Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) M199 Lindfors

Prepared by NIDDK-CR
April 17, 2024

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. 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.

The M199 study sought to understand whether distinct viral exposures alone or together with gluten increased the risk of coeliac disease autoimmunity (CDA) in genetically predisposed children.

3 Archived Datasets

A full listing of the archived datasets included in the package can be found in the Roadmap document. 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_199_lindfors_niddk_31may2012_2.sas7bdat", and "m_199_lindfors_niddk_31may2012_7" datasets.

4 Statistical Methods

Analyses were performed to replicate results for the data in the publication by Lindfors et al. [1]. To verify the integrity of the data, only descriptive statistics were computed.

5 Results

For Table 1 in the publication [1], Demographic data of the 83 nested case and control pairs, 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. The results of the replication are within expected variation to the published results.

6 Conclusions

The NIDDK Central Repository is confident that the TEDDY M199 data files to be distributed are a true copy of the study data.

7 References

[1] Lindfors K, Lin J, Lee HS, Hyöty H, Nykter M, Kurppa K, Liu E, Koletzko S, Rewers M, Hagopian W, Toppari J, Ziegler AG, Akolkar B, Krischer JP, Petrosino JF, Lloyd RE, Agardh D. Metagenomics of the Faecal Virome Indicate a Cumulative Effect of Enterovirus and Gluten Amount on the Risk of Coeliac Disease Autoimmunity in Genetically At Risk Children: The TEDDY Study. Gut, 69(8), 1416-1422, August 2020. doi: https://doi.org/10.1136/gutjnl-2019-319809

Table A: Variables used to replicate Table 1 – Demographic data of the 83 nested case and control pairs

Table Variable	**dataset.variable**
HLA genotype	m_199_lindfors_niddk_31may2012_2.hlarg m_199_lindfors_niddk_31may2012_2.outcome
Age at CDA	m_199_lindfors_niddk_31may2012_2mcda_age m_199_lindfors_niddk_31may2012_2.outcome
Developed CD during follow-up	m_199_lindfors_niddk_31may2012_2.cda_age m_199_lindfors_niddk_31may2012_7.cd m_199_lindfors_niddk_31may2012_2.outcome
Breastfeeding	m_199_lindfors_niddk_31may2012_2.ever_brstfed m_199_lindfors_niddk_31may2012_2.outcome m_199_lindfors_niddk_31may2012_2.time_to_brstfed_stop
Gluten	m_199_lindfors_niddk_31may2012_2.gluten m_199_lindfors_niddk_31may2012_2.gl_sum

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

Characteristic	Pub: Cases (n=83)	DSIC: (n=83)	Diff. (n=0)	Pub: Controls (n=83)	DSIC: Controls (n=83)	Diff. (n=0)
HLA genotype, n (%)						
DQ2/DQ2	29 (35)	29 (35)	0 (0)	11 (13)	11 (13)	0 (0)
DQ2/DQ8	39 (47)	39 (47)	0 (0)	36 (43)	36 (44)	0 (1)
DQ8/DQ8	11 (13)	11 (13)	0 (0)	20 (24)	20 (24)	0 (0)
DQ8/X	4 (5)	4 (5)	0 (0)	15 (19)	14 (17)	1 (2)
Other (ineligible)	0 (0)	0 (0)	0 (0)	1 (1)	1 (1)	0 (0)
Age at CDA, months	31 (23, 46)	31 (23, 46)	0 (0)	NA	NA	NA
Developed CD during follow-up, n (%)	28 (34)	28 (34)	0 (0)	NA	NA	NA
Breastfeeding						
Ever breastfed, n (%)	83 (100)	83 (100)	0 (0)	83 (100)	83 (100)	0 (0)
Breastfeeding stopped, months	8 (5, 11)	8 (5, 11)	0 (0)	8 (4, 12)	8 (4, 12)	0 (0)
Gluten						
Age at introduction, months	6 (5, 7)	6 (5, 7)	0 (0)	6 (5, 7)	6 (5, 7)	0 (0)
Total intake by 2 years of age (g)	8.0 (5.4, 11.0)	8.0 (5.4, 11.0)	0 (0)	7.6 (5.0, 11.8)	7.6 (5.0, 11.8)	0 (0)

Attachment A: SAS Code

```
libname m199 "X:\NIDDK\niddk-dr_studies6\TEDDY\private_created_data\M199, M208, and
M226\M_199_Lindfors_NIDDK_Submission";

/***********************/
/* TEDDY M199 Lindfors */
/* DSIC */
/***********************/

data two; set m199.m_199_lindfors_niddk_31may2012_2;
run;

*HLA;
proc freq data=two;
tables hlarg*outcome/norow nopercent;
run;

*age at CDA;
proc means data=two_2 median q1 q3;
var mcda_age;
class outcome;
run;

*developed CD during follow up;
proc freq data=two;
tables cda_age case_ind;
run;

*developed CD;
data seven; set m199.m_199_lindfors_niddk_31may2012_7;
drop country;
run;

proc sort data=two;
by maskid;
run;

proc sort data=seven;
by maskid;
run;

data cd; merge
two (in=a)
seven (in=b);
by maskid;
if a=b;
```

```
run;

proc freq data=cd;
tables cd*outcome/ norow nopercent;
run;

*breastfeeding;
proc freq data=two;
tables ever_brstfed*outcome/norow nopercent;
run;

*age breastfeeding stopped;
data two; set two;
brfd_months = time_to_brstfed_stop/30;
run;

proc means data=two n median q1 q3;
var brfd_months;
class outcome;
run;

*gluten introduction;
data two; set two;
gl_months = gluten/30;
run;

proc means data=two n median q1 q3;
var gl_months gl_sum;
class outcome;
run;
```