Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) Pub66 Hadley

Prepared by Allyson Mateja
IMS Inc.
3901 Calverton Blvd, Suite 200 Calverton, MD 20705
February 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_66_dhadley_niddk_31jan2013_1.sas7bdat" and "m_66_dhadley_niddk_31jan2013_2.sas7bdat" datasets.

4 Statistical Methods

Analyses were performed to duplicate results for the data published by Hadley et al [1] in The American Journal of Gastroenterology in 2015. To verify the integrity of the dataset, descriptive statistics were computed.

5 Results

For Table 2 in the publication [1], Demographic summary of participating children in the TEDDY cohort, 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 2. The results of the replication are an exact match to the published results.

For Table 3 in the publication [1], Proportions of nested case-control groups and replication cohort, at the time of data freeze, Table C lists the variables that were used in the replication and Table D compares the results calculated from the archived data file to the results published in Table 3. The results of the replication are almost an exact match to the published results.

For Table 6 in the publication [1], Frequency of HLA-DPB1*04:01 allele among the four common HLA-DR-DQ risk groups, Table E lists the variables that were used in the replication and Table F compares the results calculated from the archived data file to the results published in Table 6. The results of the replication are close to the published results.

6 Conclusions

The NIDDK repository is confident that the TEDDY M66 data files to be distributed are a true copy of the study data.

7 References

[1] Hadley, D., Hagopian, W., Liu, E., She, J., Simell, O., Akolkar, B., Ziegler, A., Rewers, M., Krischer, J.P., Chen, W., Onengut-Gumuscu, S., Bugawan, T.L., Rich, S.S., Erlich, H., Agardh, D., and the TEDDY study group. "HLA-DPB1*04:01 Protects Genetically Susceptible Children from Celiac Disease Autoimmunity in the TEDDY Study”. The American Journal of Gastroenterology (2015) 110:915-920.

Table A: Variables used to replicate Table 2**:** Demographic summary of participating children in the TEDDY cohort

Table Variable	**dataset.variable**
Country	m_66_dhadley_niddk_31jan2013_2.country
Female	m_66_dhadley_niddk_31jan2013_2.gender
Non-Caucasian	m_66_dhadley_niddk_31jan2013_2.caucasian
Dq2-DR3/DQ2-DR3	m_66_dhadley_niddk_31jan2013_2.hla_category

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

Country	n Manuscript	n DSIC	Diff.	Female (%) Manuscript	Female (%) DSIC	Diff.
Finland	1,833	1,833	0	49.1	49.1	0
Germany	596	596	0	50.2	50.2	0
Sweden	2,525	2,525	0	49.4	49.4	0
USA	3,722	3,722	0	49.4	49.4	0
Total	8,676	8,676	0	49.4	49.4	0

Country	Non-Caucasian (%) Manuscript	Non-Caucasian (%) DSIC	Diff.	DQ2-DR3/DQ2-DR3 (%) Manuscript	DQ2-DR3/DQ2-DR3 (%) DSIC	Diff.
Finland	0.4	0.4	0	14.9	14.9	0
Germany	0	0	0	20.4	20.4	0
Sweden	0	0	0	21.8	21.8	0
USA	28.7	28.7	0	23.4	23.4	0
Total	87.6*	12.4	N/A	20.9	20.9	0

*Note that this published value represents the total number of Caucasians. The value computed in the DSIC reflects the correct percentage of Non-Caucasians.

Table C: Variables used to replicate Table 3: Proportions of nested case-control groups and replication cohort, at the time of data freeze

Table Variable	**dataset.variable**
Cases/Controls	m_66_dhadley_niddk_31jan2013_1.tg_cc
Replication cohort	m_66_dhadley_niddk_31jan2013_1.tg_cc, m_66_dhadley_niddk_31jan2013_2.caucasian, m_66_dhadley_niddk_31jan2013_2.hla_category, m_66_dhadley_niddk_31jan2013_2.hla_dpb1_0401_p, m_66_dhadley_niddk_31jan2013_2.fid, m_66_dhadley_niddk_31jan2013_2.persist_tga, m_66_dhadley_niddk_31jan2013_2.timetotga

