Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) Pub49 Johnson

Prepared by Allyson Mateja
IMS Inc.
3901 Calverton Blvd, Suite 200 Calverton, MD 20705
August 25, 2016

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_49_sjohnson_niddk_30june2011" dataset.

4 Statistical Methods

Analyses were performed to duplicate results for the data published by Johnson et al [1] in the Journal of Clinical Epidemiology in 2014. To verify the integrity of the dataset, descriptive statistics were computed.

5 Results

For Table 1 in the publication [1], Percent of general population participants withdrawing from TEDDY in the first year, before and after high risk for early withdrawal score notification with tailored intervention was initiated, 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 an exact match to the published results.

6 Conclusions

The NIDDK repository is confident that the TEDDY M49 data files to be distributed are a true copy of the study data.

7 References

[1] Johnson, S.B., Lynch, K.F., Lee, H., Smith, L., Baxter, J., Lenmark, B., Roth, R., Simell, T., and the TEDDY study Group. "At high risk for early withdrawal: using a cumulative risk model to increate retention in the first year of the TEDDY study". Journal of Clinical Epidemiology 67 (2014) 609-611.

Table A: Variables used to replicate Table 1: Percent of general population participants withdrawing from TEDDY in the first year, before and after high risk for early withdrawal score notification with tailored intervention was initiated

Table Variable	**Dataset Variable**
TEDDY Cohort	intervention_cohort
Risk for early withdrawal	high_risk_score
Withdrawn	early_withdrawal
Country	european

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

All participants							
Cohort	Risk for early withdrawal	Total N Manuscript	Total N DSIC	Diff.	% Withdrawn Manuscript	% Withdrawn DSIC	Diff.
Comparison (C) No risk notification and not tailored intervention	Low (score < 4)	1,053	1,053	0	4.8	4.8	0
	High (score ≥ 4)	426	426	0	12.7	12.7	0
Intervention (I) Risk notification plus tailored intervention	Low (score < 4)	1,204	1,204	0	3.7	3.7	0
	High (score ≥ 4)	524	524	0	4.4	4.4	0

US participants							
Cohort	Risk for early withdrawal	Total N Manuscript	Total N DSIC	Diff.	% Withdrawn Manuscript	% Withdrawn DSIC	Diff.
Comparison (C) No risk notification and not tailored intervention	Low (score < 4)	485	485	0	4.3	4.3	0
	High (score ≥ 4)	248	248	0	11.7	11.7	0
Intervention (I) Risk notification plus tailored intervention	Low (score < 4)	552	552	0	2.0	2.0	0
	High (score ≥ 4)	294	294	0	3.7	3.7	0

European participants							
Cohort	Risk for early withdrawal	Total N Manuscript	Total N DSIC	Diff.	% Withdrawn Manuscript	% Withdrawn DSIC	Diff.
Comparison (C) No risk notification and not tailored intervention	Low (score < 4)	568	568	0	5.3	5.3	0
	High (score ≥ 4)	178	178	0	14.0	14.0	0
Intervention (I) Risk notification plus tailored intervention	Low (score < 4)	652	652	0	5.1	5.1	0
	High (score ≥ 4)	230	230	0	5.2	5.2	0

