Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) M109 Vatanen

Prepared by Sabrina Chen
IMS Inc.
3901 Calverton Blvd, Suite 200 Calverton, MD 20705
April 9, 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_109_tvatanen_niddk_31may2012.sas7bdat" dataset.

4 Statistical Methods

Analyses were performed to duplicate results for the data published by Tommi Vatanen et al [1] in *Nature* volume 562, pages589–594 (2018). To verify the integrity of the dataset, descriptive statistics were computed.

5 Results

For Table 1 in the publication [1], **Summary of TEDDY microbiome cohort**, 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 not an exact match to the published results.

6 Conclusions

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

7 References

[1] Tommi Vatanen, Eric A. Franzosa, Randall Schwager, Surya Tripathi, Timothy D. Arthur, Kendra Vehik, Ake Lernmark, William A. Hagopian, Marian J. Rewers, Jin-Xiong She, Jorma Toppari, Anette-G. Ziegler, Beena Akolkar, Jeffrey P. Krischer, the TEDDY Study Group, Christopher J. Stewart, Nadim J. Ajami, Joseph F. Petrosino, Dirk Gevers, Harri Lahdesmaki, Hera Vlamakis, Curtis Huttenhower, Ramnik J. Xavier. The human gut microbiome 1 of early onset type 1 diabetes in the TEDDY study.
Nature volume 562, pages589–594 (2018).

Table A: Variables used to replicate Table 1**:** Summary of TEDDY microbiome cohort.

Table Variable	**dataset.variable**
Country	m_109_tvatanen_niddk_31may2012.cc
Sex	m_109_tvatanen_niddk_31may2012.sex
Ethnic background	m_109_tvatanen_niddk_31may2012.race_ethnicity
Mode of birth	m_109_tvatanen_niddk_31may2012.delivery_simple
Breastfeeding	m_109_tvatanen_niddk_31may2012.time_to_brstfed_stop
Maternal characteristics	m_109_tvatanen_niddk_31may2012.maternal_diabetes m_109_tvatanen_niddk_31may2012.antibioticsduringpregnancy m_109_tvatanen_niddk_31may2012.metformin m_109_tvatanen_niddk_31may2012.glyburide m_109_tvatanen_niddk_31may2012.antihypertensives m_109_tvatanen_niddk_31may2012.insulin

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

Table 1	Manuscript	DSIC	Diff	Manuscript	DSIC	Diff	Manuscript	DSIC	Diff
	US, Colorado			**US, Georgia**			**US, Washington**		
T1D cases (samples)	14 (274)	14 (274)	0	3 (89)	3 (89)	0	8 (111)	8 (111)	0
IA cases (samples)	39 (689)	39 (689)	0	17 (252)	17 (252)	0	25 (368)	25 (368)	0
Healthy controls (samples)	61 (906)	61 (906)	0	22 (250)	22 (250)	0	36 (399)	36 (399)	0
Sex									
Male / Female	61 / 53	61 / 53	0	19 / 23	19 / 23	0	51 / 18	51 / 18	0
Ethnic background									
White, non-hispanic	86 (75.4%)	86 (75.4%)	0	41 (97 .6%)	41 (97 .6%)	0	56 (81 .2%)	56 (81 .2%)	0
Mode of birth									
Caesarean section	41 (36.0%)	41 (36.0%)	0	22 (52.4%)	22 (52.4%)	0	25 (36.2%)	25 (36.2%)	0
Probiotic supplementation									
during first 4 weeks	0			2 (4.8%)			0		
Probiotics during follow-up	22 (19.3%)			13 (31.0%)			9 (13.0%)		
Breastfeeding									
Median duration (days)	268	268	0	301	301	0	335	335	0
duration, 25 percentile	56	56	0	145	145	0	171	171	0
duration, 75 percentile	396	396	0	365	365	0	440	440	0
Number of subjects never breastfed	3	4	1	3	5	2	1	4	3
Maternal characteristics									
Maternal T1D	7 (6.1%)	7 (6.1%)	0	0	0	0	3 (4.3%)	3 (4.4%)	0(0.1)
Maternal T2D	2 (1.8%)	2 (1.8%)	0	0	0	0	0	0	0
Gestational diabetes	5 (4.4%)	5 (4.4%)	0	5 (11.9%)	5 (11.9%)	0	5 (7.2%)	5 (7.3%)	0(0.1)
Antibiotics during pregnancy	21 (18.4%)	21 (18.4%)	0	10 (23.8%)	10 (23.8%)	0	5 (7.2%)	5 (7.3%)	0(0.1)

