Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) M142 KVehik

Prepared by Sabrina Chen
IMS Inc.
3901 Calverton Blvd, Suite 200 Calverton, MD 20705
February 24, 2020

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/private_orig_data/M_142_KVehik_NIDDK_Submission folder in the data package. For this replication, variables were taken from the "m_142_kvehik_niddk_31may2012_1.sas7bdat" and "m_142_kvehik_niddk_31may2012_3.sas7bdat" dataset.

4 Statistical Methods

Analyses were performed to duplicate results for the data published by Kendra Vehik et al [1] in the Nature Medicine in 2019. To verify the integrity of the dataset, descriptive statistics were computed.

5 Results

For **Comparison of Data in the publication**, 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.

6 Conclusions

The results of the replication are almost an exact match to the published results.

7 References

[1] Kendra Vehik , Kristian F. Lynch, Matthew C. Wong, Xiangjun Tian, Matthew C. Ross, Richard A. Gibbs, Nadim J. Ajami, Joseph F. Petrosino, Marian Rewers, Jorma Toppari, Anette G. Ziegler, Jin-Xiong She, Ake Lernmark, Beena Akolkar, William A. Hagopian, Desmond A. Schatz, Jeffrey P. Krischer, Heikki Hyöty, Richard E. Lloyd and the TEDDY Study Group. Prospective virome analyses in young children at increased genetic risk for type 1 diabetes. Nature Medicine 25, pages1865–1872(2019).

Table A: Variables used to replicate data in the publication.

Table Variable	dataset.variable
Age (days) when child (case) developed Islet autoaimmunity and age (days) when matching control was selected	m_142_kvehik_niddk_31may2012_1.ia_case_endptage
Continent of residence - 0 = US, 1 = Europe	m_142_kvehik_niddk_31may2012_1.eu
Indicates whether or not the subject had serconverted for any of the three Islet Autoantibodies (GADA, IA2A and IAA) by May 31st 2012 ? 0=no, a control , 1=yes, a case	m_142_kvehik_niddk_31may2012_1.ia_case_outcome
Site of residence - 1 = US-Colorado, 2 = US-Georgia/Florida, 3 = US-Washington State, 4 = Finland, 5 = Germany, 6 = Sweden	m_142_kvehik_niddk_31may2012_1.site
Gender of child is male - 0 = no, 1 = yes	m_142_kvehik_niddk_31may2012_1.male
First degree relative status - 1=FDR, 0=GenPop	m_142_kvehik_niddk_31may2012_1.fdr
Indicates whether or not the subject was diagnosed with type 1 diabetes by May 31st 2012 ? 0=no, a control , 1=yes, a case	m_142_kvehik_niddk_31may2012_3.t1d_case_outcome
Child developed Islet autoantibodies before diagnosis - 0 = no, 1 - yes	m_142_kvehik_niddk_31may2012_3.ia_case_child

Site of residence - 1 = US-Colorado, 2 = US-Georgia/Florida, 3 = US-Washington State, 4 = Finland, 5 = Germany, 6 = Sweden	m_142_kvehik_niddk_31may2012_3.site
Gender of child is male - 0 = no, 1 = yes	m_142_kvehik_niddk_31may2012_3.male
Continent of residence - 0 = US, 1 = Europe	m_142_kvehik_niddk_31may2012_3.eu
First degree relative status -1=FDR, 0=GenPop	m_142_kvehik_niddk_31may2012_3.fdr
Age (year) child was diagnosed with type 1 diabetes and age match control was selected	m_142_kvehik_niddk_31may2012_3.t1d_age_year

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

Factors used to match each case with a control	Islet Autoantibody Cases											
	N=383											
	US						Europe					
	N = 119						N = 264					
	N (%)						N (%)					
Study site [a]	**Manuscript**	**DSIC**	**Diff**	**Manuscript**	**DSIC**	**Diff**	**Manuscript**	**DSIC**	**Diff**	**Manuscript**	**DSIC**	**Diff**
Colorado	55	55	0	46.2	46.2	0						
Georgia/Florida	28	28	0	23.5	23.5	0						
Washington state	36	36	0	30.3	30.3	0						
Finland		.			.		104	104	0	39.4	39.4	0
Germany		.			.		31	31	0	11.7	11.7	0
Sweden		.			.		129	129	0	48.9	48.9	0
Gender [a]												
Male	65	65	0	54.6	54.6	0	145	145	0	54.9	54.9	0
Female	54	54	0	45.4	45.4	0	119	119	0	45.1	45.1	0
Family history with												
type 1 diabetes [a]												
No	87	87	0	73.1	73.1	0	212	212	0	80.3	80.3	0
Yes	32	32	0	26.9	26.9	0	52	52	0	19.7	19.7	0
Age islet autoimmunity[b] appeared												
< 12 months	18	18	0	15.1	15.1	0	59	59	0	22.3	22.3	0
12 – <24 months	48	48	0	40.3	40.3	0	85	85	0	32.5	32.2	0.3

