Dataset Integrity Check for the TEDDY Pub26 BLee Data Files

Prepared by Jane Wang
IMS Inc.
3901 Calverton Blvd, Suite 200 Calverton MD 20705
July 16, 2015

Table of Contents

1 Standard Disclaimer 3
2 Study Background 3
3 Archived Datasets 3
4 Statistical Methods 4
5 Results 4
6 Conclusion 4
7 References 4

Attachment A: SAS Code 9

Table A: Variables used to replicate Table 1 5

Table B: Variables used to replicate Table 2 6

Table C: Comparison of values computed in integrity check to reference article Table 1 values: Infection and fever reported in the infection history study 7

Table D: Comparison of values computed in integrity check to reference article Table 2 values: Number of infections and fever reports per case or control in infection history study 8

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 “pub26_hlee_niddk_final” dataset.

4 Statistical Methods

Analyses were performed to duplicate results for the data published by H.-S. Lee, et al. [1] in Diabetologia in August 2013. To verify the integrity of the dataset, descriptive statistics of baseline characteristics were computed, by different case and control (Table B, Table C).

5 Results

Table A lists the variables that were used in the replication and Table B and C compares the results calculated from the archived data file to the results published in Table 1 and Table 2. The results of the replication are the same to the published results.

6 Conclusions

The NIDDK repository is confident that the TEDDY Pub26 BLee data files to be distributed are a true copy of the study data.

7 References

1. H.-S. Lee & T. Briese & C. Winkler & M. Rewers & E. Bonifacio & H. Hyoty & M. Pflueger & O. Simell & J. X. She & W. Hagopian & Å. Lernmark & B. Akolkar & J. P. Krischer & A. G. Ziegler & the TEDDY study group. Next-generation sequencing for viruses in children with rapid-onset type 1 diabetes. Diabetologia DOI 10.1007/s00125-013-2924-y

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

Table Variable	Variables Used in Replication from the Table 1 Dataset
Birth to type 1 diabetes	
Any infection	cillness_t
Fever	callfever_t
Fever without infectious illness	cfeveronly_t
Fever with any infectious illness	cfever_inf_t
Respiratory tract	crespg_t
Gastrointestinal tract	cgasg_t
Other	cotherg_t
Autoantibody-negative period	
Any infection	cillness_b
Fever	callfever_b
Fever without infectious illness	cfeveronly_b
Fever with any infectious illness	cfever_inf_b
Respiratory tract	crespg_b
Gastrointestinal tract	cgasg_b
Other	cotherg_b
Seroconversion period	
Any infection	cillness_bw
Fever	callfever_bw
Fever without infectious illness	cfeveronly_bw
Fever with any infectious illness	cfever_inf_bw
Respiratory tract	crespg_bw
Gastrointestinal tract	cgasg_bw
Other	cotherg_bw
Progression period	
Any infection	cillness_a
Fever	callfever_a
Fever without infectious illness	cfeveronly_a
Fever with any infectious illness	cfever_inf_a
Respiratory tract	crespg_a
Gastrointestinal tract	cgasg_a
Other	cotherg_a

Table B: Variables used to replicate Table 2 in the publication.

Table Variable	**Variables Used in Replication from the Table 2 Dataset**
Autoantibody-negative period	
Any infection	illness_b
Fever	allfever_b
Seroconversion period	
Any infection	illness_bw
Fever	allfever_bw
Progression period	
Any infection	illness_a
Fever	allfever_a

Table C: Comparison of values computed in integrity check to reference article Table 1 values: Most common reasons for staying in TEDDY (Reason as "Very Important by country")