Metformin during pregnancy	1 (0.9%)	1 (0.9%)	0	0	0	0	0	0	0
Glyburide during pregnancy	2 (1.8%)	2 (1.8%)	0	2 (4.8%)	2 (4.8%)	0	2 (2.9%)	2 (2.9%)	0
Antihypertensives during pregnancy	4 (3.5%)	4 (3.5%)	0	3 (7.1%)	3 (7.1%)	0	4 (5.8%)	4 (5.8%)	0
Insulin during pregnancy	9 (7.9%)	9 (7.9%)	0	0	0	0	3 (4.3%)	3 (4.4%)	0(0.1)
Table 1 (cont.)	Manuscript	DSIC	Diff	Manuscript	DSIC	Diff	Manuscript	DSIC	Diff
	Finland			**Germany**			**Sweden**		
T1D cases (samples)	34 (553)	34 (553)	0	13 (246)	13 (246)	0	29 (532)	29 (532)	0
IA cases (samples)	70 (900)	70 (900)	0	21 (292)	21 (292)	0	95 (1542)	95 (1542)	0
Healthy controls (samples)	119 (1273)	119 (1273)	0	40 (512)	40 (512)	0	137 (1725)	137 (1725)	0
Sex									
Male / Female	117 / 106	117 / 106	0	30 / 44	30 / 44	0	152 / 109	152 / 109	0
Ethnic background									
White, non-hispanic	N/ A	N/ A	0	N/ A	N/ A	0	N/ A	N/ A	0
Mode of birth									
Caesarean section	42 (18.8%)	42 (18.8%)	0	23 (31.1%)	23 (31.1%)	0	46 (17.6%)	46 (17.6%)	0
Probiotic supplementation									
during first 4 weeks	67 (30.0%)			7 (9.5%)			14 (5.4%)		
Probiotics during follow-up	162 (72.6%)			33 (44.6%)			58 (22.2%)		
Breastfeeding									
Median duration (days)	289	289	0	278	278	0	228	228	0
duration, 25 percentile	152	152	0	140	133	7	98	98	0
duration, 75 percentile	385	386	1	367	367	0	304	304	0
Number of subjects never breastfed	0	0	0	0	2	2	0	0	0
Maternal characteristics									
Maternal T1D	14 (6.3%)	14 (6.3%)	0	18 (24.3%)	18 (24.3%)	0	7 (2.7%)	7 (2.7%)	0

Maternal T2D	0	0	0	0	0	0	0	0	0
Gestational diabetes	32 (14.3%)	32 (14.4%)	0(0.1)	3 (4.1%)	3 (4.1%)	0	6 (2.3%)	6 (2.3%)	0
Antibiotics during pregnancy	40 (17.9%)	40 (17.9%)	0	13 (17.6%)	13 (17.6%)	0	29 (11.1%)	29 (11.1%)	0
Metformin during pregnancy	1 (0.4%)	2 (0.9%)	1(0.5)	0	0	0	0	0	0
Glyburide during pregnancy	0	1(0.5)	1(0.5)	0	0	0	0	1(0.4)	1(0.4)
Antihypertensives during pregnancy	5 (2.2%)	5 (2.2%)	0	3 (4.1%)	3 (4.1%)	0	0	0	0
Insulin during pregnancy	23 (10.3%)	21 (9.4%)	2(0.7)	19 (25.7%)	19 (25.7%)	0	8 (3.1%)	7 (2.7%)	1(0.4)

