Fiske';

/**********************/
/* ADDITIONAL FORMATS */
/**********************/
proc format;

 value country 1 = "1 = U.S."
               2 = "2 = Finland"
               3 = "3 = Germany"
               4 = "4 = Sweden";

 value alc     0 = "0 = No alcohol"
               1 = "1 = 1-2/month"
               2 = "2 = >2 month"
               . = "Data missing";

 value employ  0 = "0 = Reduced/quit/didn't work"
               1 = "1 = Worked all 3 trimesters"
               . = "Data missing";

 value dad    0 = "0 = Did not participate"
              1 = "1 = Participated";

 value risk   0 = "0 = Underestimate"
              1 = "1 = Accurate"
              . = "Data missing";

 value yn     0 = "0 = No"
              1 = "1 = Yes"
              . = "Data missing";

 value mdata  1 = "1: <=1 missing data points"
              2 = "2: >1 missing data point";
run;


/************/
/* GET DATA */
/************/
data demog; set data.m_10_sjohnson_niddk_final;
 /* FDR participants excluded b/c study w/d is rare in this popn */
 if fdr=0 & earlydrop in(0,1);


/* Create var: "Missing STAI score" */
 if stai=. then missstai=1; else missstai=0;

 /* Create var: "Missing data points" */
 nummiss = nmiss(smoker,nalc3,worked,anydev,stai);

 if nummiss<=1 then mdata=1; else if nummiss>1 then mdata=2;
run;


/*****************************/
/* REPLICATE ANALYSIS RESULTS */
/*****************************/
```

```
%macro frq(var,fmt,title);
proc freq data=demog; tables earlydrop*&var / missing; &fmt; title"Frequency Counts: &title"; run;
%mend frq;

proc sort data=demog; by earlydrop; run;

%macro mean_(var,title);
title"Means: &title";
proc means data=demog n mean std; by earlydrop; var &var; run;
run;
%mend mean_;

%frq(country,%str(format country country.),%str(Country));
%frq(sex,,%str(Child gender));
%mean_(maternal_age,%str(Maternal Age));
proc means data=demog n mean std; var maternal_age; run;
%frq(smoker,%str(format smoker yn.),%str(Smoking));
%frq(nalc3,%str(format nalc3 alc.),%str(Alcohol consumption at third trimester));
%frq(worked,%str(format worked employ.),%str(Employment status));
%frq(dadactive,%str(format dadactive dad.),%str(Dad participation in TEDDY));
%frq(anydev,%str(format anydev risk.),%str(Risk perception));
%mean_(stai,%str(State Anxiety Inventory Score));
proc means data=demog n mean std; var stai; run;
title'State Anxiety Inventory Score: accurate';
 proc means data=demog(where=(anydev=1)) n mean std; by earlydrop; var stai; run;
 proc means data=demog(where=(anydev=1)) n mean std; var stai; run;
title'State Anxiety Inventory Score: underestimate';
 proc means data=demog(where=(anydev=0)) n mean std; by earlydrop; var stai; run;
 proc means data=demog(where=(anydev=0)) n mean std; var stai; run;
%frq(missstai,%str(format missstai yn.),%str(STAI score missing));
%frq(mdata,%str(format mdata mdata.),%str(Missing data points));
```