Reasons for staying in TEDDY	**Case [Manuscript] N(%)**	**Case [DSIC] N(%)**	**Case [Difference] N(%)**	**Control [Manuscript] N(%)**	**Control [DSIC] N(%)**	**Control [Difference] N(%)**
Birth to type 1 diabetes						
Any infection	24(100)	24(100)	0(0)	70(97)	70(97)	0(0)
Fever	17(71)	17(71)	0(0)	62(86)	62(86)	0(0)
Fever without infectious illness	9(38)	9(38)	0(0)	22(31)	22(31)	0(0)
Fever with any infectious illness	14(58)	14(58)	0(0)	56(78)	56(78)	0(0)
Respiratory tract	24(100)	24(100)	0(0)	68(94)	68(94)	0(0)
Gastrointestinal tract	11(46)	11(46)	0(0)	29(40)	29(40)	0(0)
Other	5(21)	5(21)	0(0)	13(18)	13(18)	0(0)
Autoantibody-negative period						
Any infection	18(75)	18(75)	0(0)	53(74)	53(74)	0(0)
Fever	13(54)	13(54)	0(0)	38(53)	38(53)	0(0)
Fever without infectious illness	6(25)	6(25)	0(0)	8(11)	8(11)	0(0)
Fever with any infectious illness	9(38)	9(38)	0(0)	33(46)	33(46)	0(0)
Respiratory tract	18(75)	18(75)	0(0)	49(68)	49(68)	0(0)
Gastrointestinal tract	5(21)	5(21)	0(0)	16(22)	16(22)	0(0)
Other	2(8)	2(8)	0(0)	8(11)	8(11)	0(0)
Seroconversion period						
Any infection"	17(71)	17(71)	0(0)	52(72)	52(72)	0(0)
Fever	4(17)	4(17)	0(0)	30(42)	30(42)	0(0)
Fever without infectious illness	2(8)	2(8)	0(0)	11(15)	11(15)	0(0)
Fever with any infectious illness	3(13)	3(13)	0(0)	26(36)	26(36)	0(0)
Respiratory tract	15(63)	15(63)	0(0)	49(68)	49(68)	0(0)
Gastrointestinal tract	3(13)	3(13)	0(0)	10(14)	10(14)	0(0)

Other	1(4)	1(4)	0(0)	4(6)	4(6)	0(0)
Progression period						
Any infection	17(71)	17(71)	0(0)	48(67)	48(67)	0(0)
Fever	5(21)	5(21)	0(0)	30(42)	30(42)	0(0)
Fever without infectious illness	2(8)	2(8)	0(0)	8(11)	8(11)	0(0)
Fever with any infectious illness	5(21)	5(21)	0(0)	25(35)	25(35)	0(0)
Respiratory tract	15(63)	15(63)	0(0)	43(60)	43(60)	0(0)
Gastrointestinal tract	6(25)	6(25)	0(0)	10(14)	10(14)	0(0)
Other	2(8)	2(8)	0(0)	3(4)	3(4)	0(0)

Table D: Comparison of values computed in integrity check to reference article Table 2 values: Number of infections and fever reports per case or control in infection history study

Reasons for staying in TEDDY	**Case [Manuscript] mean (SD)**	**Case [DSIC] mean (SD)**	**Case [Difference] mean (SD)**	**Control [Manuscript] mean (SD)**	**Control [DSIC] mean (SD)**	**Control [Difference] mean (SD)**
Autoantibody-negative period						
Any infection	3.5(3.9)	3.5(3.9)	0(0)	2.8(3.6)	2.8(3.6)	0(0)
Fever	1.0(1.2)	1.0(1.2)	0(0)	1.1(1.5)	1.1(1.5)	0(0)
Seroconversion period						
Any infection	1.7(2.0)	1.7(2.0)	0(0)	2.0(1.9)	2.0(1.9)	0(0)
Fever	0.3(0.8)	0.3(0.8)	0(0)	0.8(1.3)	0.8(1.3)	0(0)
Progression period						
Any infection	1.6(1.5)	1.6(1.5)	0(0)	1.7(2.1)	1.7(2.1)	0(0)
Fever	0.4(0.9)	0.4(0.9)	0(0)	0.7(0.9)	0.7(0.9)	0(0)

