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 TEDDY M17 JBaxter Data Files

Prepared by David Ruggieri
IMS Inc.
3901 Calverton Blvd, Suite 200 Calverton MD 20705
July 21, 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 SAS data files are located in the TEDDY M17 JBaxter data package. For this replication, variables were taken from the “m_17_jbaxter_niddk_a_final” dataset.

4 Statistical Methods

Analyses were performed to duplicate results for the data published by Judith Baxter et al [1] in Contemporary Clinical Trials in 2012. To verify the integrity of the dataset, descriptive statistics were computed (Table 1).

5 Results

Table 1 in the publication [1], Description of TEDDY Study recruitment and enrollment experience in the US Clinical Centers by ethnic minority group status as of December 2009. Our Table A lists the variables we used in our replication and Table B compare the results calculated from the archived data file to the results published in Table 1. The results of the replication are identical to those listed in the table, with an exception of the family reason as a refusal to enroll.

6 Conclusions

The NIDDK repository is confident that the TEDDY data files to be distributed are within expected results.

7 References

1. Judith Baxter, Kendra Vehik, Suzanne Bennett Johnson, Barbro Lernmar , Roswith Roth, Tuula Simell for the TEDDY Study Group. Differences in Recruitment and Early Retention among Ethnic Minority Participants in a Large Pediatric Cohort: The Environmental Determinants of Diabetes in the Young (TEDDY) Study. Contemp Clin Trials. 2012 Jul;33(4): 633-640.

Table A: Variables used to replicate Table 1: Description of TEDDY Study recruitment and enrollment experience in the US Clinical Centers by ethnic minority group status as of December 2009

Table Variable	Variables used in replication from the Table 1 Dataset
Ethnic Minority Group	race_ethnic
Exclusion Flag	excluded
First Degree Relative	fdr
Primary reasons for exclusion	exclusion
Total refusing to enroll	refreason
Enrollment flag	enroll
Clinical Center	cc
Maternal age	maternal_age
Child's gender	sex

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

Ethnic Minority Group	NHW [manuscript]	NHW [DSIC]	HIS [manuscript]	HIS [DSIC]	AA [manuscript]	AA [DSIC]	OM [manuscript]	OM [DSIC]	All [manuscript]	All [DSIC]
Number of excluded children (% of HLA eligible children)	2615	2615	1136	1136	215	215	76	76	4063	4063
General Population	2532	2532	1121	1121	209	209	73	73	3955	3955
First Degree Relative	83	83	15	15	6	6	3	3	108	108
Primary reasons for exclusion: Number excluded by reason (% of children excluded)										
No response to calls/messages	1988	1988	840	840	126	126	54	54	3008	3008
Incorrect contact information	156	156	147	147	32	32	9	9	344	344
Unable to schedule visit by 4.5 months	409	409	130	130	53	53	10	10	602	602
Characteristics of TEDDY eligible and invited participants										
Total	5231	5231	1343	1343	180	180	115	115	6912	6912
General Population	4884	4884	1279	1279	169	169	111	111	6483	6483
First Degree Relative	347	347	64	64	11	11	4	4	429	429
Total refusing to enroll (% within group)	2724	2724	726	726	88	88	89	89	3647	3647
No reason given	346	346	104	104	19	19	18	18	490	490
Moving	150	150	45	45	3	3	7	7	208	208
Wait and see	123	123	38	38	4	4	8	8	173	173
Protocol too demanding	1088	1088	262	262	23	23	19	19	1399	1399
Family reasons	1019	1017	277	277	39	39	37	37	1379	1377
Characteristics of enrolled subjects										
Total	2507	2507	617	617	92	92	26	26	3265	3265
Enrollment rate (% within group)										
General Population	2241	2241	574	574	85	85	23	23	2943	2943
First Degree Relative	266	266	43	43	7	7	3	3	322	322

