# Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) Pub95 Uusitalo

**Prepared by Allyson Mateja**
**IMS Inc.**
3901 Calverton Blvd, Suite 200 Calverton, MD 20705
**October 3, 2016**

# 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 folder in the data package. For this replication, variables were taken from the "m_95_uuusitalo_niddk_31oct2014.sas7bdat" dataset.

# 4 Statistical Methods

Analyses were performed to duplicate results for the data published by Uusitalo et al [1] in JAMA Pediatrics in 2015.  To verify the integrity of the dataset, descriptive statistics were computed.

# 5 Results

For Table 1 in the publication [1], <u>Distribution of Probiotic Exposure From Dietary Supplements and Infant Formulas During the First Year of Life by Age and Country</u>, Table A lists the variables that were used in the replication and Table B compares the results calculated from the archived data file to the results published in Table 1. The results of the replication are almost an exact match to the published results, with only a few minor discrepancies.

For Table 2 in the publication [1], <u>Characteristics of Probiotic Supplement and/or Probiotic Formula Users and Nonusers</u>, Table C lists the variables that were used in the replication and Table D compares the results calculated from the archived data file to the results published in Table 2. The results of the replication are almost an exact match to the published results, with only a few discrepancies due to rounding.

For Table 3 in the publication [1], <u>First Probiotic Exposure of the Child vis Infant Formula and/or Dietary Supplement During the First Year of Life and Risk of IA</u>, Table E lists the variables that were used in the replication and Table F compares the results calculated from the archived data file to the results published in Table 3. The results of the replication are almost an exact match to the published results, with only a few minor discrepancies.

# 6 Conclusions

The NIDDK repository is confident that the TEDDY M95 data files to be distributed are a true copy of the study data.

# 7 References

[1] Uusitalo, U., Liu, X., Yang, J., Aronsson, C.A., Hummel, S., Butterworth, M., Lernmark, A., Rewers, M., Hagopian, W., She, J., Simell, O., Toppari, J., Ziegler, A.G., Akolkar, B., Krischer, J., Norris, J., Virtanen, S.M., and the TEDDY study group. "Association of Early Exposure of Probiotics and Islet Autoimmunity in the TEDDY Study". JAMA Pediatrics (2015).

**Table A:** Variables used to replicate Table 1: Distribution of Probiotic Exposure From Dietary Supplements and Infant Formulas During the First Year of Life by Age and Country

| Table Variable | Variable |
|---|---|
| Country | country |
| Probiotic use during first 12 mo | prob |
| Probiotic use during first 3 mo | early_prob_3mon |
| Age at first exposure to probiotics | prob_startdy |
| Dietary supplements | probsupp |
| Infant formula | probform |
| Timing of first probiotic exposure | early_prob_ter |

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

| Variable | United States Manuscript (n=3046) | United States DSIC (n=3046) | Diff. (n=0) | Finland Manuscript (n=1658) | Finland DSIC (n=1658) | Diff. (n=0) |
|---|---|---|---|---|---|---|
| Probiotic Use | | | | | | |
| During first 12 mo | 186 (6.1) | 186 (6.1) | 0 (0) | 869 (52.4) | 869 (52.4) | 0 (0) |
| During first 3 mo | 70 (2.3) | 73 (2.4) | 3 (0.1) | 627 (37.8) | 627 (37.8) | 0 (0) |
| Age at first exposure to probiotics, median (IQR), d | 137 (56-244) | 137 (56-244) | 0 (0-0) | 28 (14-105) | 28 (14-105) | 0 (0-0) |
| Source of first probiotic exposure among probiotic users | | | | | | |
| Dietary supplements | 124 (66.7) | 124 (66.7) | 0 (0) | 827 (95.2) | 827 (95.2) | 0 (0) |
| Infant formula | 53 (28.5) | 53 (28.5) | 0 (0) | 25 (2.9) | 25 (2.9) | 0 (0) |
| Both | 9 (4.8) | 9 (4.8) | 0 (0) | 17 (1.9) | 17 (2.0) | 0 (0.1) |
| Timing of first probiotic exposure for users only, d | | | | | | |
| 0-27 | 20 (10.8) | 20 (10.8) | 0 (0) | 344 (39.5) | 344 (39.6) | 0 (0.1) |
| 28-90 | 53 (28.5) | 53 (28.5) | 0 (0) | 283 (32.6) | 283 (32.6) | 0 (0) |
| 91-365 | 113 (60.7) | 113 (60.8) | 0 (0.1) | 242 (27.9) | 242 (27.9) | 0 (0) |

