Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) Pub101 Uusitalo

Prepared by Allyson Mateja
IMS Inc.
3901 Calverton Blvd, Suite 200 Calverton, MD 20705
October 11, 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_101_uusitalo_niddk_31jan2015.sas7bdat" dataset.

4 Statistical Methods

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

5 Results

For Table 1 in the publication [1], Characteristics of children participating in TEDDY, 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 Figure 2 in the publication [1], Proportion of mothers consuming gluten-containing foods according to tertiles in TEDDY overall and by country, Table C lists the variables that were used in the replication and Figure A compares the results calculated from the archived data file to the results published in Figure 2. The results of the replication are an exact match to published results.

For Figure 3 in the publication [1], Proportion of daily maternal intake of specified gluten-containing foods in TEDDY overall and by country, Table D lists the variables that were used in the replication and Figure B compares the results calculated from the archived data file to the results published in Figure 3. The results of the replication are almost an exact match to published results.

For Table 2 in the publication [1], Maternal sociodemographic and lifestyle characteristics of mothers to children participating in TEDDY, 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 2. The results of the replication are almost an exact match to the published results, with only a few minor discrepancies.

For Table 4 in the publication [1], HRs and 95% CIs of celiac disease autoimmunity and celiac disease adjusted for child's sex, family history of celiac disease, HLA genotype, country, and maternal education: Cox proportional hazard regression, Table G lists the variables that were used in the replication and Table H compares the results calculated from the archived data file to the results published in Table 4. The results of the replication are almost an exact match to the published results.

6 Conclusions

The NIDDK repository is confident that the TEDDY M101 data files to be distributed are a true copy of the study data.

7 References

[1] Uusitalo, U., Lee, H., Aronsson, C.A., Yang, J., Virtanen, S.M., Norris, J., Agardh, D., and the TEDDY study group. "Gluten consumption during late pregnancy and risk of celiac disease in the offspring: the TEDDY birth cohort". American Journal of Clinical Nutrition (2015) 102:1216-1221.

Table A: Variables used to replicate Table 1: Characteristics of children participating in TEDDY

Table Variable	**Variable**
Birth year	cdobyear
Male sex	female
FDR with celiac disease	celiac_fdr
HLA genotype	gehla
Country	country

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

Child characteristic	n (%) Manuscript	n (%) DSIC	Difference
Total	6546	6546	0
Birth year			
2004-2005	1088 (17)	1088 (17)	0 (0)
2006	1135 (17)	1135 (17)	0 (0)
2007	1396 (21)	1396 (21)	0 (0)
2008	1332 (20)	1332 (20)	0 (0)
2009-2010	1595 (25)	1595 (24)	0 (1)
Male sex	3352 (51)	3352 (51)	0 (0)
FDR with celiac disease	164 (3)	164 (3)	0 (0)
HLA genotype			
DR3/3	1339 (21)	1339 (20)	0 (1)
DR3/4	2578 (39)	2578 (39)	0 (0)
Others	2629 (40)	2629 (40)	0 (0)
Country			
United States	2610 (40)	2610 (40)	0 (0)
Finland	1488 (23)	1488 (23)	0 (0)
Germany	399 (6)	399 (6)	0 (0)
Sweden	2049 (31)	2049 (31)	0 (0)

Table C: Variables used to replicate Figure 2: Proportion of mothers consuming gluten-containing foods according to tertiles in TEDDY overall and by country

Chart Variable	Variable
Country	country
Tertile	mom_gluten3

Figure A: Comparison of values computed in integrity check to reference article Figure 2 values

