Persons using assistive technology may not be able to fully access information in this file. For assistance, e-mail niddk-cr@imsweb.com. Include the Web site and filename in your message.

Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) Pub36 Eliu

Prepared by Jane Wang
IMS Inc.
3901 Calverton Blvd, Suite 200 Calverton MD 20705
August 20, 2015

Table of 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 “Pub36_ELiu_NIDDK_submission” dataset.

4 Statistical Methods

Analyses were performed to duplicate results for the data published by Edwin Liu et al [1] The new England journal of medicine 371;1 July 3, 2014. To verify the integrity of the dataset, descriptive statistics of baseline characteristics were computed, by HLA DR–DQ Genotype (Table B).

5 Results

Table A lists the variables that were used in the replication and Table B compares the results calculated from the archived data file to the results published in Table 1. The results of the replication are the same to the published results.

6 Conclusions

The NIDDK repository is confident that the TEDDY Pub36 ELiu data files to be distributed are a true copy of the study data.

7 References

Edwin Liu, M.D., Hye-Seung Lee, Ph.D., Carin A. Aronsson, M.Sc.,William A. Hagopian, M.D., Ph.D., Sibylle Koletzko, M.D., Ph.D., Marian J. Rewers, M.D., M.P.H., George S. Eisenbarth, M.D., Ph.D.,Polly J. Bingley, M.D., Ezio Bonifacio, Ph.D., Ville Simell, M.Sc., Daniel Agardh, M.D., Ph.D., for the TEDDY Study Group. Risk of Pediatric Celiac Disease According to HLA Haplotype and Country. The new England journal of medicine 371;1 July 3, 2014.

Table A: Variables used to replicate Tables 1 in the publication.

Table Variable	**Variables Used in Replication from the "Table 1" Dataset**
Family history of type 1 diabetes	fdr
Family history of celiac disease	celiac_fdr
sex	female
Country	country
HLA DR–DQ Genotype	hlarg

Table B: Comparison of values computed in integrity check to reference article Table 1 values: HLA Genotypes of 6403 Children Screened for Celiac Disease

Variable	Total [Manuscript]	Total [DSIC]	Total [Difference]	DR3-DQ2/DR3-CDQ2 [Manuscript]	DR3-CDQ2/DR3-CDQ2 [Difference]	DR3-CDQ2/DR3-CDQ2 [Difference]	DR3-CDQ2/DR4-CDQ8 [Manuscript]	DR3-CDQ2/DR4-CDQ8 [DSIC]	DR3-CDQ2/DR4-CDQ8 [Difference]
All children	6403 (100)	6403(100)	0(0)	1374 (21)	1374(21)	0(0)	2612 (41)	2612(41)	0(0)
Family history of type 1 diabetes yes	535 (8)	535(8)	0(0)	109 (20)	109(20)	0(0)	238 (44)	238(44)	0(0)
Family history of type 1 diabetes No	5868 (92)	5868(92)	0(0)	1265 (22)	1265(22)	0(0)	2374 (40)	2374(40)	0(0)
Family history of celiac disease Yes	144 (2)	144(2)	0(0)	52 (36)	52(36)	0(0)	72 (50)	72(50)	0(0)
Family history of celiac disease No	6259 (98)	6259(98)	0(0)	1322 (21)	1322(21)	0(0)	2540 (41)	2540(41)	0(0)
sex Female	3118 (49)	3118(49)	0(0)	627 (20)	627(20)	0(0)	1289 (41)	1289(41)	0(0)
sex Male	3285 (51)	3285(51)	0(0)	747 (23)	747(23)	0(0)	1323 (40)	1323(40)	0(0)
Country United States	2562 (40)	2562(40)	0(0)	626 (24)	626(24)	0(0)	1064 (42)	1064(42)	0(0)
Country Finland	1461 (23)	1461(23)	0(0)	227 (16)	227(16)	0(0)	516 (35)	516(35)	0(0)
Country Germany	344 (5)	344(5)	0(0)	81 (24)	81(24)	0(0)	163 (47)	163(47)	0(0)
Country Sweden	2036 (32)	2036(32)	0(0)	440 (22)	440(22)	0(0)	869 (43)	869(43)	0(0)

