Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) M202 Mehta

Prepared by NIDDK-CR
October 13, 2022

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 type 1 diabetes (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. Data are 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.

The M202 study sought to describe gluten-free diet adherence rates in children with screening-identified celiac disease and determine adherence-related factors.

3 Archived Datasets

All 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_202_pmehta_niddk_30jun2019_1.sas7bdat" dataset.

4 Statistical Methods

Analyses were performed to replicate results for the data in the manuscript by Mehta et al. [1]. To verify the integrity of the data, only descriptive statistics were computed. The TEDDY DCC provided the NIDDK Central Repository with participant ID lists corresponding to the sub-groups analyzed in the manuscript for the purposes of this replication.

5 Results

For Table 1 in the manuscript [1], Subject characteristics at 2 and 5 years post the diagnosis of celiac disease stratified by self-reported gluten-free diet adherence, 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 in Table 1. Information on location type, child age at first form submission, gastrointestinal (GI) symptoms, and non-GI symptoms were not included in the analysis data and were excluded from this replication. For the data available, the results of the replication are within expected variation to the manuscript results.

6 Conclusions

The NIDDK Central Repository is confident that the TEDDY M202 data files to be distributed are a true copy of the study data.

7 References

[1] Mehta P, Li Q, Stahl M, Uusitalo U, Lindfors K, Butterworth MD, Kurppa K, Virtanen S, Koletzko S, Aronsson C, Hagopian WA, Rewers MJ, Toppari J, Ziegler AG, Akolkar B, Krischer JP, Agardh D, Liu E. (2022). Gluten-free Diet Adherence in Children with Screening-detected Celiac Disease using a Prospective Birth Cohort Study. (Manuscript submitted for publication)

Table A: Variables used to replicate Table 1 – Subject characteristics at 2 and 5 years post the diagnosis of celiac disease stratified by self-reported gluten-free diet adherence

Table Variable	**dataset.variable**
Sex	m_202_pmehta_niddk_30jun2019_1.sex
Country	m_202_pmehta_niddk_30jun2019_1.country
Mean child age at diagnosis in months	m_202_pmehta_niddk_30jun2019_1.cd_diag_age_days
Mean maternal age at diagnosis in years	m_202_pmehta_niddk_30jun2019_1.mom_age_at_cd
Mean paternal age at diagnosis in years	m_202_pmehta_niddk_30jun2019_1.dad_age_at_cd
Highest degree of maternal education	m_202_pmehta_niddk_30jun2019_1.education_mom_group3
Highest degree of paternal education	m_202_pmehta_niddk_30jun2019_1.education_dad_group3
Comorbid Type 1 DM	m_202_pmehta_niddk_30jun2019_1.t1d
First degree relative with celiac disease	m_202_pmehta_niddk_30jun2019_1.celiac_fdr

