Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) Pub56 Hagopian

Prepared by Jane Rideau Demuth
IMS Inc.
3901 Calverton Blvd, Suite 200 Calverton, MD 20705
August 7, 2018

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_56_whagopian_niddk_31dec2013.sas7bdat" dataset.

4 Statistical Methods

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

5 Results

For Figure 1 in the publication [1], Venn diagram showing the overlap of IAs and tTGAs as well as T1D and CD in cross-sectional prevalence of 5891 TEDDY cohort subjects at a median age of 66 months, 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 M56 data files to be distributed are a true copy of the study data.

7 References

[1] Hagopian W, Lee H-S, Liu E, et al. Co-occurrence of Type 1 Diabetes and Celiac Disease Autoimmunity. *Pediatrics.* 2017;140(5):e20171305.

Table A: Variables used to replicate Figure 1**:** Venn diagram showing the overlap of IAs and tTGAs as well as T1D and CD in cross-sectional prevalence of 5891 TEDDY cohort subjects at a median age of 66 months

Table Variable	**dataset.variable**
Diagnostic category	m_56_whagopian_niddk_31dec2013.ecat

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

Variable	Value	Manuscript n	IMS data check n	Diff. (n=0)
Diagnostic category (ecat)	CD, T1D, Islet Ab+, Persistent TG+	18	18	0
	T1D, Islet Ab+, Persistent TG+	5	5	0
	CD, Islet Ab+, Persistent TG+	20	20	0
	CD, Persistent TG+	283	283	0
	T1D, Persistent TG+	1	1	0
	T1D, Islet Ab+	107	107	0
	Islet Ab+, Persistent TG+	47	47	0
	CD Only	2	2	0
	T1D only	7	7	0
	Persistent TG+ only	524	524	0
	Islet Ab+ only	260	260	0

Attachment A: SAS Code

```
options mprint nocentre linesize=163 validvarname=upcase;

%let rundate = y2018m07d31;
%let olddate = yYYYYmMMdDD;

title "Program:
/prj/niddk/ims_analysis/TEDDY/prog_initial_analysis/M_56_WHagopian_NIDDK_Submission/DSIC.paper.re
view.&rundate..sas";
title2 "This program reviews the TEDDY M56 paper ('Co-occurrence of Type 1 Diabetes and Celiac
Disease Autoimmunity')";

/*********************************************************************

programmer:  Jane Rideau Demuth

platform:  LINUX SASv9.4

date:  30 July 2018

purpose:  See title2.

*********************************************************************/

***************;
*** formats ***;
***************;
proc format;
  value nmsgf
    . = ' '
    low-high = '###'
    ;
  value $cmsgf
    ' ' = ' '
    other = '$$$'
    ;
  value ecatf
    1='CD, T1D, Islet Ab+, Persistent TG+'
    2='T1D, Islet Ab+, Persistent TG+'
    3='CD, Islet Ab+, Persistent TG+'
    4='CD, T1D, Islet Ab+'
    5='CD, T1D, Islet Ab+, Persistent TG+'
    6='CD, T1D'
    7='CD, Persistent TG+'
    8='T1D, Persistent TG+'
    9='CD, Islet Ab+'
    10='T1D, Islet Ab+'
    11='Islet Ab+, Persistent TG+'
    12='CD Only'
    13='T1D only'
    14='Persistent TG+ only'
    15='Islet Ab+ only'
    ;

******************;
*** input files ***;
******************;
libname pcsasin
"/prj/niddk/ims_analysis/TEDDY/private_orig_data/M_56_WHagopian_NIDDK_Submission/";
data m56;
  set pcsasin.m_56_whagopian_niddk_31dec2013;
title3 "Input file:
/prj/niddk/ims_analysis/TEDDY/private_orig_data/M_56_WHagopian_NIDDK_Submission/m_56_whagopian_ni
ddk_31dec2013.sas7bdat";
proc contents data=m56 varnum;

**********************;
*** set up the data ***;
```

```
**********************;
proc freq data=m56;
  title3 'Check variables for model';
  tables ecat
         / missing list;
  format ecat ecatf.;

endsas;
```