# Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) M93 Kemppainen

**Prepared by NIDDK-CR**
**October 17, 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 manuscript.

# 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 M93 study sought to understand the pathogenic mechanism of gluten immunogenicity in patients with celiac disease.

# 3 Archived Datasets

All 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_93_kkemp_niddk_31mar2015_1.sas7bdat" dataset.

# 4 Statistical Methods

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

# 5 Results

For Table 1 in the publication [1], Proportional Hazards Model of Factors Associated With CDA or Rate of Childhood Infections on the Risk of CDA Between 1 and 4 Years, 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 M93 data files to be distributed are a true copy of the study data.

# 7 References

[1] Kemppainen KM, Lynch KF, Liu E, Lönnrot M, Simell V, Briese T, Koletzko S, Hagopian W, Rewers M, She JX, Simell O, Toppari J, Ziegler AG, Akolkar B, Krischer JP, Lernmark Å, Hyöty H, Triplett EW, Agardh D. Factors That Increase Risk of Celiac Disease Autoimmunity After a Gastrointestinal Infection in Early Life. Clinical Gastroenterology and Hepatology, 15(5), 694-702.e5, May 2017. doi: https://doi.org/10.1016/j.cgh.2016.10.033

**Table A:** Variables used to replicate Table 1 – Proportional Hazards Model of Factors Associated With CDA or Rate of Childhood Infections on the Risk of CDA Between 1 and 4 Years

| **Table Variable** | **dataset.variable** |
|---|---|
| Country | m_93_kkemp_niddk_31mar2015_1.country |
| Sex | m_93_kkemp_niddk_31mar2015_1.sex |
| HLA-DR-DQ genotype | m_93_kkemp_niddk_31mar2015_1.hla_celiac_3grps |
| Season of birth | m_93_kkemp_niddk_31mar2015_1.sept_to_feb_birth |
| First-degree relative with celiac disease | m_93_kkemp_niddk_31mar2015_1.celiac_fdr_yes |
| Maternal age | m_93_kkemp_niddk_31mar2015_1.maternal_age |
| Maternal education | m_93_kkemp_niddk_31mar2015_1.education_mom_group3 |
| Only child in household at 9 months of age | m_93_kkemp_niddk_31mar2015_1.single_child |
| Mode of delivery | m_93_kkemp_niddk_31mar2015_1.csection |
| Age at start of daycare | m_93_kkemp_niddk_31mar2015_1.daycare_grps |
| Duration of any breastfeeding | m_93_kkemp_niddk_31mar2015_1.brst_feed_grps |
| Age at introduction to gluten | m_93_kkemp_niddk_31mar2015_1.gluten_feed_grps |