| Variable | Germany Manuscript (n=506) | Germany DSIC (n=506) | Diff. (n=0) | Sweden Manuscript (n=2263) | Sweden DSIC (n=2263) | Diff. (n=0) |
|---|---|---|---|---|---|---|
| Probiotic Use | | | | | | |
| During first 12 mo | 237 (46.8) | 237 (46.8) | 0 (0) | 345 (15.2) | 345 (15.3) | 0 (0.1) |

| Variable | Germany Manuscript (n=506) | Germany DSIC (n=506) | Diff. (n=0) | Sweden Manuscript (n=2263) | Sweden DSIC (n=2263) | Diff. (n=0) |
|---|---|---|---|---|---|---|
| During first 3 mo | 123 (24.3) | 123 (24.3) | 0 (0) | 276 (12.2) | 276 (12.2) | 0 (0) |
| Age at first exposure to probiotics, median (IQR), d | 84 (14-198) | 84 (14-198) | 0 (0-0) | 35 (21-70) | 35 (21-70) | 0 (0-0) |
| Source of first probiotic exposure among probiotic users | | | | | | |
| Dietary supplements | 11 (4.6) | 11 (4.6) | 0 (0) | 274 (79.4) | 274 (79.4) | 0 (0) |
| Infant formula | 214 (90.3) | 214 (90.3) | 0 (0) | 40 (11.6) | 40 (11.6) | 0 (0) |
| Both | 12 (5.1) | 12 (5.1) | 0 (0) | 31 (9.0) | 31 (9.0) | 0 (0) |
| Timing of first probiotic exposure for users only, d | | | | | | |
| 0-27 | 71 (31.2) | 74 (31.2) | 3 (0) | 104 (30.0) | 104 (30.1) | 0 (0.1) |
| 28-90 | 49 (20.7) | 49 (20.7) | 0 (0) | 172 (50.0) | 172 (49.9) | 0 (0.1) |
| 91-365 | 114 (48.1) | 114 (48.1) | 0 (0) | 69 (20.0) | 69 (20.0) | 0 (0) |

| Variable | All Manuscript (N=7473) | All DSIC (N=7473) | Diff. (N=0) |
|---|---|---|---|
| Probiotic Use | | | |
| During first 12 mo | 1637 (21.9) | 1637 (21.9) | 0 (0) |
| During first 3 mo | 1096 (14.7) | 1099 (14.7) | 3 (0) |
| Age at first exposure to probiotics, median (IQR), d | 42 (14-152) | 42 (14-152) | 0 (0-0) |
| Source of first probiotic exposure among probiotic users | | | |
| Dietary supplements | 1236 (75.5) | 1236 (75.5) | 0 (0) |
| Infant formula | 332 (20.3) | 332 (20.3) | 0 (0) |
| Both | 69 (4.2) | 69 (4.2) | 0 (0) |
| Timing of first probiotic exposure for users only, d | | | |
| 0-27 | 542 (33.1) | 542 (33.1) | 0 (0) |
| 28-90 | 557 (34.0) | 557 (34.0) | 0 (0) |
| 91-365 | 538 (32.9) | 538 (32.9) | 0 (0) |

**Table C:** Variables used to replicate Table 2: Characteristics of Probiotic Supplement and/or Probiotic Formula Users and Nonusers

| Table Variable | Variable |
|---|---|
| Users vs. non-users | prob |
| Maternal age, y | magec |
| Maternal educational level of high school or more | mom_education |
| Birth order, first child | mom_first_child |
| Antibiotics use during pregnancy | antibiotic_use |
| Probiotics use during pregnancy | mom_prob |
| Smoking during pregnancy | smoker |
| Cesarean delivery | csection |
| Birth year | byearc |
| First degree relative with T1DM | fdr |
| HLA genotype DR3/4 | dr34 |
| Female sex | female |
| Exclusive breastfeeding at least 3 mo | excl3mon |
| Child's antibiotics use during the first 12 mo | child_antibiot_use |
| Diarrhea episode during the first 3 mo | child_diarrhea |
| Common cold during the first 3 mo | cold_flag |
| Gastroenteritis (infectious or noninfectious) during the first 12 mo | gast_flag |