Attachment A: SAS Code

```
*************************************************************************************************************************

***Program:

***Programmer: Jane Wang

***Date Created: 06/12/2015

***Purpose:

*************************************************************************************************************************;


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

title2 " ";


options nofmterr;


proc format;

 value CASE

0='control'

1='case'

;

 VALUE CCSHORT

 1='col'

 2='geo'

 3='was'
```

```
 4='fin'
 5='ger'
 6='swe'
 ;
 value esm
 1='in virus study'
 ;

value fdr
 0='gen pop'
 1='fdr'
 ;
value yesno
0='no'
1='yes'
;


libname sas_data "/prj/niddk/ims_analysis/TEDDY/private_orig_data/Pub26_HLee_niddk_submission/";
%include '/prj/niddk/ims_analysis/sas_macros/redaction_data_summary.sas';
*** Data from the Primary outcome 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_pub26_table1.csv';
```

```
FILENAME table2  '/prj/niddk/ims_analysis/TEDDY/private_created_data/teddy_pub26_table2.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_pub26_table1_dsic.csv';
FILENAME out_t2  '/prj/niddk/ims_analysis/TEDDY/private_created_data/teddy_pub26_table2_dsic.csv';

*** Reading in the analysis datasets used for the DSIC;
data pub26_hlee_niddk_final           ; set sas_data.pub26_hlee_niddk_final     ;

%FreqMeans(dsn=pub26_hlee_niddk_final                    , id_var=, cutoff=32, printcases=yes, ncases=20, libnm=work);

%macro baseline_freq1(dataset_name,var_name);

       *** Creating a frequency table in the format of Table 1 in the primary outcome paper;
  proc freq data = &dataset_name ;
       table (&var_name.)*outcome ;
       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;

  data &var_name._cross(keep =  outcome Frequency colpercent table_name);
```

```
  set &var_name._cross;
  if &var_name = 1 and outcome ne .;
  length table_name $30.;
  table_name  ="&var_name";

 proc sort data = &var_name._cross;
   by table_name outcome;

data &var_name._cross_1(drop =  outcome Frequency colpercent i);
  set &var_name._cross;
  by table_name;
  array temp1(2) control_count case_count ;
  array temp2(2) control_pert case_pert ;
  retain case_count control_count case_pert control_pert;
  if first.table_name then do i = 1 to 2;
    temp1(i) = .;
    temp2(i) = .;
  end;
    temp1(_n_) = Frequency;
    temp2(_n_) = colpercent;
  if last.table_name;
```

```
%mend;
%macro baseline_means(dataset_name,var_name);

       *** Creating a frequency table in the format of Table 1 in the primary outcome paper;
  proc means data = &dataset_name mean Std ;
       var &var_name.;
       by outcome;
       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 Summary = work.&var_name._means;
  run;

  data &var_name._means;
    set &var_name._means;
    length table_name $30.;
    table_name  ="&var_name";

   proc sort data = &var_name._means;
     by table_name outcome;
  data &var_name._means_1(drop = outcome &var_name._Mean &var_name._StdDev i);
    set &var_name._means;
```

```
    by table_name;
    array temp1(2) control_mean case_mean ;
    array temp2(2) control_std case_std ;
    retain case_mean control_mean case_std control_std;
    if first.table_name then do i = 1 to 2;
      temp1(i) = .;
      temp2(i) = .;
    end;
      temp1(_n_) = round(&var_name._Mean,0.1);
      temp2(_n_) = round(&var_name._StdDev,0.1);
    if last.table_name;
%mend;
*******************************************************************************;
************** Check Table 1 ************************************************;
*******************************************************************************;

*** Running the baseline_freq on the  categorical variables in the Table 1 manuscript file;
%baseline_freq1(pub26_hlee_niddk_final,cillness_t         );