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

| **Variable** | **Publication: Total (n=6327)** | **DSIC: Total (n=6327)** | **Diff. (n=0)** |
|---|---:|---:|---:|
| Country, n (%) | | | |
| United States | 2545 (40.2) | 2545 (40.2) | 0 (0) |
| Finland | 1436 (22.7) | 1436 (22.7) | 0 (0) |
| Germany | 338 (5.3) | 338 (5.3) | 0 (0) |
| Sweden | 2008 (31.7) | 2008 (31.7) | 0 (0) |
| Sex, n (%) | | | |
| Male | 3253 (51.4) | 3253 (51.4) | 0 (0) |
| Female | 3074 (48.6) | 3074 (48.6) | 0 (0) |
| HLA-DR-DQ genotype, n (%) | | | |
| DQ8/8 or DQ4/DR4-DQ8 | 2385 (37.7) | 2385 (37.7) | 0 (0) |
| DQ2/DQ8 | 2581 (40.8) | 2581 (40.8) | 0 (0) |
| DQ2/DQ2 | 1361 (21.5) | 1361 (21.5) | 0 (0) |
| Season of birth, n (%) | | | |
| Summer (March-August) | 1613 (49.4) | 3123 (49.4) | 1510 (0) |
| Winter (September-February) | 1591 (50.6) | 3204 (50.6) | 1613 (0) |
| First-degree relative with celiac disease, n (%) | | | |
| No | 6125 (96.8) | 6125 (96.8) | 0 (0) |
| Yes | 202 (3.2) | 202 (3.2) | 0 (0) |
| Maternal age, years, median (IQR) | 31 (27-34) | 31 (27-34) | 0 (0) |
| Maternal education, n (%) | | | |
| High school | 1157 (18.7) | 1157 (18.7) | 0 (0) |
| Trade school or some college | 1484 (24.0) | 1484 (24.0) | 0 (0) |
| College degree or more | 3542 (57.3) | 3542 (57.29) | 0 (0.01) |
| Only child in household at 9 months of age, n (%) | | | |
| No | 3576 (57.8) | 3576 (57.8) | 0 (0) |
| Yes | 2608 (42.2) | 2608 (42.2) | 0 (0) |
| Mode of delivery, n (%) | | | |
| Vaginal | 4709 (74.5) | 4709 (74.5) | 0 (0) |
| Caesarian section | 1614 (25.5) | 1614 (25.5) | 0 (0) |
| Age at start of daycare, n (%) | | | |
| < 4 months | 2312 (36.5) | 2312 (36.5) | 0 (0) |
| 4 to < 8 months | 1292 (20.4) | 1292 (20.4) | 0 (0) |
| 8 to < 12 months | 683 (10.8) | 683 (10.8) | 0 (0) |
| ≥ 12 months | 2040 (32.2) | 2040 (32.2) | 0 (0) |
| Duration of any breastfeeding, n (%) | | | |
| < 4 months | 1681 (26.6) | 1681 (26.6) | 0 (0) |
| 4 to < 8 months | 1471 (23.2) | 1471 (23.2) | 0 (0) |
| 8 to < 12 months | 1679 (26.4) | 1679 (26.4) | 0 (0) |
| ≥ 12 months | 1440 (22.8) | 1440 (22.8) | 0 (0) |

| **Variable** | **Publication: Total (n=6327)** | **DSIC: Total (n=6327)** | **Diff. (n=0)** |
|---|---:|---:|---:|
| Age at introduction to gluten, n (%) | | | |
| ≤ 4 months | 1226 (19.4) | 1226 (19.4) | 0 (0) |
| 5 months | 1436 (22.7) | 1436 (22.7) | 0 (0) |
| 6 months | 1444 (22.8) | 1444 (22.8) | 0 (0) |
| ≥ 7 months | 2220 (35.1) | 2220 (35.1) | 0 (0) |

# Attachment A: SAS Code

```
libname m93 "X:\NIDDK\niddk-dr_studies6\TEDDY\private_orig_data\M_93_KKemppainen_NIDDK_Submission";

/**********************************/
/* DSIC for M93 Kemppainen et al. */
/**********************************/

data one; set m93.m_93_kkemp_niddk_31mar2015_1;
run;

*Country;
proc freq data=one;
tables country;
run;

*Sex;
proc freq data=one;
tables sex;
run;

*HLA-DR-DQ genotype;
proc freq data=one;
tables hla_celiac_3grps;
run;

*seaon of birth;
proc freq data=one;
tables sept_to_feb_birth;
run;

*FDR with celiac disease;
proc freq data=one;
tables celiac_fdr_yes;
run;

*Maternal age;
proc means data=one n median q1 q3;
var maternal_age;
run;

*Maternal education;
proc freq data=one;
tables education_mom_group3;
run;
```

```
*only child in household at 9 months;
proc freq data=one;
tables single_child;
run;

*mode of delivery;
proc freq data=one;
tables csection;
run;

*age at start of daycare;
proc freq data=one;
tables daycare_grps;
run;

*duration of breastfeeding;
proc freq data=one;
tables brst_feed_grps/missing;
run;

*age at introduction to gluten;
proc freq data=one;
tables gluten_feed_grps/missing;
run;
```