Variable	DR4-CDQ8/DR4-CDQ8 [Manuscript]	DR4-CDQ8/DR4-CDQ8 [DSIC]	DR4-CDQ8/DR4-CDQ8 [Difference]	DR4-CDQ8/DR8-CDQ4 [Manuscript]	DR4-CDQ8/DR8-CDQ4 [DSIC]	DR4-CDQ8/DR8-CDQ4 [Difference]
All children	1303 (20)	1303(20)	0(0)	1114 (17)	1114(17)	0(0)
Family history of type 1 diabetes yes	122 (23)	122(23)	0(0)	66 (12)	66(12)	0(0)
Family history of type 1 diabetes No	1181 (20)	1181(20)	0(0)	1048 (18)	1048(18)	0(0)
Family history of celiac disease Yes	11 (8)	11(8)	0(0)	9 (6)	9(6)	0(0)
Family history of celiac disease No	1292 (21)	1292(21)	0(0)	1105 (18)	1105(18)	0(0)
sex Female	661 (21)	661(21)	0(0)	541 (17)	541(17)	0(0)
sex Male	642 (20)	642(20)	0(0)	573 (17)	573(17)	0(0)
Country United States	533 (21)	533(21)	0(0)	339 (13)	339(13)	0(0)
Country Finland	245 (17)	245(17)	0(0)	473 (32)	473(32)	0(0)
Country Germany	68 (20)	68(20)	0(0)	32 (9)	32(9)	0(0)
Country Sweden	457 (22)	457(22)	0(0)	270 (13)	270(13)	0(0)