24 - <36 months	32	32	0	26.9	26.9	0	62	62	0	23.5	23.5	0
36 - < 48 months	10	10	0	8.4	8.4	0	36	36	0	13.6	13.6	0
48 – 72 months	11	11	0	9.2	9.2	0	22	22	0	8.3	8.3	0
Age T1D diagnosis												
< 12 months	-						-					
12 – <24 months	-						-					
24 - <36 months	-						-					
36 - < 48 months	-						-					
48 – 72 months	-						-					

	Type 1 Diabetes Cases											
	N=112											
	US						**Europe**					
	N = 29						**N = 83**					
Study site [a]	**Manuscript**	**DSIC**	**Diff**	**Manuscript**	**DSIC**	**Diff**	**Manuscript**	**DSIC**	**Diff**	**Manuscript**	**DSIC**	**Diff**
Colorado	16	16	0	55.2	55.2	0						
Georgia/Florida	5	5	0	17.2	17.2	0						
Washington state	8	8	0	27.6	27.6	0						
Finland							35	35	0	42.2	42.2	0
Germany							18	18	0	21.7	21.7	0
Sweden							30	30	0	36.1	36.1	0
Gender [a]												

Male	17	17	0	58.6	58.6	0	42	42	0	50.6	50.6	0
Female	12	12	0	41.4	41.4	0	41	41	0	49.4	49.4	0
Family history with												
type 1 diabetes [a]												
No	17	17	0	58.6	58.6	0	55	55	0	66.3	66.3	0
Yes	12	12	0	41.4	41.4	0	28	28	0	33.7	33.7	0
Age islet autoimmunity[b] **appeared**												
< 12 months	8	8	0	33.3	33.3	0	26	26	0	41.3	41.3	0
12 – <24 months	13	13	0	54.2	54.2	0	32	32	0	50.8	50.8	0
24 - <36 months	3	3	0	12.5	12.5	0	2	2	0	3.2	3.2	0
36 - < 48 months	0	0	0	0	0	0	3	3	0	4.8	4.8	0
48 – 72 months	0	0	0	0	0	0	0	0	0	0	0	0
Age T1D diagnosis												
< 12 months	1	1	0	3.4	3.4	0	4	4	0	4.8	4.8	0
12 – <24 months	10	10	0	34.5	34.5	0	27	27	0	32.5	32.5	0
24 - <36 months	5	5	0	17.2	17.2	0	24	24	0	28.9	28.9	0
36 - < 48 months	8	8	0	27.6	27.6	0	10	10	0	12	12	0
48 – 72 months	5	5	0	17.2	17.2	0	18	18	0	21.7	21.7	0