Manuscript

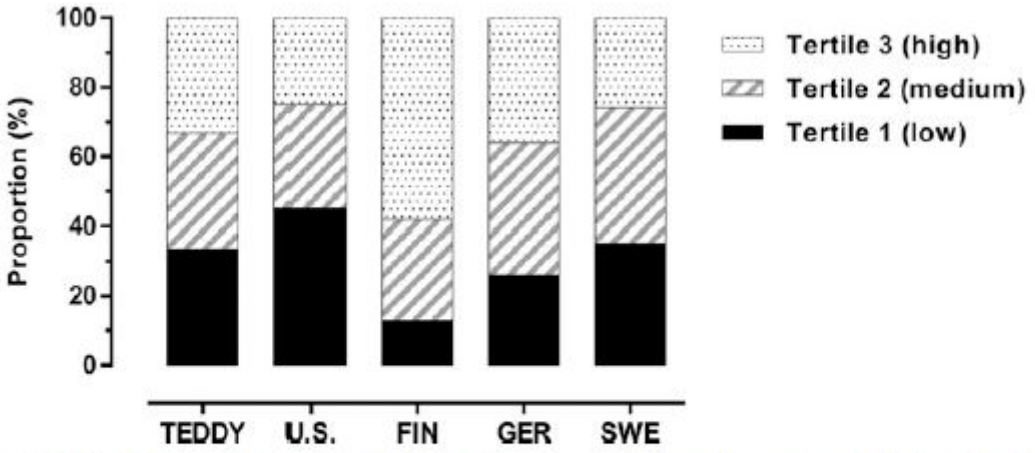
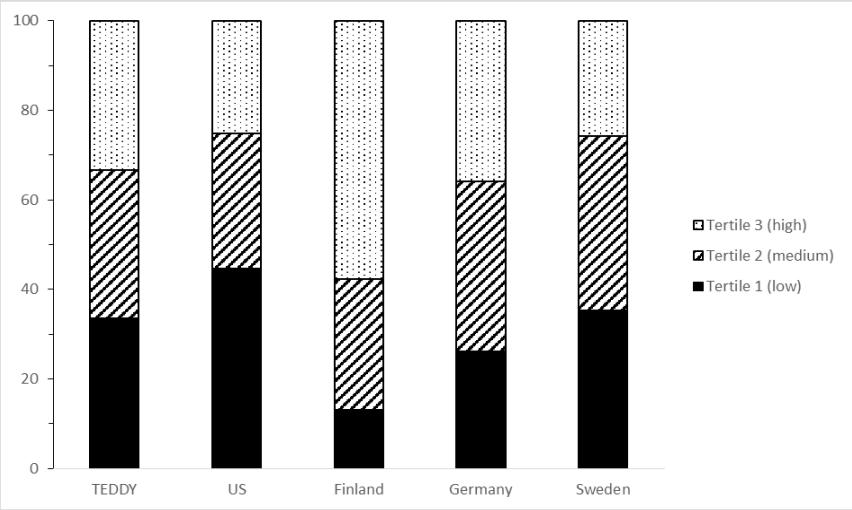


FIGURE 2 Proportion of mothers consuming gluten-containing foods according to tertiles in TEDDY overall and by country: United States (n = 2610), Finland (n = 1488), Germany (n = 399), and Sweden (n = 2049). TEDDY, The Environmental Determinants of Diabetes in the Young.

DSIC



Proportion of mothers consuming gluten-containing food according to tertiles in TEDDY overall and by country: United States (n = 2610), Finland (n = 1488), Germany (n = 399), and Sweden (n = 2049). TEDDY, The Environmental Determinants of Diabetes in the Young

Table D: Variables used to replicate Figure 3: Proportion of daily maternal intake of specified gluten-containing foods in TEDDY overall and by country

Chart Variable	**Variable**
Country	country
Total gluten	mom_gluten
Pasta	pasta_dy
Cereals	cereals_dy
Pizza	pizza_dy
Pastries	pastries_dy
Savory Pastries	meatpastries_dy
Cookies	cookies_dy
Bread	bread_dy

Figure B: Comparison of values computed in integrity check to reference article Figure 3 values