Attachment A: SAS Code

```
**** TEDDY M49 DSIC;
****
**** Programmer: Allyson Mateja;
**** Date: August 5, 2016;

title '/prj/niddk/ims_analysis/TEDDY/prog_initial_analysis/teddy_integrity_check_m49.sas';
title2 ' ';

proc format;
       value riskf 0 = 'Low (score < 4)'
                   1 = 'High (score >= 4)';

       value cohortf 0 = 'Comparison'
                     1 = 'Intervention';

       value yesnof 0 = 'No'
                    1 = 'Yes';

libname privorig '/prj/niddk/ims_analysis/TEDDY/private_orig_data/M_49_SJohnson_NIDDK_Submission/';

data m49_data;
       set privorig.m_49_sjohnson_niddk_30june2011;

proc contents data = m49_data;

proc freq data = m49_data;
       tables intervention_cohort*high_risk_score /list missing;
       format intervention_cohort cohortf. high_risk_score riskf.;
       title3 'Table 1 - All Participants - Total N';

proc freq data = m49_data;
       tables early_withdrawal /list missing;
       format early_withdrawal yesnof.;
       where intervention_cohort = 0 and high_risk_score = 0;
       title3 'Table 1 - All Participants, % Withdrawn, Comparison Cohort, Low risk';

proc freq data = m49_data;
       tables early_withdrawal /list missing;
       format early_withdrawal yesnof.;
       where intervention_cohort = 0 and high_risk_score = 1;
       title3 'Table 1 - All Participants, % Withdrawn, Comparison Cohort, High risk';

proc freq data = m49_data;
       tables early_withdrawal /list missing;
       format early_withdrawal yesnof.;
       where intervention_cohort = 1 and high_risk_score = 0;
       title3 'Table 1 - All Participants, % Withdrawn, Intevention Cohort, High risk';
```

```
proc freq data = m49_data;
       tables early_withdrawal /list missing;
       format early_withdrawal yesnof.;
       where intervention_cohort = 1 and high_risk_score = 1;
       title3 'Table 1 - All Participants, % Withdrawn, Intervention Cohort, High risk';

proc freq data = m49_data;
       tables intervention_cohort*high_risk_score /list missing;
       where european = 0;
       format intervention_cohort cohortf. high_risk_score riskf.;
       title3 'Table 1 - US Participants - Total N';

proc freq data = m49_data;
       tables early_withdrawal /list missing;
       format early_withdrawal yesnof.;
       where intervention_cohort = 0 and high_risk_score = 0 and european = 0;
       title3 'Table 1 - US Participants, % Withdrawn, Comparison Cohort, Low risk';

proc freq data = m49_data;
       tables early_withdrawal /list missing;
       format early_withdrawal yesnof.;
       where intervention_cohort = 0 and high_risk_score = 1 and european = 0;
       title3 'Table 1 - US Participants, % Withdrawn, Comparison Cohort, High risk';

proc freq data = m49_data;
       tables early_withdrawal /list missing;
       format early_withdrawal yesnof.;
       where intervention_cohort = 1 and high_risk_score = 0 and european = 0;
       title3 'Table 1 - US Participants, % Withdrawn, Intevention Cohort, High risk';

proc freq data = m49_data;
       tables early_withdrawal /list missing;
       format early_withdrawal yesnof.;
       where intervention_cohort = 1 and high_risk_score = 1 and european = 0;
       title3 'Table 1 - US Participants, % Withdrawn, Intervention Cohort, High risk';

proc freq data = m49_data;
       tables intervention_cohort*high_risk_score /list missing;
       where european = 1;
       format intervention_cohort cohortf. high_risk_score riskf.;
       title3 'Table 1 - European Participants - Total N';

proc freq data = m49_data;
       tables early_withdrawal /list missing;
       format early_withdrawal yesnof.;
       where intervention_cohort = 0 and high_risk_score = 0 and european = 1;
       title3 'Table 1 - European Participants, % Withdrawn, Comparison Cohort, Low risk';

proc freq data = m49_data;
       tables early_withdrawal /list missing;
```

```
        format early_withdrawal yesnof.;
        where intervention_cohort = 0 and high_risk_score = 1 and european = 1;
        title3 'Table 1 - European Participants, % Withdrawn, Comparison Cohort, High risk';

proc freq data = m49_data;
        tables early_withdrawal /list missing;
        format early_withdrawal yesnof.;
        where intervention_cohort = 1 and high_risk_score = 0 and european = 1;
        title3 'Table 1 - European Participants, % Withdrawn, Intevention Cohort, High risk';

proc freq data = m49_data;
        tables early_withdrawal /list missing;
        format early_withdrawal yesnof.;
        where intervention_cohort = 1 and high_risk_score = 1 and european = 1;
        title3 'Table 1 - European Participants, % Withdrawn, Intervention Cohort, High risk';
```