Attachment A: SAS Code

```
options nocenter validvarname=upcase;

title '/prj/niddk/ims_analysis/TEDDY/prog_initial_analysis/m_142_dsic.sas';
run;

libname dat '/prj/niddk/ims_analysis/TEDDY/private_orig_data/M_142_KVehik_NIDDK_Submission';


proc format;
 value val
 .     = "no value"
 other = "   value"
 ;

 value oneplus
 . = "no value"
 0 = "0"
 1-high = "1+"
 ;

 value zerohi
 . = "no value"
 0-high = "0-high"
 ;

 value sitef
 1 = '(1) US-Colorado'
 2 = '(2) US-Georgia/Florida'
 3 = '(3) US-Washington State'
 4 = '(4) Finland'
 5 = '(5) Germany'
 6 = '(6) Sweden'
 ;

 value malef
 0 = '(B) Female'
 1 = '(A) Male'
 ;

 value fdr
 1='(B) FDR'
 0='(A) GenPop'
 ;

 value ageislet
 1 = '    <12 months'
 2 = '12 =<24 months'
 3 = '24 =<36 months'
 4 = '36 =<48 months'
 5 = '48 = 72 months'
 ;

run;



* produce n and %;
%macro npercent(ds, rownum, var, varf, subset,  subsetname);
  proc freq data=&ds noprint;
    where &subset = 1;
    tables &var/list missing out=tbl1&subsetname;
    format &var &varf..;
  run;

  data tbl1&subsetname;
    length covar covarf $100;
    set tbl1&subsetname;
    covar = "&var";
    covarf = put(&var,&varf..);
    rownum = &rownum;
  run;

  data prnt&subsetname;
    set prnt&subsetname tbl1&subsetname;
  run;

%mend;



data analy;
  set dat.m_142_kvehik_niddk_31may2012_1;
  * convert to months;
  if (. < (ia_case_endptage/30.5) < 12) then ia_case_endptage_mon = 1;
  else if (12 <= (ia_case_endptage/30.5) < 24) then ia_case_endptage_mon = 2;
  else if (24 <= (ia_case_endptage/30.5) < 36) then ia_case_endptage_mon = 3;
  else if (36 <= (ia_case_endptage/30.5) < 48) then ia_case_endptage_mon = 4;
  else if (48 <= (ia_case_endptage/30.5) <= 72.5) then ia_case_endptage_mon = 5;  *NOTE: one subject was 72.3 months;

  * create subsets;
  if ia_case_outcome=1 and eu=0 then subset_us=1;
  if ia_case_outcome=1 and eu=1 then subset_eu=1;

run;

proc contents data=analy;
run;

proc freq data=analy;
  tables ia_case_endptage_mon*ia_case_endptage/list missing;
  tables subset_us*subset_eu*ia_case_outcome* eu/list missing;
  title3 "file 1 - checking";
run;


proc freq data=analy;
  tables site male fdr ia_case_endptage_mon /list missing;
  run;

* n and percent;
data prntusa;
  set _null_;
run;

%npercent(analy, 1, site     , sitef          , subset_us , usa);
%npercent(analy, 2, male     , malef   , subset_us , usa);
%npercent(analy, 3, fdr      , fdr     , subset_us , usa);
%npercent(analy, 4, ia_case_endptage_mon, ageislet     , subset_us    , usa);

proc print data=prntusa;
  var covar covarf count percent;
run;


data prnteu;
  set _null_;
run;

%npercent(analy, 1, site     , sitef          , subset_eu , eu);
%npercent(analy, 2, male     , malef   , subset_eu , eu);
%npercent(analy, 3, fdr      , fdr     , subset_eu , eu);
%npercent(analy, 4, ia_case_endptage_mon, ageislet     , subset_eu    , eu);
```

```
proc print data=prnteu;
  var covar covarf count percent;
run;



data analy3;
  set dat.m_142_kvehik_niddk_31may2012_3;
  * create subsets;
  if T1D_CASE_OUTCOME=1 and eu=0 then subset_us=1;
  if T1D_CASE_OUTCOME=1 and eu=1 then subset_eu=1;

  if IA_CASE_CHILD=1 and T1D_CASE_OUTCOME =1 and eu=0 then subset_us_ia=1;
  if IA_CASE_CHILD=1 and T1D_CASE_OUTCOME =1 and eu=1 then subset_eu_ia=1;

run;

proc contents data=analy3;
title3 "analy3";
run;

proc freq data=analy3;
  where T1D_CASE_OUTCOME =1;
  tables site male eu fdr  T1D_AGE_YEAR/list missing;
  run;

proc freq data=analy3;
  where IA_CASE_CHILD=1 and T1D_CASE_OUTCOME =1;
  tables IA_CASE_YEAR/missing;
  run;


* n and percent;
data prntusa3;
  set _null_;
run;

%npercent(analy3, 1, site     , sitef         , subset_us , usa3);
%npercent(analy3, 2, male     , malef  , subset_us , usa3);
%npercent(analy3, 3, fdr      , fdr    , subset_us , usa3);
%npercent(analy3, 4, IA_CASE_YEAR, ageislet  , subset_us_ia , usa3);
%npercent(analy3, 5, T1D_AGE_YEAR, ageislet  , subset_us    , usa3);

proc print data=prntusa3;
```

```
  var covar covarf count percent;
run;


data prnteu3;
  set _null_;
run;

%npercent(analy3, 1, site     , sitef         , subset_eu , eu3);
%npercent(analy3, 2, male     , malef  , subset_eu , eu3);
%npercent(analy3, 3, fdr      , fdr    , subset_eu , eu3);
%npercent(analy3, 4, IA_CASE_YEAR, ageislet  , subset_eu_ia , eu3);
%npercent(analy3, 5, T1D_AGE_YEAR, ageislet  , subset_eu    , eu3);

proc print data=prnteu3;
  var covar covarf count percent;
run;



proc sort data=prntusa;
  by rownum covarf;
run;

proc sort data=prnteu (rename=(COUNT=count_eu PERCENT=percent_eu));
  by rownum covarf;
run;

data st1_ia;
  merge prntusa (in=in1 keep=rownum covarf count percent)
        prnteu  (in=in2 keep=rownum covarf count_eu percent_eu);
  by rownum covarf;
  if in1 or in2;

  percent = put(percent,8.1);
  percent_eu = put(percent_eu,8.1);
run;

proc print data=st1_ia;
title3 "Supplemental Table 1 - Islet Autoantibody Cases";
run;


proc sort data=prntusa3;
  by rownum covarf;
```

```
run;

proc sort data=prnteu3 (rename=(COUNT=count_eu PERCENT=percent_eu));
  by rownum covarf;
run;

data st1_t1d;
  merge prntusa3 (in=in1 keep=rownum covarf count percent)
        prnteu3  (in=in2 keep=rownum covarf count_eu percent_eu);
  by rownum covarf;
  if in1 or in2;

  percent = put(percent,8.1);
  percent_eu = put(percent_eu,8.1);
run;

proc print data=st1_t1d;
title3 "Supplemental Table 1 - Type 1 Diabetes Cases";
run;
```