Manuscript

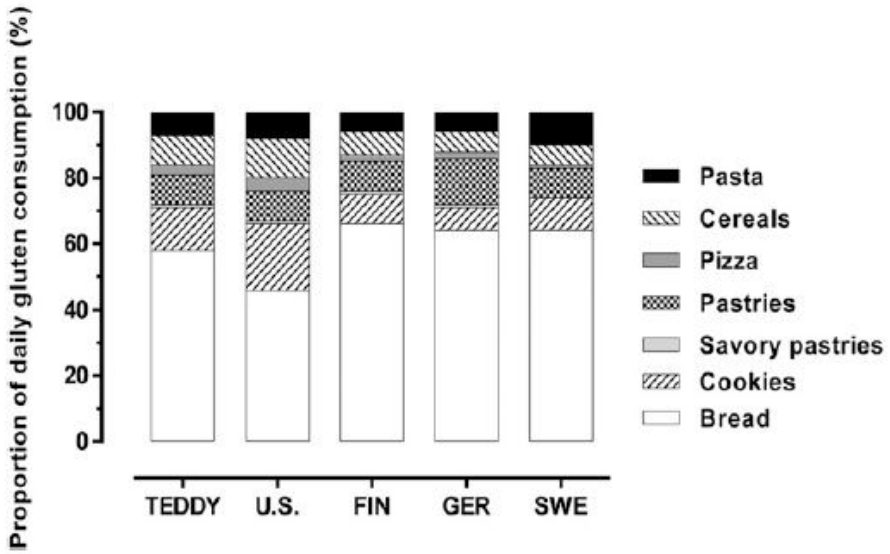
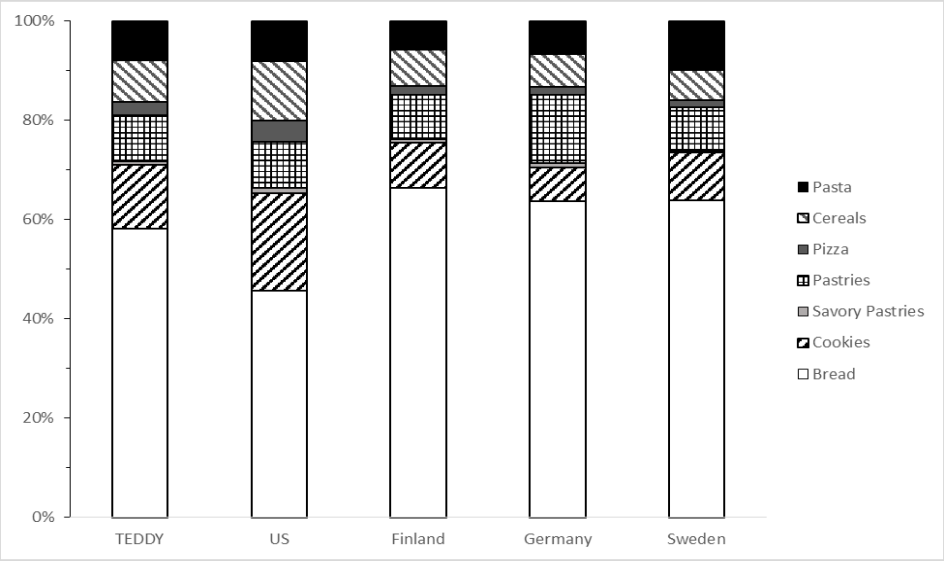


FIGURE 3 Proportion of daily maternal intake of specified gluten-containing foods in TEDDY overall and by country: United States (n = 2610), Finland (n = 1488), Germany (n = 399), and Sweden (n = 2049). TEDDY, The Environmental Determinants of Diabetes in the Young.

DSIC



Proportion of daily maternal intake of specified gluten-containing foods in TEDDY overall and by country: United States (n = 2610), Finland (n=1488), Germany (n = 399), and Sweden (n = 2049). TEDDY, The Environmental Determinants of Diabetes in the Young

Table E: Variables used to replicate Table 2: Maternal sociodemographic and lifestyle characteristics of mothers to children participating in TEDDY

Table Variable	**Variable**
Age at delivery	maternal_age
Gluten consumption	mom_gluten
Education	mom_education
Smoking during pregnancy	rsmoker
Alcohol consumption during pregnancy	drinker
Country	country

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

Variable	TEDDY Manuscript (n=6546)	TEDDY DSIC (n=6546)	Diff. (n=0)	United States Manuscript (n=2610)	United States DSIC (n=2610)	Diff. (n=0)
Age at delivery, y	31 (27-34)	31 (27-34)	0 (0-0)	31 (27-35)	31 (27-35)	0 (0-0)
Gluten consumption, servings/d	3.9 (2.7-5.5)	3.9 (2.7-5.5)	0 (0-0)	3.4 (2.3-4.9)	3.4 (2.3-4.9)	0 (0-0)
Education, n (%)						
>High school	5284 (81)	5284 (81)	0 (0)	2235 (86)	2235 (86)	0 (0)
≤High school	1134 (17)	1134 (17)	0 (0)	336 (13)	336 (13)	0 (0)
Missing	128 (2)	128 (2)	0 (0)	39 (1)	39 (1)	0 (0)
Smoking during pregnancy, n (%)						
Yes	712 (11)	712 (11)	0 (0)	215 (8)	215 (8)	0 (0)
No	5828 (89)	5828 (89)	0 (0)	2389 (92)	2389 (92)	0 (0)
Missing	6 (<1)	6 (<1)	0 (0)	6 (<1)	6 (<1)	0 (0)
Alcohol consumption during pregnancy, n (%)						
Yes	2290 (35)	2290 (35)	0 (0)	1046 (40)	1046 (40)	0 (0)
No	4253 (65)	4253 (65)	0 (0)	1561 (60)	1561 (60)	0 (0)
Missing	3 (<1)	3 (<1)	0 (0)	3 (<1)	3 (<1)	0 (0)

