Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) M169 Hård af Segerstad

Prepared by Sabrina Chen
IMS Inc.
3901 Calverton Blvd, Suite 200 Calverton, MD 20705
May 29, 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_169_hard_niddk_31july2016_2.sas7bdat” dataset.

4 Statistical Methods

Analyses were performed to duplicate results for the data published by Elin M. Hård af Segerstad et al [1] in *Nutrients* 2018, *10*(5), 550. To verify the integrity of the dataset, descriptive statistics were computed.

5 Results

For Table 1 in the publication [1], **Characteristics of the identified cases with celiac disease in the Swedish TEDDY birth cohort used as matching factors in a nested 1-3 case-control study**, 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 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] Elin M. Hård af Segerstad, Hye-Seung Lee, Carin Andrén Aronsson , Jimin Yang, Ulla Uusitalo, Ingegerd Sjöholm, Marilyn Rayner , Kalle Kurppa, Suvi M. Virtanen, Jill M. Norris, Daniel Agardh, and on behalf of the TEDDY Study Group. Daily Intake of Milk Powder and Risk of Celiac Disease in Early Childhood: A Nested Case-Control Study. *Nutrients* 2018, *10*(5), 550.

Table A: Variables used to replicate Table 1: Characteristics of the identified cases with celiac disease in the Swedish TEDDY birth cohort used as matching factors in a nested 1-3 case-control study.

Table Variable	**dataset.variable**
Gender	m_169_hard_niddk_31july2016_2.female
Birth year	m_169_hard_niddk_31july2016_2.birthyear
HLA genotype	m_169_hard_niddk_31july2016_2.hlarg
Outcome	m_169_hard_niddk_31july2016_2.outcome

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

Table 1	Manuscript	DSIC	Diff	Manuscript	DSIC	Diff
	Cases (n)			**Cases (%)**		
- Female	131	131	0	63.3	63.3	0
- Male	76	76	0	63.3	63.3	0
Birth year						
- 2004	11	11	0	5.3	5.3	0
- 2005	39	39	0	18.8	18.8	0
- 2006	28	28	0	13.5	13.5	0
- 2007	41	41	0	19.8	19.8	0
- 2008	37	37	0	17.9	17.9	0
- 2009	46	46	0	22.2	22.2	0
- 2010	5	5	0	2.4	2.4	0
HLA-genotype						
- DQ2/DQ8	64	64	0	22.2	22.2	0
- DQ8/DQ8	35	35	0	30.9	30.9	0
- DQ2/DQ2	100	100	0	16.9	16.9	0
- Other	8	8	0	48.3	48.3	0

Attachment A: SAS Code

```
options nocenter validvarname=upcase;

title 'prj/niddk/ims_analysis/TEDDY/prog_initial_analysis/m_169_dsic.sas';
run;

* DSIC for TEDDY M106a. Reproduce Table 1 of M_169_Hard_NIDDK_Manuscript.pdf ;

***********;
* INPUT    ;
***********;

libname sasfile1 '/prj/niddk/ims_analysis/TEDDY/private_orig_data/M_169_Hard_NIDDK_Submission/';


************;
* MACROS    ;
************;
%macro readin(lib, ds);
  data &ds;
    set sasfile&lib..&ds;
  run;

  proc contents data=&ds;
  title3 "&ds";
  run;
%mend;


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

**********;
* FORMATS ;
**********;

proc format;

  value hlaf
  . = 'HLA ineligible'
  1 = 'DR3/4'
  2 = 'DR4/4'
  3 = 'DR4/8'
  4 = 'DR3/3'
  5 = 'DR4/4b'
  6 = 'DR4/1'
  7 = 'DR4/13'
  9 = 'DR4/9'
  10 = 'DR3/9'
  ;

  value hlagpf
  . = 'HLA ineligible'
  1 = 'DR3/4'
  2 = 'DR4/4'
  4 = 'DR3/3'
  99 = 'Other'
  ;

  value female
  0 = 'Male'
  1 = 'Female'
  ;

run;


%readin(1, m_169_hard_niddk_31july2016_1);
%readin(1, m_169_hard_niddk_31july2016_2);
```

```
proc freq data=m_169_hard_niddk_31july2016_2;
  table outcome/missing;
run;

proc freq data= m_169_hard_niddk_31july2016_2;
  where outcome=1;
  tables birthyear female hlarg/missing;
run;


data analy;
  set m_169_hard_niddk_31july2016_2;
  subset_all = 1;

  if hlarg in(3,6,7,9,10) then hlarg_gp = 99;
  else hlarg_gp = hlarg;
run;

proc freq data=analy;
  tables hlarg_gp*hlarg/list missing;
run;


** Table 1;

* n percent;
data prntcase;
  set _null_;
run;

%npercent(1   , female    , female    , outcome , case);
%npercent(2   , birthyear  , best      , outcome , case);
%npercent(3   , hlarg_gp   , hlagpf    , outcome , case);

data prntcase;
  set prntcase;
  percent  = round(percent  , 0.1);
run;

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