Clinical center (% within group)										
Colorado	768	768	440	440	9	9	6	6	1224	122
Georgia/Florida	734	734	56	56	66	66	5	5	879	879
Washington	1005	1005	121	121	17	17	15	15	1162	1162
Maternal age mean years (SD)	30.9	30.9	27.4	27.4	27.4	27.4	30	30	30.1	30.1
Child's gender										
% Female	1206	1206	333	333	45	45	12	12	1610	1610

```
%let flnm = %sysfunc(getoption(sysin));
title "File saved as &FLNM.";
title2 "Check the tables on the check.m_17_niddk.BaxterJ_JClinEpi_EthDiffs paper";

options nofmterr;

/*****************************************************************************************************************************************
Programmer: Dave Ruggieri
Date: 18 July 2014
Billing Code:
Requestor:

Function/Notes: Check the tables on the check.m_17_niddk.BaxterJ_JClinEpi_EthDiffs paper.
*****************************************************************************************************************************************/
*  Input file directory  *;
**************************;
libname inlib "/prj/niddk/ims_analysis/TEDDY/private_orig_data/teddy_m_17_jbaxter/";

*** Input files ***;
data afinal;
  set inlib.m_17_jbaxter_niddk_a_final;

data bfinal;
  set inlib.m_17_jbaxter_niddk_b_final;

data cfinal;
  set inlib.m_17_jbaxter_niddk_c_final;

*************;
*  Formats  *;
*************;
proc format;
  value ethn
    1='HIS'
    2='NHW'
    3='AA'
    4,5='OM'
    ;

  value eth2n
    1='NHW'
    2='HIS'
    3='AA'
    4,5='OM'
    ;

  value fdrfmt
    0 = 'General Population'
    1 = 'First Degree Relative'
```

```
    ;

  value excfmt
    1 = 'No test result'
    2 = 'Appt not scheduled'
    3 = 'Appt not in window'
    4 = 'Incorrect contact'
    5 = 'No response'
    6 = 'Illness or birth defect'
    7 = 'Refuse repository'
    8 = 'No reason given'
    ;

  value refreas
    1 = 'No reason'
    2 = 'Moving'
    3 = 'Wait and See'
    4 = 'Protocol too demanding'
    5 = 'Family Reasons'
    6 = 'Unknown'
    ;

  value ccfmt
    1   = 'Colorado'
    2   = 'Georgia'
    3   = 'Washington'
    4   = 'FIN'
    5   = 'GER'
    6   = 'SWE'
    133 = 'NBD'
    134 = 'CHP'
    ;

*************;
*  Table 1  *;
*************;
%macro tab1 (subst, subval, subdesc, frqvar, frqfmt);
proc freq data = afinal;
  where &SUBST. = &SUBVAL.;
  tables race_ethnic &FRQVAR.
         /missing list;
  format race_ethnic ethn. &FRQVAR. &FRQFMT..;
  title4 "Distribution of ethnicity and &FRQVAR.";
  title5 "Dataset is subset to &SUBDESC. patients";

proc freq data = afinal;
  where &SUBST. = &SUBVAL. and race_ethnic ^= .;
  tables race_ethnic*&FRQVAR.
         /missing list;
```

```
  format race_ethnic ethn. &FRQVAR. &FRQFMT..;
  title4 "Distribution of ethnicity by &FRQVAR.";
  title5 "Dataset is subset to &SUBDESC. patients";
%mend tab1;

proc sort data = afinal;
  by race_ethnic;

%tab1(subst=excluded,subval=1,subdesc=excluded,frqvar=fdr,frqfmt=fdrfmt);

proc freq data = afinal;
  tables exclusion*inelig_cat1*inelig_cat2*inelig_cat3*inelig_cat4*inelig_cat5*excluded
         /missing list;
  format exclusion excfmt.;
  title4 'Exclusion by various ineligibility flags';

proc freq data = afinal;
       where excluded = 1 and race_ethnic ^= .;
  tables exclusion /missing list;
  format exclusion excfmt.;
  title4 'Exclusion flag where patients are excluded and have race';

%tab1(subst=excluded,subval=1,subdesc=excluded,frqvar=exclusion,frqfmt=excfmt);
%tab1(subst=excluded,subval=0,subdesc=not excluded,frqvar=fdr,frqfmt=fdrfmt);
%tab1(subst=excluded,subval=0,subdesc=not excluded,frqvar=refreason,frqfmt=refreas);
%tab1(subst=enroll,subval=1,subdesc=enrolled,frqvar=fdr,frqfmt=fdrfmt);
%tab1(subst=enroll,subval=1,subdesc=enrolled,frqvar=cc,frqfmt=ccfmt);

proc means data = afinal;
  where enroll = 1 and race_ethnic ^= .;
  by race_ethnic;
  var maternal_age;
  format race_ethnic ethn.;
  title4 'Maternal age mean years stratified by race.';
  title5 'Subset to enrolled patients who have non-missing race';

proc means data = afinal;
  where enroll = 1 and race_ethnic ^= .;
  var maternal_age;
  format race_ethnic ethn.;
  title4 'Maternal age mean years.';
  title5 'Subset to enrolled patients who have non-missing race';

proc freq data = afinal;
  where enroll = 1;
  tables sex
         sex*race_ethnic
         /missing list;
  format race_ethnic ethn.;
```

```
title4 'Gender by ethnic race';
title5 'Subset to enrolled patients';
```