Table Variable	dataset.variable
tTGA	m_66_dhadley_niddk_31jan2013_1.tg_cc, m_66_dhadley_niddk_31jan2013_2.persist_tga
Female	m_66_dhadley_niddk_31jan2013_2.gender
Non-Caucasian	m_66_dhadley_niddk_31jan2013_2.caucasian
Dq2-DR3/DQ2-DR3	m_66_dhadley_niddk_31jan2013_2.hla_category

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

	n Manuscript	n DSIC	Diff.	tTGA (%) Manuscript	tTGA (%) DSIC	Diff.	Female (%) Manuscript	Female (%) DSIC	Diff.
Cases	248	248	0	100	100	0	61.3	61.3	0
Controls	248	248	0	0	0	0	61.3	61.3	0
Replication cohort	4,514	4,515	1	9.6	9.6	0	47.34	47.33	0.01

	Caucasian (%) Manuscript	Caucasian (%) DSIC	Diff.	DR3-DQ2/DR3-DQ2 (%) Manuscript	DR3-DQ2/DR3-DQ2 (%) DSIC	Diff.
Cases	95.6	95.6	0	52.8	52.8	0
Controls	97.2	97.2	0	24.2	24.2	0
Replication cohort	100	100	0	20.8	20.9	0.1

Table E: Variables used to replicate Table 6: Frequency of HLA-DPB1*04:01 allele among the four common HLA-DR-DQ risk groups

Table Variable	dataset.variable
HLA Category	m_66_dhadley_niddk_31jan2013_2.hla_category
Number of copies and genotype frequency of HLA-DPB1*04:01	m_66_dhadley_niddk_31jan2013_2.hla_dpb1_0401_p

Table F: Comparison of values computed in integrity check to reference article Table 6 values

HLA Category	0 copies of HLA-DPB1*04:01 Manuscript	0 copies of HLA-DPB1*04:01 DSIC	Diff.	1 copy of HLA-DPB1*04:01 Manuscript	1 copy of HLA-DPB1*04:01 DSIC	Diff.
DR4-DQ8/DR3-DQ2	630 (35.5%)	633 (34.7%)	3 (0.8%)	897 (49.1%)	896 (49.0%)	1 (0.1%)
DR4-DQ8/DR4-DQ8	247 (26.1%)	246 (26.1%)	1 (0%)	463 (49.0%)	463 (49.2%)	0 (0.2%)
DR4-DQ8/DR8-DQ4	247 (30.8%)	253 (31.5%)	6 (0.7%)	398 (49.6%)	394 (49.1%)	4 (0.5%)
DR3-DQ2/DR3-DQ2	398 (42.3%)	401 (42.5%)	3 (0.2%)	425 (45.2%)	425 (45.0%)	0 (0.2%)
Total	1,522 (33.7%)	1,533 (34.0%)	11 (0.3%)	2,183 (48.3%)	2,178 (48.2%)	5 (0.1%)

HLA Category	2 copies of HLA-DPB1*04:01 Manuscript	2 copies of HLA-DPB1*04:01 DSIC	Diff.	Total Manuscript	Total DSIC	Diff.
DR4-DQ8/DR3-DQ2	300 (16.4%)	298 (16.3%)	2 (0.1%)	1,827	1,827	0
DR4-DQ8/DR4-DQ8	235 (24.9%)	232 (24.7%)	3 (0.2%)	945	941	4
DR4-DQ8/DR8-DQ4	157 (19.6%)	156 (19.4%)	1 (0.2%)	802	803	1
DR3-DQ2/DR3-DQ2	117 (12.4%)	118 (12.5%)	1 (0.1%)	940	944	4
Total	809 (17.9%)	804 (17.8%)	5 (0.1%)	4,514	4,515	1

