Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) M21 Lynch

Prepared by Sabrina Chen
IMS Inc.
3901 Calverton Blvd, Suite 200 Calverton, MD 20705
April 9, 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_21_klynch_niddk_31aug2016.sas7bdat" dataset.

4 Statistical Methods

Analyses were performed to duplicate results for the data published by KLynch et al [1] in *Journal of Autoimmunity* in 2018. To verify the integrity of the dataset, descriptive statistics were computed.

5 Results

For Table 1 in the publication [1], Gestational Respiratory Infection (G-RI) in relation to IAA and GADA as the first appearing b-cell autoantibodies by genetic factors, 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 similar to the published results.

6 Conclusions

The NIDDK repository is confident that the TEDDY M21 data files to be distributed are a true copy of the study data.

7 References

Kristian F.Lynch, Hye-SeungLee, CarinaTörn, KendraVehik, Jeffrey P.Krischer, Helena EldingLarsson, Michael J.Haller, William A.Hagopian, Marian J.Rewers, Jin-Xiong She, Olli G.Simell, JormaToppari, Anette-G.Ziegler, BeenaAkolkar, HeikkiHyöty, EzioBonifacio, ÅkeLernmark, on behalf of TEDDY Study Group. Gestational respiratory infections interacting with offspring HLA and CTLA-4 modifies incident b-cell autoantibodies. *Journal of Autoimmunity* 86 (2018) 93-103.

Table A: Variables used to replicate Table 1: Gestational Respiratory Infection (G-RI) in relation to IAA and GADA as the first appearing b-cell autoantibodies by genetic factors

Table Variable	**dataset.variable**
Mother had a gestational respiratory infection - 0 = no, 1 = yes	m_21_klynch_niddk_31aug2016.resp_gest_inf
Indicates whether or not the subject had serconverted for Islet Autoantibodies and had only GADA at seroconversion ? 0=no, 1=yes	m_21_klynch_niddk_31aug2016.ia_gad_only
Indicates whether or not the subject had serconverted for Islet Autoantibodies and had only IAA at seroconversion ? 0=no, 1=yes	m_21_klynch_niddk_31aug2016.ia_iaa_only
First degree relative status - 0=General Pop, 1=FDR father or sibling but not mother	m_21_klynch_niddk_31aug2016.fdr
HLA genotypes groups (1 = DR-DQ 3-2/4-8, 2= DR-DQ 4-8/4-8, 3= DR-DQ 4-8/8-4, 4 = HLA-DR-DQ 3-2 or 4-8/X, 5 = HLA-DR-DQ 3-2/3-2)	m_21_klynch_niddk_31aug2016.hla_5grps
Child had SNP rs2476601 in PTPN22 - 0 = no, only major alleles; 1 = yes, minor allele	m_21_klynch_niddk_31aug2016.rs2476601_ptpn22
Child had SNP rs689 in INS-23 Hph1 - 0 = no, only major alleles; 1 = yes, minor allele	m_21_klynch_niddk_31aug2016.rs689_ins
Child had SNP rs231775 in CTLA4 - 0 = no, only major alleles; 1 = yes, minor allele	m_21_klynch_niddk_31aug2016.rs231775_ctla4
Child had SNP rs2292239 in ERBB3 - 0 = no, only major alleles; 1 = yes, minor allele	m_21_klynch_niddk_31aug2016.rs2292239_erbb3
Child had SNP rs3184504 in SH2B3 - 0 = no, only major alleles; 1 = yes, minor allele	m_21_klynch_niddk_31aug2016.rs3184504_sh2b3
Child had SNP rs4948088_CLEC16A - 0 = no, only major alleles; 1 = yes, minor allele	m_21_klynch_niddk_31aug2016.rs12708716_clec16a
Child had SNP rs10517086 - 0 = no, only major alleles; 1 = yes, minor allele	m_21_klynch_niddk_31aug2016.rs10517086
Child had SNP rs4948088 in COBL - 0 = no, only major alleles; 1 = yes, minor allele	m_21_klynch_niddk_31aug2016.rs4948088_cobl

Table Variable	**dataset.variable**
Child had SNP rs2816316 in RGS1 - 0 = no, only major alleles; 1 = yes, minor allele	m_21_klynch_niddk_31aug2016.rs2816316_rgs1

