Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) Pub102 Beyerlein

Prepared by Sabrina Chen
IMS Inc.
3901 Calverton Blvd, Suite 200 Calverton, MD 20705
March 4, 2019

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_102_shummel.sas7bdat” datasets.

4 Statistical Methods

Analyses were performed to duplicate results for the data published by Hummel et al [1] Diabetes Care in 2017. To verify the integrity of the dataset, descriptive statistics were computed.

5 Results

For Table 1 in the publication [1], Characteristics of children included in the analysis, 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 identical to the published results.

6 Conclusions

The NIDDK repository is confident that the TEDDY M102 data files to be distributed are a true copy of the study data.

7 References

[1] Hummel et al. "First Infant Formula Type and Risk of Islet Autoimmunity in The Environmental Determinants of Diabetes in the Young (TEDDY) Study". Diabetes Care. 2017 Mar; 40(3): 398–404.

Table 1: Variables used to replicate Table 1**:** Characteristics of the data analyzed

Table Variable	**dataset.variable**
Developed any islet autoantibodies	m_102_schummel.Persist_Conf_Ab
Developed multiple islet autoantibodies	m_102_schummel.Two_Or_More_Persistent
Male Child	m_102_schummel.sex
HLA genotype	m_102_schummel.hla_risk
Having a first degree relative w/ type 1 diabetes	m_102_schummel.fdr
C-section	m_102_schummel.csection
Country	m_102_schummel.country
First formula introduced during the first 3 month	m_102_schummel.formula_cat_1_3mo

Table 2: Comparison of values computed in integrity check to reference article Table 1 values

Variable	**Manuscript (n=8,506)**	**DSIC (n=8,506)**	**Diff (n=0)**
Developed any islet autoantibodies	686 (8.1)	686 (8.1)	0(0)
Developed multiple islet autoantibodies	410 (4.8)	410(4.8)	0(0)
Male child	4313 (50.7)	4313 (50.7)	0(0)
HLA genotype			
DR3/4	3319 (39.0)	3319 (39.0)	0(0)
DR4/4	1664 (19.6 %)	1664 (19.6)	0(0)
DR3/3	1782 (21.0%)	1782 (21.0)	0(0)
other	1741 (20.5%)	1741 (20.5)	0(0)
Having a first degree relative with type 1 diabetes	922 (10.8 %)	922 (10.8 %)	0(0)
C-section	2205 (25.9 %)	2205 (25.9 %)	0(0)
Country			
USA	3632 (42.7 %)	3632 (42.7 %)	0(0)
Finland	1805 (21.2 %)	1805 (21.2 %)	0(0)
Germany	572 (6.7 %)	572 (6.7 %)	0(0)
Sweden	2497 (29.4 %)	2497 (29.4 %)	0(0)

Variable	Manuscript (n=8,506)	DSIC (n=8,506)	Diff (n=0)
First formula introduced during the first 3 months			
Non-hydrolyzed cow's milk-based formula	5523 (64.9 %)	5523 (64.9 %)	0(0)
Extensively hydrolyzed cow's milk-based formula	266 (3.1 %)	266 (3.1 %)	0(0)
Partially hydrolyzed cow's milk-based formula	274 (3.2 %)	274 (3.2 %)	0(0)
Other formula*	214 (2.5 %)	214 (2.5 %)	0(0)
No formula, no cow's milk	2198 (25.8 %)	2198 (25.8 %)	0(0)
No formula, regular cow's milk	31 (0.4%)	31 (0.4%)	0(0)

Attachment A: SAS Code

```
*** TEDDY M102 DSIC;
*** Programmer: Sabrina Chen
*** Date: 6/14/18;

options nocenter validvarname=upcase;

title '/prj/niddk/ims_analysis/TEDDY/prog_initial_analysis/m_102_tbls1_2.check.sas';
run;


libname dat '/prj/niddk/ims_analysis/TEDDY/private_orig_data/M_102_SHummel_NIDDK_Submission';


proc format;
 value noyes
 .  = "no"
 other = "yes"
 ;
run;


data m_102_shummel;
       set dat.m_102_shummel;
run;

proc contents data = m_102_shummel;
run;


data m_102_shummel (keep= bf_3mo birth_season country csection exclude exclusion_reason fdr
first_formula_duration_cat formstart_dy_1 formula_cat_1 formula_cat_2
                          formula_cat_1_1w formula_cat_1_3mo fu_any_ab fu_gada_only fu_miaa_only
fu_mult_ab gada_only hla_risk last_visit mask_id miaa_only mom_fdr
                          persist_conf_ab sex switch_3mo two_or_more_persistent dr33 dr34 dr44);
  set m_102_shummel;
  * create binary var for DR ;
  if HLA_RISK = "DR 3/3" then dr33 = 1;
  else if HLA_RISK = "DR 3/4" then dr34 = 1;
  else if HLA_RISK = "DR 4/4" then dr44 = 1;
run;

proc freq data=m_102_shummel;
  tables HLA_RISK*dr33*dr34*dr44/list missing;
  title3 'check binary HLA var';
run;

proc freq data=m_102_shummel;
  tables exclude*EXCLUSION_REASON/list missing;
  title3 'Check population - any exclusions?';
run;


* dups?;
proc sort data= m_102_shummel;
  by mask_id;
run;

data dups;
  set m_102_shummel;
  by mask_id;
  if not(first.mask_id and last.mask_id);
run;

data _null_;
  set dups;
  abort;
run;
```

```
* table 1;
proc freq data=m_102_shummel;
  tables bf_3mo
    birth_season
    country
    csection
    exclude
    exclusion_reason
    fdr
    first_formula_duration_cat
    formstart_dy_1
    formula_cat_1
    formula_cat_2
    formula_cat_1_1w
    formula_cat_1_3mo
    fu_any_ab
    fu_gada_only
    fu_miaa_only
    fu_mult_ab
    gada_only
    hla_risk
    last_visit
    miaa_only
    mom_fdr
    persist_conf_ab
    sex
    switch_3mo
    two_or_more_persistent/missing;
    format fu_mult_ab fu_any_ab gada_only fu_gada_only fu_miaa_only miaa_only noyes.;
title3 'print of all vars';
run;


* table 1;
proc freq data=m_102_shummel;
  tables sex
      hla_risk
      fdr
      csection
      country
      formula_cat_1_3mo /missing;
      title3 'Table 1';
run;

* table 2;
proc freq data=m_102_shummel;
  tables (fdr dr34 dr44 dr33 sex country csection  persist_conf_ab two_or_more_persistent
switch_3mo bf_3mo gada_only mom_fdr miaa_only)*formula_cat_1_3mo/missing;
title3 'Table 2';
run;

proc freq data=m_102_shummel;
  where first_formula_duration_cat ne '';
  tables first_formula_duration_cat*formula_cat_1_3mo/missing;
run;

proc sort data=m_102_shummel;
  by formula_cat_1_3mo;
run;

proc univariate data=m_102_shummel;
  by formula_cat_1_3mo;
  var formstart_dy_1;
run;
```