Variable	Finland Manuscript (n=1488)	Finland DSIC (n=1488)	Diff. (n=0)	Germany Manuscript (n=399)	Germany DSIC (n=399)	Diff. (n=0)
Age at delivery, y	30 (27-33)	30 (27-33)	0 (0-0)	32 (29-35)	32 (29-35)	0 (0-0)
Gluten consumption, servings/d	5.3 (3.9-6.9)	5.3 (3.9-6.9)	0 (0-0)	4.3 (3.1-5.5)	4.3 (3.1-5.5)	0 (0-0)
Education, n (%)						
>High school	1350 (91)	1350 (91)	0 (0)	342 (86)	342 (86)	0 (0)
≤High school	99 (7)	99 (7)	0 (0)	37 (9)	37 (9)	0 (0)
Missing	39 (2)	39 (3)	0 (1)	20 (5)	20 (5)	0 (0)
Smoking during pregnancy, n (%)						
Yes	192 (13)	192 (13)	0 (0)	63 (16)	63 (16)	0 (0)
No	1296 (87)	1296 (87)	0 (0)	336 (84)	336 (84)	0 (0)
Missing	0	0	0	0	0	0
Alcohol consumption during pregnancy, n (%)						
Yes	457 (31)	457 (31)	0 (0)	203 (51)	203 (51)	0 (0)
No	1031 (69)	1031 (69)	0 (0)	196 (49)	296 (49)	0 (0)
Missing	0	0	0	0	0	0

Variable	Sweden Manuscript (n=2049)	Sweden DSIC (n=2049)	Diff. (n=0)
Age at delivery, y	31 (28-34)	31 (28-34)	0 (0-0)
Gluten consumption, servings/d	3.7 (2.8-4.9)	3.7 (2.7-4.9)	0 (0.1-0)
Education, n (%)			
>High school	1357 (66)	1357 (66)	0 (0)
≤High school	662 (32)	662 (32)	0 (0)
Missing	30 (2)	30 (1)	0 (1)
Smoking during pregnancy, n (%)			
Yes	242 (12)	242 (12)	0 (0)
No	1807 (88)	1807 (88)	0 (0)

Variable	Sweden Manuscript (n=2049)	Sweden DSIC (n=2049)	Diff. (n=0)
Missing	0	0	0
Alcohol consumption during pregnancy, n (%)			
Yes	584 (29)	584 (29)	0 (0)
No	1465 (71)	1465 (72)	0 (1)
Missing	0	0	0

Table G: Variables used to replicate Table 4: HRs and 95% CIs of celiac disease autoimmunity and celiac disease adjusted for child's sex, family history of celiac disease, HLA genotype, country, and maternal education: Cox proportional hazard regression

Table Variable	**Variable**
Celiac disease autoimmunity	htg_conf
Celiac disease	celiac_disease
Sex	female
Family history of CD	celiac_fdr
HLA Genotype	gehla
Country	country
Maternal education	mom_education
Maternal gluten consumption during pregnancy	mom_gluten3

Table H: Comparison of values computed in integrity check to reference article Table 4 values

Risk Factor	Celiac disease autoimmunity Manuscript n (%)	Celiac disease autoimmunity DSIC n (%)	Diff.	Celiac disease Manuscript n (%)	Celiac disease DSIC n (%)	Diff.
Sex						
Male	435 (13)	435 (13)	0 (0)	133 (4)	133 (4)	0 (0)
Female	580 (18)	580 (18)	0 (0)	226 (7)	226 (7)	0 (0)
Family history of CD						
No	946 (15)	946 (15)	0 (0)	315 (5)	315 (5)	0 (0)
Yes	69 (42)	69 (42)	0 (0)	44 (27)	44 (27)	0 (0)
HLA genotype						
Other	194 (7)	194 (7)	0 (0)	58 (2)	58 (2)	0 (0)
DR3/3	435 (32)	435 (32)	0 (0)	184 (14)	184 (14)	0 (0)
DR3/4	386 (15)	386 (15)	0 (0)	117 (5)	117 (5)	0 (0)
Country						
United States	348 (13)	348 (13)	0 (0)	105 (4)	105 (4)	0 (0)
Finland	207 (14)	207 (14)	0 (0)	58 (4)	58 (4)	0 (0)
Germany	56 (14)	56 (14)	0 (0)	14 (4)	14 (4)	0 (0)
Sweden	404 (20)	404 (20)	0 (0)	182 (9)	182 (9)	0 (0)
Maternal education						
≤High school	151 (13)	151 (13)	0 (0)	29 (5)*	59 (5)	30 (0)
>High school	851 (16)	851 (16)	0 (0)	297 (6)	297 (6)	0 (0)
Maternal gluten consumption during pregnancy						
Low	331 (15)	331 (15)	0 (0)	119 (5)	119 (5)	0 (0)
Middle	343 (16)	343 (16)	0 (0)	121 (6)	121 (6)	0 (0)
High	341 (16)	341 (16)	0 (0)	119 (5)	119 (5)	0 (0)

