Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) M68 LFerrat

Prepared by Sabrina Chen
IMS Inc.
3901 Calverton Blvd, Suite 200 Calverton, MD 20705
Aug 7, 2020

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/private_orig_data/M_68_LFerrat_NIDDK_Submission folder in the data package. For this replication, variables were taken from the “mp68_demographics_masked.csv”, “mp68_family_histories_masked.csv", and “mp68_subject_list.csv" datasets.

4 Statistical Methods

Analyses were performed to duplicate results for the data published by Lauric Ferrat et al [1] in Nature Medicine 2020. To verify the integrity of the dataset, descriptive statistics were computed.

5 Results

For **Comparison of Data in the publication**, 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.

6 Conclusions

The results of the replication are almost an exact match to the published results.

7 References

[1] Ferrat, L.A., Vehik, K., Sharp, S.A. *et al.* A combined risk score enhances prediction of type 1 diabetes among susceptible children. *Nat Med* (2020). https://doi.org/10.1038/s41591-020-0930-4.

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

Table Variable	**dataset.variable**
Id	mp68_subject_list.mp68_maskid
Country	mp68_demographics_masked.country_cd
FDR	mp68_demographics_masked.fdr
HLA Category	mp68_demographics_masked.hla_category
Sex	mp68_demographics_masked.sex
C Section	mp68_demographics_masked.c_section
T1D	mp68_demographics_masked.t1d
Father's diabetes type	mp68_family_histories_masked.fatherdiabetestype
Mother's diabetes type	mp68_family_histories_masked.motherdiabetestype
Sibling's diabetes type	mp68_family_histories_masked.sibling1diabetestype
Sibling's diabetes type	mp68_family_histories_masked.sibling2diabetestype
Sibling's diabetes type	mp68_family_histories_masked.sibling3diabetestype
Sibling's diabetes type	mp68_family_histories_masked.sibling4diabetestype
Sibling's diabetes type	mp68_family_histories_masked.sibling5diabetestype
Sibling's diabetes type	mp68_family_histories_masked.sibling6diabetestype
Sibling's diabetes type	mp68_family_histories_masked.sibling7diabetestype
Sibling's diabetes type	mp68_family_histories_masked.sibling8diabetestype
Sibling's diabetes type	mp68_family_histories_masked.sibling9diabetestype

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

		non T1D			T1D		
		Manuscript	**DSIC**	**Diff**	**Manuscript**	**DSIC**	**Diff**
Country	USA	3143	3143	0	103	103	0
	Finland	1612	1612	0	89	89	0
	Germany	507	507	0	28	28	0
	Sweden	2231	2231	0	85	85	0
First degree relative with T1D	no	6691	6691	0	221	221	0
	yes	802	802	0	84	84	0
Mother T1D	no, missing	7214	7214	0	283	283	0
	yes	279	279	0	22	22	0
Father T1D	no, missing	7131	7131	0	260	260	0
	yes	362	362	0	45	45	0
Siblings T1D	no, missing	7381	7376	5	280	280	0
	yes	112	117	-5	25	25	0
HLA genotype	other	340	340	0	17	17	0
	DR4/DR3 (DR4*030X/0302*DR3*0501/0201)	2835	2835	0	166	166	0
	DR4/DR4 (DR4*030X/0302*DR4*030X/0302)	1464	1464	0	55	55	0
	DR4/DR8 (DR4*030X/0302*DR8*0401/0402)	1277	1277	0	42	42	0
	DR3/DR3 (DR3*0501/0201*DR3*0501/0201)	1577	1577	0	25	25	0
Sex	Female	3684	3684	0	148	148	0
	Male	3809	3809	0	157	157	0
Caesarean section	no	5542	5542	0	223	223	0
	yes	1951	1951	0	82	82	0