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

Characteristic	2 years post CD diagnosis						5 years post CD diagnosis					
	Pub: Adherent (n=290)	**DSIC: Adherent (n=290)**	**Diff. (n=0)**	**Pub: Nonadherent (n=110)**	**DSIC: Nonadherent (n=110)**	**Diff. (n=0)**	**Pub: Adherent (n=199)**	**DSIC: Adherent (n=199)**	**Diff. (n=0)**	**Pub: Nonadherent (n=97)**	**DSIC: Nonadherent (n=97)**	**Diff. (n=0)**
Sex (%)												
Female	187 (65)	187 (65)	0 (0)	65 (59)	65 (59)	0 (0)	132 (66)	132 (66)	0 (0)	58 (60)	58 (60)	0 (0)
Male	103 (36)	103 (36)	0 (0)	45 (41)	45 (41)	0 (0)	67 (34)	67 (34)	0 (0)	39 (40)	39 (40)	0 (0)
Country (%)												
Finland	51 (18)	51 (18)	0 (0)	23 (21)	23 (21)	0 (0)	37 (19)	37 (19)	0 (0)	18 (19)	18 (19)	0 (0)
Germany	11 (4)	11 (4)	0 (0)	5 (5)	5 (5)	0 (0)	9 (5)	9 (5)	0 (0)	4 (4)	4 (4)	0 (0)
Sweden	148 (51)	148 (51)	0 (0)	50 (45)	50 (45)	0 (0)	101 (51)	101 (51)	0 (0)	48 (49)	48 (49)	0 (0)
United States	80 (28)	80 (28)	0 (0)	32 (29)	32 (29)	0 (0)	52 (26)	52 (26)	0 (0)	27 (28)	27 (28)	0 (0)
Mean child age at diagnosis in months (± SD)	55.5 (23.5)	56 (23.5)	0.5 (0)	60.4 (23.5)	61 (23.5)	0.6 (0)	47.9 (18.5)	48.4 (18.5)	0.5 (0)	48.6 (18.6)	49.0 (18.6)	0.4 (0)
Mean maternal age at diagnosis in years (± SD)	35.9 (4.9)	35.9 (4.9)	0 (0)	35.7 (5.3)	35.7 (5.3)	0 (0)	35.3 (4.9)	35.3 (4.9)	0 (0)	34.7 (4.9)	34.7 (4.9)	0 (0)
Mean paternal age at diagnosis in years (± SD)	38.8 (5.4)	38.8 (5.4)	0 (0)	38.8 (8.0)	38.9 (8.0)	0.1 (0)	37.8 (5.7)	37.8 (5.7)	0 (0)	38.8 (7.2)	38.8 (7.2)	0 (0)
Highest degree of maternal education (%)												
Primary education – some trade school	51 (18)	51 (18)	0 (0)	22 (20)	22 (20)	0 (0)	34 (17)	34 (17)	0 (0)	18 (19)	18 (19)	0 (0)
Graduated trade school or some college	49 (17)	49 (17)	0 (0)	26 (24)	26 (24)	0 (0)	29 (15)	29 (15)	0 (0)	22 (23)	22 (23)	0 (0)
Graduated college or higher degree	188 (65)	188 (65)	0 (0)	58 (52)	58 (52)	0 (0)	56 (58)	134 (67)	78 (9)	134 (67)	56 (58)	78 (9)
Missing Data	2 (1)	2 (1)	0 (0)	4 (4)	4 (4)	0 (0)	2 (1)	2 (1)	0 (0)	1 (1)	1 (1)	0 (0)
Highest degree of paternal education (%)												
Primary education – some trade school	76 (26)	76 (26)	0 (0)	35 (32)	35 (32)	0 (0)	49 (25)	49 (25)	0 (0)	30 (31)	30 (31)	0 (0)
Graduated trade school or some college	68 (24)	68 (23)	0 (1)	21 (19)	21 (19)	0 (0)	48 (24)	48 (24)	0 (0)	17 (18)	17 (18)	0 (0)
Graduated college or higher degree	142 (49)	142 (49)	0 (0)	47 (43)	47 (43)	0 (0)	99 (50)	99 (50)	0 (0)	46 (47)	46 (47)	0 (0)
Missing Data	4 (1)	4 (1)	0 (0)	7 (6)	7 (6)	0 (0)	3 (2)	3 (2)	0 (0)	4 (4)	4 (4)	0 (0)
Comorbid Type 1 DM (%)	6 (2)	6 (2)	0 (0)	5 (5)	5 (5)	0 (0)	2 (1)	2 (1)	0 (0)	3 (3)	3 (3)	0 (0)
First degree relative with celiac disease (%)	60 (21)	60 (21)	0 (0)	14 (13)	14 (13)	0 (0)	47 (24)	47 (24)	0 (0)	13 (13)	13 (13)	0 (0)