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

	COVARF	Total			IAA Event			GAD Event		
		DSIC	Manuscript	Diff	DSIC	Manuscript	Diff	DSIC	Manuscript	Diff
All children	Yes	7058	7058	0	211	211	0	206	206	0
General Population (GP)	No	6530	6530	0	171	171	0	175	175	0
First Degree Relative (FDR)	Yes	528	528	0	40	44	-4	31	31	0
HLA-DQ	DR-DQ 4-8/4-8	1396	1396	0	36	36	0	31	31	0
	DR-DQ 3-2/4-8	2778	2778	0	101	101	0	103	103	0
	DR-DQ 4-8/8-4	1221	1221	0	49	49	0	21	21	0
	HLA-DR-DQ 3-2/3-2	1510	1510	0	16	16	0	50	50	0
rs2476601	No	4991	4991	0	144	144	0	146	146	0
(PTPN22-A)	Yes	1265	1265	0	64	64	0	57	57	0
rs689	No	3404	3404	0	153	153	0	106	106	0
(INS-A)	Yes	2853	2853	0	55	55	0	98	98	0
rs231775	No	1971	1971	0	72	72	0	52	52	0
(CTLA4-G)	Yes	4286	4286	0	136	136	0	152	152	0
rs2292239	No	2657	1657	1000	70	70	0	74	74	0
(ERBB3-T)	Yes	3104	3104	0	124	124	0	118	118	0
rs3184504	No	1813	1813	0	50	50	0	40	40	0
(SH2B3-T)	Yes	3948	3948	0	144	144	0	152	152	0
rs12708716	No	2528	2528	0	91	91	0	96	96	0
(CLEC16A_G)	Yes	3218	3218	0	103	103	0	96	96	0
rs10517086-A	No	2951	2951	0	93	93	0	105	105	0
	Yes	2810	2810	0	101	101	0	87	87	0
rs4948088	No	5240	5240	0	185	185	0	178	178	0

	COVARF	Total			IAA Event			GAD Event		
		DSIC	Manuscript	Diff	DSIC	Manuscript	Diff	DSIC	Manuscript	Diff
(COBL-A)	Yes	521	521	0	9	9	0	14	14	0
rs2816316	No	3845	3845	0	129	129	0	130	130	0
(RGS1-C)	Yes	1916	1916	0	65	65	0	62	62	0