Attachment A: SAS Code

```
options nocenter validvarname=upcase;

title '/prj/niddk/ims_analysis/TEDDY/prog_initial_analysis/m_68_dsic.sas';
run;

proc import datafile="/prj/niddk/ims_analysis/TEDDY/private_orig_data/M_68_LFerrat_NIDDK_Submission/mp68_demographics_masked.csv"
  dbms=csv replace
  out=work.dems;
run;

proc import datafile="/prj/niddk/ims_analysis/TEDDY/private_orig_data/M_68_LFerrat_NIDDK_Submission/mp68_family_histories_masked.csv"
  dbms=csv replace
  out=work.fam;
run;

proc import datafile="/prj/niddk/ims_analysis/TEDDY/private_orig_data/M_68_LFerrat_NIDDK_Submission/mp68_subject_list.csv"
  dbms=csv replace
  out=work.subject;
run;




proc format;
 value val
 .     = "no value"
 other = "   value"
 ;

 value oneplus
 . = "no value"
 0 = "0"
 1-high = "1+"
 ;

 value zerohi
 . = "no value"
 0-high = "0-high"
 ;

 value countryf
 1 = 'US'
 2 = 'Finland'
 3 = 'Germany'
 4 = 'Sweden'
 ;
```

```
 value fdr
 1='(B) FDR'
 0='(A) GenPop'
 ;

 value hlaf
 -1='HLA*Results*Pending'
 0 ='Not*Eligible'
 1 ='DR4*030X/0302*DR3*0501/0201'
 2 ='DR4*030X/0302*DR4*030X/0302'
 3 ='DR4*030X/0302*DR4*030X/020X'
 4 ='DR4*030X/0302*DR8*0401/0402'
 5 ='DR4*030X/0302*DR1*0101/0501'
 6 ='DR4*030X/0302*DR13*0102/0604'
 7 ='DR4*030X/0302*DR4*030X/0304'
 8 ='DR4*030X/0302*DR9*030X/0303'
 9 ='DR3*0501/0201*DR3*0501/0201'
 10='DR3*0501/0201*DR9*030X/0303'
 99='Results*Under*Review'
;

 value hlagpf
 -1,0,3,7,5,6,8,10='other'
/* -1='HLA*Results*Pending'
 0='Not*Eligible' */
 1     ='DR4/DR3 (DR4*030X/0302*DR3*0501/0201)'
 2     ='DR4/DR4 (DR4*030X/0302*DR4*030X/0302)'
 /*3 ='DR4*030X/0302*DR4*030X/020X'  */
 4     ='DR4/DR8 (DR4*030X/0302*DR8*0401/0402)'
/* 5='DR4*030X/0302*DR1*0101/0501'
   6='DR4*030X/0302*DR13*0102/0604'
   7 ='DR4*030X/0302*DR4*030X/0304'
   8='DR4*030X/0302*DR9*030X/0303' */
 9     ='DR3/DR3 (DR3*0501/0201*DR3*0501/0201)'
/* 10    ='DR3/DR9' /*'DR3*0501/0201*DR9*030X/0303'*/
 99    ='Results*Under*Review'
;

  value t1df
  0 = "non T1D"
  1 = "T1D"
  ;

  value famt1df
  .,0 = "no, missing"
  1   = "yes"
  ;
```

```
  value yesno
  0 = "no"
  1 = "yes"
  ;

run;



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



proc contents data=subject;
run;

proc contents data=dems;
run;

proc freq data=dems;
  tables status*exclude*T1D/list missing;
format MP68_MASKID  val.;
title3 "dems file";
run;

proc contents data=fam;
  title3 "family histories file";
run;
```

```
proc sort data=subject;
  by mp68_maskid;
run;

proc sort data=dems;
  by mp68_maskid;
run;

proc sort data=fam;
  by mp68_maskid;
run;

proc freq data=fam;
  tables sibling8diabetestype sibling9diabetestype/missing;
  tables siblingdiabetestype1_1 *    siblingdiabetestype2_1   *    siblingdiabetestype3_1  *     siblingdiabetestype4_1    *
siblingdiabetestype5_1/list missing;
  tables siblingdiabetic1 * siblingdiabetic2 * siblingdiabetic3 * siblingdiabetic4 * siblingdiabetic5 * siblingdiabetic6 * siblingdiabetic7 *
siblingdiabetic8 * siblingdiabetic9/list missing;
  title3 "family histories file - checking";
run;

* It is the first line which appears for each child (information known at birth of the child). ;
data famfirst;
  set fam;
  by mp68_maskid;

  array sibling (9) $ sibling1diabetestype sibling2diabetestype sibling3diabetestype sibling4diabetestype sibling5diabetestype
                      sibling6diabetestype sibling7diabetestype sibling8diabetestype sibling9diabetestype;

  if first.mp68_maskid then do;
    siblingt1d = 0;
    fathert1d = 0;
    mothert1d = 0;

    do i=1 to 9;
      if sibling(i) = "Type 1" then siblingt1d=1;
    end;

    if fatherdiabetestype = "Type 1" then fathert1d=1;
    if motherdiabetestype = "Type 1" then mothert1d=1;

    output;
  end;
run;

proc freq data=famfirst;
  tables siblingt1d*sibling1diabetestype* sibling2diabetestype* sibling3diabetestype* sibling4diabetestype* sibling5diabetestype*
         sibling6diabetestype* sibling7diabetestype/list missing nocum nopercent;
```

```
  tables siblingt1d/missing;
  tables fathert1d * fatherdiabetestype/list missing;
  tables mothert1d * motherdiabetestype/list missing;
  title3 "family histories file - check Table vars";
run;


data combine;
  merge subject  (in=in1 keep=mp68_maskid)
        dems     (in=in2 keep=mp68_maskid country_cd fdr hla_category sex c_section t1d)
        famfirst (in=in3 keep=mp68_maskid fatherdiabetic motherdiabetic fatherdiabetestype motherdiabetestype fathert1d mothert1d
                          sibling1diabetestype sibling2diabetestype sibling3diabetestype sibling4diabetestype sibling5diabetestype
                          sibling6diabetestype sibling7diabetestype sibling8diabetestype sibling9diabetestype siblingt1d);
  by mp68_maskid;
  if in1;
run;

proc freq data=combine;
  tables t1d*(country_cd fdr mothert1d fathert1d  siblingt1d hla_category sex c_section)/list missing nopercent nocum;
  format  country_cd countryf. fdr yesno.  mothert1d fathert1d  siblingt1d famt1df. hla_category hlagpf. t1d t1df.;
  title3 "Table S6";
run;
```