Attachment A: SAS Code

```
libname m202 "X:\NIDDK\niddk-
dr_studies6\TEDDY\private_orig_data\M_202_PMehta_NIDDK_Submission";

/**************************/
/* DSIC for TEDDY M202 */
/* Submission Mehta et al. */
/**************************/

*temp datasets;
data one; set m202.m_202_pmehta_niddk_30jun2019_1;
run;

data two; set m202.m_202_pmehta_niddk_30jun2019_2;
run;

data three; set m202.m_202_pmehta_niddk_30jun2019_3;
run;

*identifying the 2-yr group;
proc sort data=one;
by maskid;
run;

proc sort data=work.mp202_niddk_2yr_list;
by MaskID;
run;

data one_1;
merge
one (in=a)
work.mp202_niddk_2yr_list (in=b);
by maskid;
if b=1;
run ;

*identifying the adherent vs. non-adherent within the 2-yr group;
data one_2; set one_1;
adh = 0;
if rate_adh < 1 then adh = 0; *non-adherent group;
if rate_adh = 1 then adh = 1; *adherent group;
run ;

proc freq data=one_2;
tables adh;
```

```
run;

*sex for the 2-yr group;
proc freq data=one_2;
tables sex*adh/norow nopercent;
run;

*country for the 2-yr group;
proc freq data=one_2;
tables country*adh/norow nopercent;
run;

*Location for the 2-yr group;
*mean child age at diagnosis for the 2-yr group;
data one_3; set one_2;
diag_age_mo = (cd_diag_age_days/365.25)*12;
run ;

proc means data=one_3 mean std;
var diag_age_mo;
class adh;
run;

*mean maternal age at diagnosis in years for the 2-yr group;
proc means data=one_3 mean std;
var mom_age_at_cd;
class adh;
run ;

*mean paternal age at diagnosis in years for the 2-yr group;
proc means data=one_3 mean std;
var dad_age_at_cd;
class adh;
run;

*highest degree of maternal education;
proc freq data=one_3;
tables education_mom_group3*adh/norow nopercent missing;
run;

*highest degree of paternal education for the 2-yr group;
proc  freq data=one_3;
tables education_dad_group3*adh/norow nopercent missing;
run;

*t1d for the 2-yr group;
proc freq data=one_3;
tables t1d*adh/norow nopercent;
```

```
run;

*first degree relative with celiac disease in 2-yr group;
proc freq data=one_3;
tables celiac_fdr*adh/norow nopercent;
run;

*5-yr group;
proc sort data=work.mp202_niddk_5yr_list;
by maskid;
run;

data one_4;
merge
one (in=a)
work.mp202_niddk_5yr_list (in=b);
by maskid;
if b=1;
run;

*adherent vs non-adherent in the 5 yr group;
data one_5; set one_4;
adh=0;
if rate_adh < 1 then adh = 0;
if rate_adh = 1 then adh = 1;
run;

proc freq data=one_5;
tables adh;
run;

*Sex for the 5-yr group;
proc freq data=one_5;
tables sex*adh/norow nopercent;
run;

*country for the 5-yr group;
proc freq data=one_5;
tables country*adh/norow nopercent;
run;

*mean child age at diagnosis for the 5-yr group;
data one_6; set one_5;
age_diag_mo = (cd_diag_age_days/365.25)*12;
run;

proc means data=one_6 mean std;
var age_diag_mo;
```

```
class adh;
run;

*mean maternal age at diagnosis for the 5-yr group;
proc means data=one_6 mean std;
var mom_age_at_cd;
class adh;
run;

*mean paternal age at diagnosis for the 5-yr group;
proc means data=one_6 mean std;
var dad_age_at_cd;
class adh;
run;

*higest degree of maternal education for 5-yr group;
proc freq data=one_6;
tables education_mom_group3*adh/ norow nopercent missing;
run;

*highest degree of paternal education for 5-yr group;
proc freq data=one_6;
tables education_dad_group3*adh/norow nopercent missing;
run;

*t1d;
proc freq data=one_6;
tables t1d*adh/ norow nopercent;
run;

*fdr with celiac disease;
proc freq data=one_6;
tables celiac_fdr*adh/norow nopercent;
run;
```