%baseline_freq1(pub26_hlee_niddk_final,callfever_t        );
%baseline_freq1(pub26_hlee_niddk_final,cfeveronly_t       );
%baseline_freq1(pub26_hlee_niddk_final,cfever_inf_t       );
%baseline_freq1(pub26_hlee_niddk_final,crespg_t           );
```

```
%baseline_freq1(pub26_hlee_niddk_final,cgasg_t           );
%baseline_freq1(pub26_hlee_niddk_final,cotherg_t         );
%baseline_freq1(pub26_hlee_niddk_final,cillness_b        );
%baseline_freq1(pub26_hlee_niddk_final,callfever_b       );
%baseline_freq1(pub26_hlee_niddk_final,cfeveronly_b      );
%baseline_freq1(pub26_hlee_niddk_final,cfever_inf_b      );
%baseline_freq1(pub26_hlee_niddk_final,crespg_b          );
%baseline_freq1(pub26_hlee_niddk_final,cgasg_b           );
%baseline_freq1(pub26_hlee_niddk_final,cotherg_b         );
%baseline_freq1(pub26_hlee_niddk_final,cillness_bw       );
%baseline_freq1(pub26_hlee_niddk_final,callfever_bw      );
%baseline_freq1(pub26_hlee_niddk_final,cfeveronly_bw     );
%baseline_freq1(pub26_hlee_niddk_final,cfever_inf_bw     );
%baseline_freq1(pub26_hlee_niddk_final,crespg_bw         );
%baseline_freq1(pub26_hlee_niddk_final,cgasg_bw          );
%baseline_freq1(pub26_hlee_niddk_final,cotherg_bw        );
%baseline_freq1(pub26_hlee_niddk_final,cillness_a        );
%baseline_freq1(pub26_hlee_niddk_final,callfever_a       );
%baseline_freq1(pub26_hlee_niddk_final,cfeveronly_a      );
%baseline_freq1(pub26_hlee_niddk_final,cfever_inf_a      );
%baseline_freq1(pub26_hlee_niddk_final,crespg_a          );
%baseline_freq1(pub26_hlee_niddk_final,cgasg_a           );
```

```
%baseline_freq1(pub26_hlee_niddk_final,cotherg_a          );
data table1_compare;
  set
    cillness_t_cross_1
    callfever_t_cross_1
    cfeveronly_t_cross_1
    cfever_inf_t_cross_1
    crespg_t_cross_1
    cgasg_t_cross_1
    cotherg_t_cross_1
    cillness_b_cross_1
    callfever_b_cross_1
    cfeveronly_b_cross_1
    cfever_inf_b_cross_1
    crespg_b_cross_1
    cgasg_b_cross_1
```

```
cotherg_b_cross_1

cillness_bw_cross_1

callfever_bw_cross_1

cfeveronly_bw_cross_1

cfever_inf_bw_cross_1

crespg_bw_cross_1

cgasg_bw_cross_1

cotherg_bw_cross_1

cillness_a_cross_1

callfever_a_cross_1

cfeveronly_a_cross_1

cfever_inf_a_cross_1

crespg_a_cross_1

cgasg_a_cross_1

cotherg_a_cross_1
```

```
    ;

proc print data = table1_compare;

*** Importing the Table 1 Data taken from the primary outcome paper;
data table1_data;
  infile table1 delimiter = ',' MISSOVER DSD  firstobs=2 ls=1080;
  length character $100.   table_name $ 30.;
  input character      $ table_name $ case_char_ $ control_char_ $ or_char $        p $
;

data table1_data(drop = or_char p);
  set table1_data;
  ordernum = _n_;
  case_char_ = compress(case_char_) ;
  control_char_ = compress(control_char_);
  case_count_ = input(substr(case_char_,1,index(case_char_, '(')-1),8.);
  control_count_ =input( substr(control_char_,1,index(control_char_, '(')-1),8.);
  case_pert_ = input(substr(case_char_,index(case_char_, '(')+1, length(case_char_)-index(case_char_, '(')-1),8.);
  control_pert_ = input(substr(control_char_,index(control_char_, '(')+1, length(control_char_)-index(control_char_, '(')-1),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;
  case_char = compress(put(case_count,8.) || '(' || put(case_pert,8.) || ')');
  control_char = compress(put(control_count,8.) || '(' || put(control_pert,8.) || ')');
  case_diff = compress(put((case_count-case_count_),8.) || '(' || put((round(case_pert,1)-case_pert_),8.) || ')');
  control_diff = compress(put((control_count-control_count_),8.) || '(' || put((round(control_pert,1)-control_pert_),8.) || ')');
  label
  character           = "Reasons for staying in TEDDY"
  case_char_          = "Case [Manuscript]"
```

```
  case_char           = "Case [DSIC]      "