Attachment A: SAS Code

```
*************************************************************************************************************************
***Program:
***Programmer: Jane Wang
***Date Created: 08/19/2015
***Purpose:
*************************************************************************************************************************;

title1 "%sysfunc(getoption(sysin))";
title2 " ";

options nofmterr;
options nofmterr;
proc format;
 value YESNO
 0=no
 1=yes
;

value COUNTRY
1='US'
2='FIN'
3='GER'
4='SWE'
;
value HLARGC
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';


libname sas_data "/prj/niddk/ims_analysis/TEDDY/private_orig_data/Pub36_ELiu_NIDDK_submission/";

*** File containing macro for examining each dataset ***;
%include '/prj/niddk/ims_analysis/sas_macros/redaction_data_summary.sas';
data pub36_eliu_niddk         ; set sas_data.pub36_eliu_niddk     ;


*** Data from the Primary hlarg paper that was converted to .csv format so that the DSIC data could be easily compared;
FILENAME table1  '/prj/niddk/ims_analysis/TEDDY/private_created_data/teddy_pub36_table1.csv';

*** Output CSV files that will be converted to .xls before being added to the DSIC document;
FILENAME out_t1  '/prj/niddk/ims_analysis/TEDDY/private_created_data/teddy_pub36_table1_dsic.csv';

*** Reading in the analysis datasets used for the DSIC;

%macro baseline_freq1(dataset_name,var_name,);

       *** Creating a frequency table in the format of Table 1 in the primary hlarg paper;
```

```
proc freq data = &dataset_name ;
      table (&var_name.)*hlarg ;
      title3 "Frequency table of the &var_name. variable in the analysis dataset";

      *** Outputting the frequency data to work.&var_name._cross using the ODS output;
ods output CrossTabFreqs = work.&var_name._cross;
proc print data = &var_name._cross;
data &var_name._cross(keep =  hlarg Frequency Rowpercent table_name &var_name);
  set &var_name._cross;
  if hlarg ne . and Rowpercent ne .;
  length table_name $30.;
  table_name  = "&var_name" ;

  proc sort data = &var_name._cross;
   by   &var_name hlarg;
proc print data = &var_name._cross;

data &var_name._cross;
  set &var_name._cross;
  by &var_name;
  retain count 0;
  if first.&var_name then count = 0;
  count = count + 1;

proc print data = &var_name._cross;


data &var_name._cross_1(drop =   Frequency Rowpercent i);
  set &var_name._cross;
  by  &var_name ;
  array temp1(4) count34 count44 count48 count33 ;
  array temp2(4) pert34 pert44 pert48 pert33 ;
  retain count34 count44 count48 count33 pert34 pert44 pert48 pert33 ;
  if first.&var_name then do i = 1 to 4;
    temp1(i) = .;
    temp2(i) = .;
  end;
    temp1(count) = Frequency;
    temp2(count) = Rowpercent;
  if last.&var_name;

proc print data = &var_name._cross_1;

proc freq data = &dataset_name ;
      table &var_name/  out = &var_name._freq;
      title3 "Frequency table of the &var_name. variable in the analysis dataset";
data &var_name._freq;
  set &var_name._freq;
  length table_name $30.;
  table_name  = "&var_name" ;
```

```
    rename count = countall PERCENT =PERtall;

proc print data =  &var_name._freq;

data &var_name._c;
  merge &var_name._freq &var_name._cross_1;
  by table_name &var_name;
  rename &var_name = var_level;
  table_name = compress(table_name || put(&var_name,1.));
proc print data =  &var_name._c;


%mend;


*******************************************************************************;
*************** Check Table 1 *************************************************;
*******************************************************************************;

*** Running the baseline_freq on the  categorical variables in the Table 1 manuscript file;

*%baseline_freq1(pub36_eliu_niddk,hlarg        );
proc freq data = pub36_eliu_niddk;
  tables hlarg/out = outall;

data outall ;
  set outall;
*    rename count = countall PERCENT =PERtall;
    length table_name $30.;
    table_name  = "hlarg" ;


proc print data = outall;

data outall ;
  set outall;
  by table_name;
    array temp1(4) count34 count44 count48 count33 ;
    array temp2(4) pert34 pert44 pert48 pert33 ;
    retain count34 count44 count48 count33 pert34 pert44 pert48 pert33 ;
    if first.table_name then do i = 1 to 4;
      temp1(i) = .;
      temp2(i) = .;
    end;
      temp1(_n_) = count;
      temp2(_n_) = PERCENT;
    if last.table_name;