Attachment A: SAS Code

```
options nocenter mprint validvarname=upcase;
title '/prj/niddk/ims_analysis/TEDDY/prog_initial_analysis/m_109_dsic.sas';
run;

*********;
* INPUT  ;
*********;
libname pclib v9 '/prj/niddk/ims_analysis/TEDDY/private_orig_data/M_109_TVatanen_NIDDK_Submission/';


proc format;

  value missn
  .     = 'No Value'
  other = '   Value'
  ;

  value $missc
  ' '   = 'No Value'
  other = '   Value'
  ;

run;


data m109;
  set pclib.m_109_tvatanen_niddk_31may2012;

  if time_to_brstfed_stop ne "NA" then time_to_brstfed_stop_n  = input(time_to_brstfed_stop,8.);
  if time_to_excl_stop ne "NA" then   time_to_excl_stop_n     = input(time_to_excl_stop   ,8.);

  if probiotic_start_week ne "NA" then probiotic_start_week_n  = input(probiotic_start_week,8.);
  if probiotic_stop_week ne "NA" then probiotic_stop_week_n   = input(probiotic_stop_week ,8.);
  if probiotic_start_week_n ne . and probiotic_stop_week_n ne . then delta_probiotic = probiotic_stop_week_n -
probiotic_start_week_n;

  if 0 < delta_probiotic <= 4 then delta_probiotic_le4 = 1;
```

```
  else if delta_probiotic > 4 then delta_probiotic_le4 = 2;

  if ANTIBIOTICSDURINGPREGNANCY in('Amoxicillins', 'Cephalosporins', 'Macrolides', 'Penicillins') then
ANTIBIOTICSDURINGPREGNANCY_flag = 1;
  else if ANTIBIOTICSDURINGPREGNANCY = '' then ANTIBIOTICSDURINGPREGNANCY_flag = 0;
  else if ANTIBIOTICSDURINGPREGNANCY ne "" then abort;

  if      cc = 'COL' then cc_ord= '1 COL';
  else if cc = 'GEO' then cc_ord= '2 GEO';
  else if cc = 'WAS' then cc_ord= '3 WAS';
  else if cc = 'FIN' then cc_ord= '4 FIN';
  else if cc = 'GER' then cc_ord= '5 GER';
  else if cc = 'SWE' then cc_ord= '6 SWE';
  else if cc ne '' then abort;

run;

proc contents data=m109;
run;

proc freq data=m109;
  format age_at_collection shannon_div num_wgs_reads mask_id sample_mask_id time_since_abx time_to_abx missn.
         age_first_pos time_to_brstfed_stop age_mult_persist age_first_ia2a ia_casecontrol_outcome ia_casecontrol_ind
age_t1d age_first_gad age_first_miaa birth_weight $missc.;
       title3 "raw freqs";
run;

proc freq data=m109;
  tables cc_ord*cc/list missing;
  tables delta_probiotic_le4*delta_probiotic/list missing;
  tables antibioticsduringpregnancy_flag*antibioticsduringpregnancy/list missing;
  title3 "checking regrouped variables";
run;


proc sort data=m109 out=m109per nodupkey;
  by mask_id sex race_ethnicity cc t1d_sero_control delivery_simple maternal_diabetes time_to_brstfed_stop
time_to_excl_stop probiotic_start_week probiotic_stop_week
     antibioticsduringpregnancy metformin glyburide antihypertensives insulin;
run;
```

```
** table 1;
proc freq data=m109;
  tables t1d_sero_control  *cc_ord/ missing;
  title3 "table 1 - sample counts";
run;

proc freq data=m109per;
  tables t1d_sero_control  *cc_ord/ missing;
  tables sex               *cc_ord/ missing;
  tables race_ethnicity    *cc_ord/ missing;
  tables delivery_simple   *cc_ord/ missing;
/*  tables delta_probiotic_le4   *cc_ord/ missing;  */
  tables maternal_diabetes *cc_ord/ missing;
  tables antibioticsduringpregnancy_flag *cc_ord/ missing;
  tables metformin     *cc_ord/ missing;
  tables glyburide     *cc_ord/ missing;
  tables antihypertensives *cc_ord/ missing;
  tables insulin       *cc_ord/ missing;
  title3 "table 1 - subject counts";
  run;

proc sort data=m109per;
  by cc_ord;
run;

proc univariate data=m109per;
  by cc_ord;
  var time_to_brstfed_stop_n;
run;

** table 2;
proc freq data=m109per;
  tables antibioticsduringpregnancy*cc_ord/missing;
title3 "table 2 ??";
run;

proc freq data=m109;
  tables antibioticsduringpregnancy*cc_ord/missing;
title3 "table 2 ??";
```

```
run;
```