**Table D:** Comparison of values computed in integrity check to reference article Table 2 values

| Characteristic | Users Manuscript (n=1637) | Users DSIC (n=1637) | Diff. (n=0) | Nonusers Manuscript (n=5836) | Nonusers DSIC (n=5836) | Diff. (n=0) |
|---|---|---|---|---|---|---|
| Maternal age, y | | | | | | |
| ≤24 | 151 (9.2) | 151 (9.2) | 0 (0) | 746 (12.8) | 746 (12.8) | 0 (0) |
| 25-29 | 520 (31.8) | 520 (31.8) | 0 (0) | 1664 (28.5) | 1664 (28.5) | 0 (0) |
| 30-34 | 598 (36.5) | 598 (36.5) | 0 (0) | 2062 (35.3) | 2062 (35.3) | 0 (0) |
| ≥35 | 368 (22.5) | 368 (22.5) | 0 (0) | 1364 (23.4) | 1364 (23.4) | 0 (0) |
| Maternal educational level of high school or more | 1407 (87.7) | 1407 (87.7) | 0 (0) | 4534 (79.7) | 4534 (79.7) | 0 (0) |
| Birth order, first child | 799 (50.2) | 799 (50.2) | 0 (0) | 2442 (43.1) | 2442 (43.1) | 0 (0) |
| Antibiotics use during pregnancy | 384 (23.7) | 384 (23.7) | 0 (0) | 1321 (22.9) | 1321 (22.9) | 0 (0) |
| Probiotics use during pregnancy | 116 (7.1) | 116 (7.1) | 0 (0) | 169 (2.9) | 169 (2.9) | 0 (0) |
| Smoking during pregnancy | 180 (11.1) | 180 (11.1) | 0 (0) | 704 (12.2) | 704 (12.2) | 0 (0) |
| Cesarean delivery | 393 (24.0) | 393 (24.0) | 0 (0) | 1542 (26.4) | 1542 (26.4) | 0 (0) |
| Birth year | | | | | | |
| 2004-2005 | 163 (10.0) | 163 (10.0) | 0 (0) | 1067 (18.3) | 1067 (18.3) | 0 (0) |
| 2006 | 228 (13.9) | 228 (13.9) | 0 (0) | 1084 (18.6) | 1084 (18.6) | 0 (0) |
| 2007 | 344 (21.0) | 344 (21.0) | 0 (0) | 1225 (21.0) | 1225 (21.0) | 0 (0) |
| 2008 | 379 (23.1) | 379 (23.2) | 0 (0.1) | 1149 (19.7) | 1149 (19.7) | 0 (0) |
| 2009-2010 | 523 (32.0) | 523 (32.0) | 0 (0) | 1311 (22.4) | 1311 (22.5) | 0 (0.1) |
| First degree relative with T1DM | 203 (12.4) | 203 (12.4) | 0 (0) | 639 (11.0) | 639 (11.0) | 0 (0) |
| HLA genotype DR3/4 | 628 (38.4) | 628 (38.4) | 0 (0) | 2305 (39.5) | 2305 (39.5) | 0 (0) |
| Female sex | 797 (48.7) | 797 (48.7) | 0 (0) | 2862 (49.0) | 2862 (49.0) | 0 (0) |
| Exclusive breastfeeding at least 3 mo | 392 (24.0) | 392 (24.0) | 0 (0) | 1456 (25.0) | 1456 (25.0) | 0 (0) |
| Child's antibiotics use during the first 12 mo | 910 (55.6) | 910 (55.6) | 0 (0) | 2435 (41.7) | 2435 (41.7) | 0 (0) |
| Diarrhea episode during the first 3 mo | 169 (10.3) | 169 (10.3) | 0 (0) | 494 (8.5) | 494 (8.5) | 0 (0) |
| Common cold during the first 3 mo | 910 (55.6) | 910 (55.6) | 0 (0) | 3432 (58.8) | 3432 (58.8) | 0 (0) |
| Gastroenteritis (infectious or noninfectious) during the first 12 mo | 582 (35.6) | 582 (35.6) | 0 (0) | 1723 (29.5) | 1723 (29.5) | 0 (0) |