  case_diff           = "Case [Difference]"
  control_char_     = "Control [Manuscript]"
  control_char      = "Control [DSIC]      "
  control_diff      = "Control [Difference]"
  ;
proc sort data = table1_combine;
  by ordernum;

*******************************************************************************;
*************** Check Table 2 ************************************************;
*******************************************************************************;

*** Running the baseline_freq on the  categorical variables in the Table 2 manuscript file;
proc sort data = pub26_hlee_niddk_final;
  by outcome;
%baseline_means(pub26_hlee_niddk_final, illness_b      );
%baseline_means(pub26_hlee_niddk_final, allfever_b     );
%baseline_means(pub26_hlee_niddk_final, illness_bw     );
%baseline_means(pub26_hlee_niddk_final, allfever_bw    );
%baseline_means(pub26_hlee_niddk_final, illness_a      );
%baseline_means(pub26_hlee_niddk_final, allfever_a     );
```

```
data table2_compare;
  set
illness_b_means_1
allfever_b_means_1
illness_bw_means_1
allfever_bw_means_1
illness_a_means_1
allfever_a_means_1
;
proc print data = table2_compare;

*** Importing the Table 1 Data taken from the primary outcome paper;
data table2_data(drop = or_char p);
  infile table2 delimiter = ',' MISSOVER DSD  firstobs=3 ls=1080;
  length character $100.   table_name $ 30. case_char_ control_char_ $ 10.;
  input character      $ table_name $ case_char_ $ control_char_ $ or_char $        p $
;
  ordernum = _n_;
  case_char_ = compress(case_char_) ;
  control_char_ = compress(control_char_);
```

```
  case_mean_ = input(substr(case_char_,1,index(case_char_, '(')-1),8.);
  control_mean_ =input( substr(control_char_,1,index(control_char_, '(')-1),8.);
  case_std_ = input(substr(case_char_,index(case_char_, '(')+1, length(case_char_)-index(case_char_, '(')-1),8.);
  control_std_ = input(substr(control_char_,index(control_char_, '(')+1, length(control_char_)-index(control_char_, '(')-1),8.);
  if case_char_ ne '';

proc print data = table2_data;
  title3 'table 2 from paper';
proc sort data = table2_data;
  by table_name ;
proc sort data = table2_compare;
  by table_name ;

data table2_combine;
  merge  table2_data (in = in2) table2_compare (in = in1);
  by table_name;
  if in1 and in2;

data table2_combine;
  set table2_combine;
  case_char = compress(put(case_mean,8.1) || '(' || put(case_std,8.1) || ')');
  control_char = compress(put(control_mean,8.1) || '(' || put(control_std,8.1) || ')');
```

```
  case_diff = compress(put((case_mean-case_mean_),8.) || '(' || put((round(case_std,.1)-case_std_),8.) || ')');
  control_diff = compress(put((control_mean-control_mean_),8.) || '(' || put((round(control_std,.1)-control_std_),8.) || ')');
  label
  character           = "Reasons for staying in TEDDY"
  case_char_          = "Case [Manuscript]"
  case_char           = "Case [DSIC]      "
  case_diff           = "Case [Difference]"
  control_char_     = "Control [Manuscript]"
  control_char      = "Control [DSIC]      "
  control_diff      = "Control [Difference]"
  ;
proc sort data = table2_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
case_char_
```

```
case_char

case_diff

control_char_

control_char

control_diff


;

        title "DSIC Check of Table 1 Infection and fever reported in the infection history study";

run;


proc sort data = table2_combine;

  by ordernum;



*** Outputting the data to a csv format to be added to the DSIC;

ods csv file = out_t2;

run;


proc print data = table2_combine NOOBS label;

        var

character

case_char_
```

```
case_char

case_diff

control_char_

control_char

control_diff


;

        title "DSIC Check of Table 2 Number of infections and fever reports per case or control in infection history study";

run;



endsas;
```