data outall ;
  set outall;
  countall = sum(count34, count44, count48, count33 );
```

```
  PERtall = round(sum(pert34, pert44, pert48, pert33),1);
  proc print data = outall;

%baseline_freq1(pub36_eliu_niddk,fdr        );
%baseline_freq1(pub36_eliu_niddk,celiac_fdr        );
%baseline_freq1(pub36_eliu_niddk,female       );
%baseline_freq1(pub36_eliu_niddk,country       );



data table1_compare(drop = hlarg  COUNT  PERCENT  i var_level);
  set outall fdr_c celiac_fdr_c female_c country_c     ;
  pert33   = round(pert33,1.);
  pert34   = round(pert34,1.);
  pert44   = round(pert44,1.);
  pert48   = round(pert48,1.);
  PERtall   = round(PERtall,1.);

proc print data = table1_compare;

*** Importing the Table 1 Data taken from the primary hlarg paper;
data table1_data;
  infile table1 delimiter = ',' MISSOVER DSD  firstobs=1 ls=1080;
  length character $100.   table_name $ 30. charall_ char33_ char34_ char44_        char48_ $ 25.;
  input character      $ table_name $ charall_   $ char33_ $ char34_ $ char44_ $    char48_ $ p$
;

data table1_data;
  set table1_data;
  ordernum = _n_;
  countall_ = input(substr(charall_,1,index(charall_, '(')-1),8.);
  pertall_  = input(substr(charall_,index(charall_, '(')+1, length(charall_)-index(charall_, '(')-1),8.);

  count33_ = input(substr(char33_,1,index(char33_, '(')-1),8.);
  pert33_  = input(substr(char33_,index(char33_, '(')+1, length(char33_)-index(char33_, '(')-1),8.);

  count34_ = input(substr(char34_,1,index(char34_, '(')-1),8.);
  pert34_  = input(substr(char34_,index(char34_, '(')+1, length(char34_)-index(char34_, '(')-1),8.);

  count44_ = input(substr(char44_,1,index(char44_, '(')-1),8.);
  pert44_  = input(substr(char44_,index(char44_, '(')+1, length(char44_)-index(char44_, '(')-1),8.);

  count48_ = input(substr(char48_,1,index(char48_, '(')-1),8.);
  pert48_  = input(substr(char48_,index(char48_, '(')+1, length(char48_)-index(char48_, '(')-2),8.);


proc print data = table1_data;
  title3 'table 1 from paper';


proc sort data = table1_data;
```

```
  by table_name ;
proc sort data = table1_compare;
  by table_name ;

data table1_combine;
  merge  table1_data (in = in2) table1_compare (in = in1);
  by table_name;
  if in1 and in2;

data table1_combine;
  set table1_combine;


  charall = compress(put(countall,8.) || '(' || put(PERtall,8.) || ')');
  char33 = compress(put(count33,8.) || '(' || put(pert33,8.) || ')');
  char34 = compress(put(count34,8.) || '(' || put(pert34,8.) || ')');
  char44 = compress(put(count44,8.) || '(' || put(pert44,8.) || ')');
  char48 = compress(put(count48,8.) || '(' || put(pert48,8.) || ')');

  diffall = compress(put(countall-countall_,8.) || '('  || put(PERtall-pertall_,8.) || ')');
  diff33  = compress(put(count33 -count33_ ,8.)  || '(' || put(pert33 -pert33_ ,8.) || ')');
  diff34  = compress(put(count34 -count34_ ,8.)  || '(' || put(pert34 -pert34_ ,8.) || ')');
  diff44  = compress(put(count44 -count44_ ,8.)  || '(' || put(pert44 -pert44_ ,8.) || ')');
  diff48  = compress(put(count48 -count48_ ,8.)  || '(' || put(pert48 -pert48_ ,8.) || ')');


  label
  character           = "Variable"
  charall_          = "Total [Manuscript]"
  charall           = "Total [DSIC]      "
  diffall           = "Total [Difference]"

  char33_           = "DR3-DQ2/DR3-DQ2 [Manuscript]"
  char33            = "DR3-DQ2/DR3-DQ2 [DSIC]      "
  diff33            = "DR3-DQ2/DR3-DQ2 [Difference]"

  char34_           = "DR3-DQ2/DR4-DQ8 [Manuscript]"
  char34            = "DR3-DQ2/DR4-DQ8 [DSIC]      "
  diff34            = "DR3-DQ2/DR4-DQ8 [Difference]"

  char44_           = "DR4-DQ8/DR4-DQ8 [Manuscript]"
  char44            = "DR4-DQ8/DR4-DQ8 [DSIC]      "
  diff44            = "DR4-DQ8/DR4-DQ8 [Difference]"

  char48_           = "DR4-DQ8/DR8-DQ4 [Manuscript]"
  char48            = "DR4-DQ8/DR8-DQ4 [DSIC]      "
  diff48            = "DR4-DQ8/DR8-DQ4 [Difference]"


  ;
```

```
proc sort data = table1_combine;
  by ordernum;


*** Outputting the data to a csv format to be added to the DSIC;
ods csv file = out_t1;
run;

proc print data = table1_combine NOOBS label;
       var
character     charall_ charall  diffall  char33_  char33   diff33 char34_  char34   diff34  char44_  char44   diff44  char48_  char48
diff48

;
       title "DSIC Check of Table 1 HLA Genotypes of 6403 Children Screened for Celiac Disease.*";
run;
```