**Table E:** Variables used to replicate Table 3: First Probiotic Exposure of the Child vis Infant Formula and/or Dietary Supplement During the First Year of Life and Risk of IA

| Table Variable | Variable |
|---|---|
| IA | persist_conf_ab |
| Country | country |
| Timing of first probiotic exposure, d | early_prob_ter |
| FDR with T1DM | fdr |
| High-risk HLA-DR-DR3/4 | dr34 |
| Female sex | female |

**Table F:** Comparison of values computed in integrity check to reference article Table 3 values

| Variable | No. (%) of Infants Developed IA Manuscript (n=601) | No. (%) of Infants Developed IA DSIC (n=601) | Diff. (n=0) | No. (%) of Infants Did Not Develop IA Manuscript (n=6872) | No. (%) of Infants Did Not Develop IA DSIC (n=6872) | Diff. (n=0) |
|---|---|---|---|---|---|---|
| Country | | | | | | |
| United States | 201 (33.4) | 201 (33.4) | 0 (0) | 2845 (41.4) | 2845 (41.4) | 0 (0) |
| Finland | 151 (25.1) | 151 (25.1) | 0 (0) | 1507 (21.9) | 1507 (21.9) | 0 (0) |
| Germany | 46 (7.7) | 46 (7.7) | 0 (0) | 460 (6.7) | 460 (6.7) | 0 (0) |
| Sweden | 203 (33.8) | 203 (33.8) | 0 (0) | 2060 (30.0) | 2060 (30.0) | 0 (0) |
| Timing of first probiotic exposure, d | | | | | | |
| 0-27 | 34 (5.7) | 34 (5.7) | 0 (0) | 506 (7.4) | 508 (7.4) | 2 (0) |
| 28-90 | 41 (6.8) | 42 (7.0) | 1 (0.2) | 515 (7.5) | 515 (7.5) | 0 (0) |
| 91-365 | 57 (9.5) | 57 (9.5) | 0 (0) | 481 (7.0) | 481 (7.0) | 0 (0) |
| After 1 year or no exposure | 469 (78.0) | 468 (77.9) | 1 (0.1) | 5370 (78.1) | 5368 (78.1) | 2 (0) |
| FDR with T1DM | 126 (21.0) | 126 (21.0) | 0 (0) | 716 (10.4) | 716 (10.4) | 0 (0) |
| High-risk HLA-DR-DR3/4 | 304 (50.6) | 304 (50.6) | 0 (0) | 2629 (38.3) | 2629 (38.3) | 0 (0) |
| Female sex | 262 (43.6) | 262 (43.6) | 0 (0) | 3397 (49.4) | 3397 (49.4) | 0 (0) |