Attachment A: SAS Code

```
**** TEDDY M66 DSIC;
****
**** Programmer: Allyson Mateja;
**** Date: August 5, 2016;

proc format;
        value countryf 1 = 'USA'
                       2 = 'Finland'
                       3 = 'Germany'
                       4 = 'Sweden';
        value genderf 1 = 'M'
                      2 = 'F';
        value ccf 0 = 'Control'
                  1 = 'Case'
                  2 = 'Replication cohort';

title '/prj/niddk/ims_analysis/TEDDY/prog_initial_analysis/teddy_integrity_check_m66.sas';
title2 ' ';

libname privorig '/prj/niddk/ims_analysis/TEDDY/private_orig_data/m_66_dhadley_niddk_submission_01_27_2017/';

data m66_data_part1;
        set privorig.m_66_dhadley_niddk_31jan2013_1;

proc contents data = m66_data_part1;

data m66_data_part2;
        set privorig.m_66_dhadley_niddk_31jan2013_2;

proc contents data = m66_data_part2;

proc sort data = m66_data_part1;
        by maskid;

proc sort data = m66_data_part2;
        by maskid;

data m66_data_part2;
        set m66_data_part2;
        if caucasian = . then caucasian = 0;

proc freq data = m66_data_part2;
        tables country /list missing ;
        format country countryf.;
        title3 'Table 2 - Country';

proc sort data = m66_data_part2;
```

```
       by country;

proc freq data = m66_data_part2;
       tables gender;
       format gender genderf.;
       title3 'Table 2 - Gender';

proc freq data = m66_data_part2;
       tables gender;
       by country;
       format country countryf. gender genderf.;

proc freq data = m66_data_part2;
       tables caucasian;
       title3 'Table 2 - Non-Caucasian';

proc freq data = m66_data_part2;
       tables caucasian;
       by country;
       format country countryf.;

data m66_data_part2;
       set m66_data_part2;
       if hla_category = 9 then dq2_dr3 = 1;
       else if hla_category ne 0 then dq2_dr3 = 0;

proc freq data = m66_data_part2;
       tables dq2_dr3;
       title3 'Table 2 - DQ2-DR3/DQ2-DR3';

proc freq data = m66_data_part2;
       tables dq2_dr3;
       by country;
       format country countryf.;

proc sort data = m66_data_part2;
       by maskid;

data m66_data;
       merge m66_data_part1 (in=val1)
             m66_data_part2 (in=val2);
       by maskid;
       if substr(fid, 1, 1) = 'T' then fid = substr(fid, 6);
       if val2 then output;

proc freq data = m66_data;
       tables tg_cc;

data replication_cohort case_control;
       set m66_data;
```

```
        if tg_cc = . and hla_dpb1_0401_p in ('0', '1', '2') and caucasian = 1 and hla_category in (1,2,4,9) and fid not in ('', ' ')
and persist_tga ne . and timetotga ne . nad timetotga >= 365 then output replication_cohort;
        else if tg_cc ne . then output case_control;

proc sort data = replication_cohort nodupkey;
        by fid;

proc sort data = case_control;
        by tg_cc;

proc freq data = case_control;
        tables tg_cc;
        by tg_cc;
        format tg_cc ccf.;
        title3 'Table 3 - tTGA';

proc freq data = replication_cohort;
        tables persist_tga;

proc freq data = case_control;
        tables gender;
        by tg_cc;
        format tg_cc ccf. gender genderf.;
        title3 'Table 3 - Female';

proc freq data = replication_cohort;
        tables gender;
        format gender genderf.;

proc freq data = case_control;
        tables caucasian;
        by tg_cc;
        format tg_cc ccf.;
        title3 'Table 3 - Caucasian';

proc freq data = replication_cohort;
        tables caucasian;

proc freq data = case_control;
        tables dq2_dr3;
        by tg_cc;
        format tg_cc ccf.;
        title3 'Table 3 - DR3-DQ2/DR3-DQ2';

proc freq data = replication_cohort;
        tables dq2_dr3;

proc freq data = replication_cohort;
        tables hla_category*hla_dpb1_0401_p;
        title3 'Table 6';
```