Attachment A: SAS Code

```
options nocenter validvarname=upcase;

title '/prj/niddk/ims_analysis/TEDDY/prog_initial_analysis/m_21_dsic.sas';
run;

/*
                              */
/* "The entire dataset is the analytical dataset.
                              */
/* It was an apriori hypothesis only to pull singleton children of non-diabetic mothers with HLA eligibility and determined islet
autoantibodies by 6 years of age (n=7472)    */
/*
                              */
/* The variable or exposure of interest is gestational respiratory infection called resp_gest_inf (0=no, 1=yes, .=missing)
                                     */
/* And there is 7058 children available data with 414 with missing data on this variable.
                              */
/*
                              */
/* But the 414 may have data on other infections that were not the main part of the manuscript.
                              */
/*
                              */
/* The outcomes of interest are IA_IAA_only and IA_GAD_only."
                                     */
/*
                              */



***********;
* INPUT    ;
***********;
libname pcsas '/prj/niddk/ims_analysis/TEDDY/private_orig_data/M_21_KLynch_NIDDK_Submission/';


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


%macro univ(rownum, var, subset, subsetname);

  proc univariate data=analy outtable= univ&subsetname noprint;
    where &subset=1;
    var &var
       ;
  run;

  data univ&subsetname;
    length covarf $100;
    set univ&subsetname;
        covarf = "&subset";
        rownum = &rownum;
  run;
```

```
  data prntuniv&subsetname;
    set prntuniv&subsetname univ&subsetname;
  run;

%mend;



*************;
* FORMATS    ;
*************;
proc format;
  value novalue
      . = "No Value"
  other = "   Value"
  ;

  value $nochar
  " "   = "No Value"
  other = "   Value"
  ;

  value negpos
  0 = "Negative"
  1 = "Positive"
  ;

  value yesno
  . = "Missing"
  0 = "No"
  1 = "Yes"
  ;

  value gender
  0 = "Male"
  1 = "Female"
  ;

  value hlagp
  1 = 'DR-DQ 3-2/4-8'
  2 = 'DR-DQ 4-8/4-8'
  3 = 'DR-DQ 4-8/8-4'
  4 = 'HLA-DR-DQ 3-2 or 4-8/X'
  5 = 'HLA-DR-DQ 3-2/3-2'
  ;
```

```
run;



%readin(pcsas, m_21_klynch_niddk_31aug2016);


proc freq data=m_21_klynch_niddk_31aug2016;
  tables resp_gest_inf/missing;
run;

proc freq data=m_21_klynch_niddk_31aug2016;
  where resp_gest_inf ne .;
  tables ia_iaa_only ia_gad_only fdr hla_5grps rs2476601_ptpn22
rs10517086
rs12708716_clec16a
rs2292239_erbb3
rs231775_ctla4
rs2476601_ptpn22
rs2816316_rgs1
rs3184504_sh2b3
rs4948088_cobl
rs689_ins
/missing;
  run;

proc freq data=m_21_klynch_niddk_31aug2016;
  where resp_gest_inf ne . and IA_IAA_ONLY=1;
tables  IA_IAA_ONLY*(fdr hla_5grps rs2476601_ptpn22
rs10517086
rs12708716_clec16a
rs2292239_erbb3
rs231775_ctla4
rs2476601_ptpn22
rs2816316_rgs1
rs3184504_sh2b3
rs4948088_cobl
rs689_ins)
/list missing;


proc freq data=m_21_klynch_niddk_31aug2016;
  where resp_gest_inf ne . and ia_gad_only=1;
tables  ia_gad_only*(fdr hla_5grps rs2476601_ptpn22
```

```
rs10517086
rs12708716_clec16a
rs2292239_erbb3
rs231775_ctla4
rs2476601_ptpn22
rs2816316_rgs1
rs3184504_sh2b3
rs4948088_cobl
rs689_ins)
/list missing;



data analy;
  set m_21_klynch_niddk_31aug2016;

  if resp_gest_inf ne . then subset_all = 1;

  if resp_gest_inf ne . and IA_IAA_ONLY=1 then subset_iaaevent=1;
  if resp_gest_inf ne . and IA_IAA_ONLY=1 and rs2476601_ptpn22   ne . then rs247_IAAsubset = 1;
  if resp_gest_inf ne . and IA_IAA_ONLY=1 and rs689_ins      ne . then rs689_IAAsubset = 1;
  if resp_gest_inf ne . and IA_IAA_ONLY=1 and rs231775_ctla4     ne . then rs231_IAAsubset = 1;
  if resp_gest_inf ne . and IA_IAA_ONLY=1 and rs2292239_erbb3       ne . then rs229_IAAsubset = 1;
  if resp_gest_inf ne . and IA_IAA_ONLY=1 and rs3184504_sh2b3    ne . then rs318_IAAsubset = 1;
  if resp_gest_inf ne . and IA_IAA_ONLY=1 and rs12708716_clec16a ne . then rs127_IAAsubset = 1;
  if resp_gest_inf ne . and IA_IAA_ONLY=1 and rs10517086         ne . then rs105_IAAsubset = 1;
  if resp_gest_inf ne . and IA_IAA_ONLY=1 and rs4948088_cobl     ne . then rs494_IAAsubset = 1;
  if resp_gest_inf ne . and IA_IAA_ONLY=1 and rs2816316_rgs1     ne . then rs281_IAAsubset = 1;

  if resp_gest_inf ne . and ia_gad_only=1 then subset_gadevent=1;
  if resp_gest_inf ne . and ia_gad_only=1 and rs2476601_ptpn22   ne . then rs247_gadsubset = 1;
  if resp_gest_inf ne . and ia_gad_only=1 and rs689_ins      ne . then rs689_gadsubset = 1;
  if resp_gest_inf ne . and ia_gad_only=1 and rs231775_ctla4     ne . then rs231_gadsubset = 1;
  if resp_gest_inf ne . and ia_gad_only=1 and rs2292239_erbb3       ne . then rs229_gadsubset = 1;
  if resp_gest_inf ne . and ia_gad_only=1 and rs3184504_sh2b3    ne . then rs318_gadsubset = 1;
  if resp_gest_inf ne . and ia_gad_only=1 and rs12708716_clec16a ne . then rs127_gadsubset = 1;
  if resp_gest_inf ne . and ia_gad_only=1 and rs10517086         ne . then rs105_gadsubset = 1;
  if resp_gest_inf ne . and ia_gad_only=1 and rs4948088_cobl     ne . then rs494_gadsubset = 1;
  if resp_gest_inf ne . and ia_gad_only=1 and rs2816316_rgs1     ne . then rs281_gadsubset = 1;

run;


** Table 1;
```

```
* n percent;
data prntall;
  set _null_;
run;

%npercent(1   , subset_all        , yesno , subset_all , all);
%npercent(2   , fdr                     , yesno , subset_all , all);
%npercent(3   , hla_5grps         , hlagp , subset_all , all);
%npercent(5   , rs2476601_ptpn22    , yesno , subset_all , all);
%npercent(6   , rs689_ins           , yesno , subset_all , all);
%npercent(7   , rs231775_ctla4      , yesno , subset_all , all);
%npercent(8   , rs2292239_erbb3     , yesno , subset_all , all);
%npercent(9   , rs3184504_sh2b3     , yesno , subset_all , all);
%npercent(10  , rs12708716_clec16a  , yesno , subset_all , all);
%npercent(11  , rs10517086          , yesno , subset_all , all);
%npercent(12  , rs4948088_cobl      , yesno , subset_all , all);
%npercent(13  , rs2816316_rgs1      , yesno , subset_all , all);

proc print data=prntall;
  where covarf ne "Missing";
  var rownum covar covarf count    percent;
run;


* subset events col;
data prntiaaevents;
  set _null_;
run;

%npercent(1   , IA_IAA_ONLY       , yesno , subset_iaaevent , iaaevents);
%npercent(2   , fdr                     , yesno , subset_iaaevent , iaaevents);
%npercent(3   , hla_5grps         , hlagp , subset_iaaevent , iaaevents);
%npercent(5   , rs2476601_ptpn22    , yesno , rs247_IAAsubset , iaaevents);
%npercent(6   , rs689_ins           , yesno , rs689_IAAsubset , iaaevents);
%npercent(7   , rs231775_ctla4      , yesno , rs231_IAAsubset , iaaevents);
%npercent(8   , rs2292239_erbb3     , yesno , rs229_IAAsubset , iaaevents);
%npercent(9   , rs3184504_sh2b3     , yesno , rs318_IAAsubset , iaaevents);
%npercent(10  , rs12708716_clec16a  , yesno , rs127_IAAsubset , iaaevents);
%npercent(11  , rs10517086          , yesno , rs105_IAAsubset , iaaevents);
%npercent(12  , rs4948088_cobl      , yesno , rs494_IAAsubset , iaaevents);
%npercent(13  , rs2816316_rgs1      , yesno , rs281_IAAsubset , iaaevents);

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