## Attachment A: SAS Code

```
*** TEDDY M95 Integrity Check;
*** Date: 9/29/16;
*** Programmer: Allyson Mateja;

libname sas_data '/prj/niddk/ims_analysis/TEDDY/private_orig_data/M_95_UUusitalo_NIDDK_Submission/';

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

        value timingf 1 = '0-27'
                      2 = '28-90'
                      3 = '91-365'
                      4 = 'After 1 year or no exposure';

        value agef 1 = '<= 24'
                   2 = '25-29'
                   3 = '30-34'
                   4 = '>= 35';

        value yearf 1 = '2004-2005'
                    2 = '2006'
                    3 = '2007'
                    4 = '2008'
                    5 = '2009-2010';

data m95data;
        set sas_data.m_95_uuusitalo_niddk_31oct2014;

proc contents data = m95data;

data m95data;
        set m95data;
        if exclude = 0;

proc freq data = m95data;
        tables country;
        format country countryf.;
        title 'Table 1 - Country';

proc sort data = m95data;
        by country;

proc freq data = m95data;
        tables prob /list missing;
```

```
        by country;
        format country countryf.;
        title 'Table 1 - Probiotic Use During first 12 mon';

proc freq data = m95data;
        tables prob /list missing;

proc freq data = m95data;
        tables early_prob_3mon /list missing;
        by country;
        format country countryf.;
        title 'Table 1 - Probiotic Use During first 3 mon';

proc freq data = m95data;
        tables early_prob_3mon /list missing;

proc means data = m95data n median p25 p75;
        var prob_startdy;
        where prob=1;
        class country;
        types () country;
        format country countryf.;
        title 'Table 1 - Age at first exposure to probiotics';

proc freq data = m95data;
        tables probsupp*probform /list missing;
        where prob = 1;
        by country;
        format country countryf.;
        title 'Table 1 - Source of first probiotic exposure among probiotic users';

proc freq data = m95data;
        tables probsupp*probform /list missing;
        where prob=1;

proc freq data = m95data;
        tables early_prob_ter /list missing;
        where prob = 1;
        by country;
        format country countryf. early_prob_ter timingf.;
        title 'Table 1 - Timing of first probiotic exposure for users only';

proc freq data = m95data;
        tables early_prob_ter /list missing;
        where prob=1;
        format early_prob_ter timingf.;

proc freq data = m95data;
        tables prob;
        title 'Table 2 - Users vs. Nonusers';
```

```
proc sort data = m95data;
      by prob;

proc freq data = m95data;
      tables magec;
      by prob;
      format magec agef.;
      title 'Table 2 - Maternal Age';

proc freq data = m95data;
      tables mom_education;
      by prob;
      title 'Table 2 - Maternal educational level of high school or more';

proc freq data = m95data;
      tables mom_first_child;
      by prob;
      title 'Table 2 - Birth order, first child';

proc freq data = m95data;
      tables antibiotic_use;
      by prob;
      title 'Table 2 - Antibiotics use during pregnancy';

proc freq data = m95data;
      tables mom_prob;
      by prob;
      title 'Table 2 - Probiotics use during pregnancy';

proc freq data = m95data;
      tables smoker;
      by prob;
      title 'Table 2 - Smoking during pregnancy';

proc freq data = m95data;
      tables csection;
      by prob;
      title 'Table 2 - Cesarean delivery';

proc freq data = m95data;
      tables byearc;
      format byearc yearf.;
      by prob;
      title 'Table 2 - Birth year';

proc freq data = m95data;
      tables fdr;
      by prob;
      title 'Table 2 - First degree relative with T1DM';

proc freq data = m95data;
```

```
        tables dr34;
        by prob;
        title 'Table 2 - HLA genotype DR3/4';

proc freq data = m95data;
        tables female;
        by prob;
        title 'Table 2 - Female sex';

proc freq data = m95data;
        tables excl3mon;
        by prob;
        title 'Table 2 - Exclusive breastfeeding at least 3 mo';

proc freq data = m95data;
        tables child_antibiot_use;
        by prob;
        title 'Table 2 - Child antibiotics use during the first 12 mo';

proc freq data = m95data;
        tables child_diarrhea;
        by prob;
        title 'Table 2 - Diarrhea episode during the first 3 mo';

proc freq data = m95data;
        tables cold_flag;
        by prob;
        title 'Table 2 - Common cold during the first 3 mo';

proc freq data = m95data;
        tables gast_flag;
        by prob;
        title 'Table 2 - Gastroenteritis during the first 12 mo';

proc freq data = m95data;
        tables persist_conf_ab;
        title 'Table 3 - Developed IA vs. Did Not Develop IA';

proc sort data = m95data;
        by persist_conf_ab;

proc freq data = m95data;
        tables country;
        by persist_conf_ab;
        format country countryf.;
        title 'Table 3 - Country';

proc freq data = m95data;
        tables early_prob_ter;
        by persist_conf_ab;
        format early_prob_ter timingf.;
```

```
        title3 'Table 3 - Timing of first probiotic exposure';

proc freq data = m95data;
        tables fdr;
        by persist_conf_ab;
        title3 'Table 3 - First degree relative with T1DM';

proc freq data = m95data;
        tables dr34;
        by persist_conf_ab;
        title3 'Table 3 - High-risk HLA-DR-DR3/4';

proc freq data = m95data;
        tables female;
        by persist_conf_ab;
        title3 'Table 3 - Female sex';
```