*This number from the publication is a typo and the calculated value in the DSIC of 59 (5) is correct.

Attachment A: SAS Code

```
*** TEDDY M101 Analysis DSIC;
*** Programmer: Allyson Mateja;
*** Date: 10/4/16;

proc format;
       value byearf 1 = '2004-2005'
                    2 = '2006'
                    3 = '2007'
                    4 = '2008'
                    5 = '2009-2010';
       value sexf 0 = 'M'
                  1 = 'F';
       value hlaf 1 = 'DR3/3'
                  2 = 'DR3/4'
                  3 = 'Others';
       value countryf 1 = 'United States'
                      2 = 'Finland'
                      3 = 'Germany'
                      4 = 'Sweden';
       value tertilef 0 = 'Tertile 1 (low)'
                      1 = 'Tertile 2 (medium)'
                      2 = 'Tertile 3 (high)';

libname sas_data '/prj/niddk/ims_analysis/TEDDY/private_orig_data/M_101_UUusitalo_NIDDK_Submission/';

data m101data;
       set sas_data.m_101_uuusitalo_niddk_31jan2015;

proc contents data=m101data;

proc freq data=m101data;
       tables cdobyear;
       format cdobyear byearf.;
       title 'Table 1 - Birth year';

proc freq data = m101data;
       tables female;
       format female sexf.;
       title 'Table 1 - Gender';

proc freq data = m101data;
       tables celiac_fdr;
       title 'Table 1 - FDR with celiac disease';

proc freq data = m101data;
       tables gehla;
       format gehla hlaf.;
```

```
        title 'Table 1 - HLA genotype';

proc freq data = m101data;
        tables country;
        format country countryf.;
        title 'Table 1 - Country';

proc sort data = m101data;
        by country;

proc freq data = m101data;
        tables mom_gluten3;
        by country;
        format country countryf. mom_gluten3 tertilef.;
        title 'Figure 2';

proc freq data = m101data;
        tables mom_gluten3;
        format mom_gluten3 tertilef.;

proc means data = m101data n sum;
        var mom_gluten pasta_dy cereals_dy pizza_dy pastries_dy meatpastries_dy cookies_dy bread_dy;
        class country;
        types () country;
        title 'Figure 3';

proc means data = m101data n median p25 p75;
        var maternal_age;
        class country;
        types () country;
        title 'Table 2 - Age at delivery';

proc means data = m101data n median p25 p75;
        var mom_gluten;
        class country;
        types () country;
        title 'Table 2 - Gluten consumption, servings/d';

proc freq data = m101data;
        tables mom_education /missing;
        title 'Table 2 - Education';

proc freq data = m101data;
        tables mom_education /missing;
        by country;

proc freq data = m101data;
        tables rsmoker /missing;
        title 'Table 2 - Smoking during pregnancy';

proc freq data = m101data;
```

```
        tables rsmoker /missing;
        by country;

proc freq data = m101data;
        tables drinker /missing;
        title 'Table 2 - Alcohol consumption during pregnancy';

proc freq data = m101data;
        tables drinker /missing;
        by country;

proc freq data = m101data;
        tables htg_conf;
        where female = 0;
        title 'Table 4 - Celiac disease autoimmunity - Male';

proc freq data = m101data;
        tables htg_conf;
        where female = 1;
        title 'Table 4 - Celiac disease autoimmunity - Female';

proc freq data = m101data;
        tables celiac_disease /list missing;
        where female = 0;
        title 'Table 4 - Celiac Disease - Male';

proc freq data = m101data;
        tables celiac_disease /list missing;
        where female = 1;
        title 'Table 4 - Celiac Disease - Female';

proc freq data = m101data;
        tables htg_conf;
        where celiac_fdr = 0;
        title 'Table 4 - Celiac disease autoimmunity - Family history of CD, No';

proc freq data = m101data;
        tables htg_conf;
        where celiac_fdr = 1;
        title 'Table 4 - Celiac disease autoimmunity - Family history of CD, Yes';

proc freq data = m101data;
        tables celiac_disease /list missing;
        where celiac_fdr = 0;
        title 'Table 4 - Celiac Disease - Family history of CD, No';