```
data prntgadevents;
  set _null_;
run;

%npercent(1   , IA_gad_ONLY       , yesno , subset_gadevent , gadevents);
%npercent(2   , fdr                     , yesno , subset_gadevent , gadevents);
%npercent(3   , hla_5grps         , hlagp , subset_gadevent , gadevents);
%npercent(5   , rs2476601_ptpn22    , yesno , rs247_gadsubset , gadevents);
%npercent(6   , rs689_ins           , yesno , rs689_gadsubset , gadevents);
%npercent(7   , rs231775_ctla4      , yesno , rs231_gadsubset , gadevents);
%npercent(8   , rs2292239_erbb3     , yesno , rs229_gadsubset , gadevents);
%npercent(9   , rs3184504_sh2b3     , yesno , rs318_gadsubset , gadevents);
%npercent(10  , rs12708716_clec16a  , yesno , rs127_gadsubset , gadevents);
%npercent(11  , rs10517086          , yesno , rs105_gadsubset , gadevents);
%npercent(12  , rs4948088_cobl      , yesno , rs494_gadsubset , gadevents);
%npercent(13  , rs2816316_rgs1      , yesno , rs281_gadsubset , gadevents);

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


proc sort data=prntall;
  where covarf ne "Missing";
  by rownum covar covarf;
run;

proc sort data=prntiaaevents (rename=(count=count_iaa));
  by rownum covar covarf;
run;

proc sort data=prntgadevents (rename=(count=count_gad));
  by rownum covar covarf;
run;

data table1;
  merge prntall       (in=in1 keep= rownum covar covarf count)
        prntiaaevents (in=in2 keep= rownum covar covarf count_iaa)
        prntgadevents (in=in3 keep= rownum covar covarf count_gad)
        ;
  by rownum covar covarf;
  if in1 or in2 or in3;
```

```
  label count    = 'Total'
        count_iaa = 'IAA Event'
        count_gad = 'GAD Event'
        ;
run;

proc print data=table1 label;
  var rownum covar covarf count count_iaa count_gad;
run;

ods listing close;
ods phtml file="/prj/niddk/ims_analysis/TEDDY/private_created_data/teddy_m_21_table1.xls";

proc print data=table1 label;
  var rownum covar covarf count count_iaa count_gad;
run;
```