proc freq data = m101data;
        tables celiac_disease /list missing;
        where celiac_fdr = 1;
        title 'Table 4 - Celiac Disease - Family history of CD, Yes';
```

```
proc freq data = m101data;
       tables htg_conf;
       where gehla = 1;
       title 'Table 4 - Celiac disease autoimmunity - HLA genotype DR3/3';

proc freq data = m101data;
       tables htg_conf;
       where gehla = 2;
       title 'Table 4 - Celiac disease autoimmunity - HLA genotype DR3/4';

proc freq data = m101data;
       tables htg_conf;
       where gehla = 3;
       title 'Table 4 - Celiac disease autoimmunity - HLA genotype Other';

proc freq data = m101data;
       tables celiac_disease /list missing;
       where gehla = 1;
       title 'Table 4 - Celiac Disease - HLA genotype DR3/3';

proc freq data = m101data;
       tables celiac_disease /list missing;
       where gehla = 2;
       title 'Table 4 - Celiac Disease - HLA genotype DR3/4';

proc freq data = m101data;
       tables celiac_disease;
       where gehla = 3;
       title 'Table 4 - Celiac disease autoimmunity - HLA genotype Other';

proc freq data = m101data;
       tables htg_conf;
       where country = 1;
       title 'Table 4 - Celiac disease autoimmunity - US';

proc freq data = m101data;
       tables htg_conf;
       where country = 2;
       title 'Table 4 - Celiac disease autoimmunity - Finland';

proc freq data = m101data;
       tables htg_conf;
       where country = 3;
       title 'Table 4 - Celiac disease autoimmunity - Germany';

proc freq data = m101data;
       tables htg_conf;
       where country = 4;
       title 'Table 4 - Celiac disease autoimmunity - Sweden';

proc freq data = m101data;
```

```
        tables celiac_disease /list missing;
        where country = 1;
        title 'Table 4 - Celiac Disease - US';

proc freq data = m101data;
        tables celiac_disease /list missing;
        where country = 2;
        title 'Table 4 - Celiac Disease - Finland';

proc freq data = m101data;
        tables celiac_disease;
        where country = 3;
        title 'Table 4 - Celiac disease autoimmunity - Germany';

proc freq data = m101data;
        tables celiac_disease;
        where country = 4;
        title 'Table 4 - Celiac disease autoimmunity - Sweden';

proc freq data = m101data;
        tables htg_conf;
        where mom_education = 0;
        title 'Table 4 - Celiac disease autoimmunity - Maternal education <= high school';

proc freq data = m101data;
        tables htg_conf;
        where mom_education = 1;
        title 'Table 4 - Celiac disease autoimmunity - Maternal education > high school';

proc freq data = m101data;
        tables celiac_disease /list missing;
        where mom_education = 0;
        title 'Table 4 - Celiac Disease - Maternal education <= high school';

proc freq data = m101data;
        tables celiac_disease /list missing;
        where mom_education = 1;
        title 'Table 4 - Celiac Disease - Maternal education > high school';

proc freq data = m101data;
        tables htg_conf;
        where mom_gluten3 = 0;
        title 'Table 4 - Celiac disease autoimmunity - Maternal gluten consumption low';

proc freq data = m101data;
        tables htg_conf;
        where mom_gluten3 = 1;
        title 'Table 4 - Celiac disease autoimmunity - Maternal gluten consumption middle';

proc freq data = m101data;
        tables htg_conf;
```

```
        where mom_gluten3 = 2;
        title 'Table 4 - Celiac disease autoimmunity - Maternal gluten consumption high';

proc freq data = m101data;
        tables celiac_disease /list missing;
        where mom_gluten3 = 0;
        title 'Table 4 - Celiac Disease - Maternal gluten consumption low';

proc freq data = m101data;
        tables celiac_disease /list missing;
        where mom_gluten3 = 1;
        title 'Table 4 - Celiac Disease - Maternal gluten consumpetion middle';

proc freq data = m101data;
        tables celiac_disease /list missing;
        where mom_gluten3 = 2;
        title 'Table 4 - Celiac Disease - Maternal gluten consumpetion high';
```