Dataset Integrity Check for The Environmental Determinants of Diabetes in the Young (TEDDY) Pub52 Kemppainen

Prepared by Allyson Mateja
IMS Inc.
3901 Calverton Blvd, Suite 200 Calverton, MD 20705
December 13, 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_52_kkemppainen_31july2011.csv" dataset.

4 Statistical Methods

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

5 Results

For Figure 1A in the publication [1], Table A 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 1A. The results of the replication are an exact match to the published results.

6 Conclusions

The NIDDK repository is confident that the TEDDY M52 data files to be distributed are a true copy of the study data.

7 References

[1] Kemppainen, K.M., Ardissone, A.N., Davis-Richardson, A.G., Fagen, J.R., Gano, K.A., Leon-Novelo, L.G., Vehik, K., Casella, G., Simell, O., Ziegler, A.G., Rewers, M.J., Lernmark, A., Hagopian, W., She, J., Krischer, J.P., Akolkar, B., Schatz, D.A., Atkinson, M.A., Triplett, E.W., and the TEDDY study group. "Early Childhood Gut Microbiomes Show Strong Geographic Differences Among Subjects at High Risk for Type 1 Diabetes". Diabetes Care (2015) 38:329-332.

Table A: Variables used to replicate Figure 1A

Figure Variable	**Variable**
Country, Months after birth	site_month
Roseburia	Roseburia
Eubacterium	Eubacterium
Akkermansia	Akkermansia
Serratia	Serratia
Stretococcus	Stretococcus
Escheria	Escheria
Ruminococcus	Ruminococcus
Faecalibacteriu,	Faecalibacteriu,
Veillonella	Veillonella
Bifidobacterium	Bifidobacterium
Clostridium	Clostridium
Bacteroides	Bacteroides

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

Manuscript:

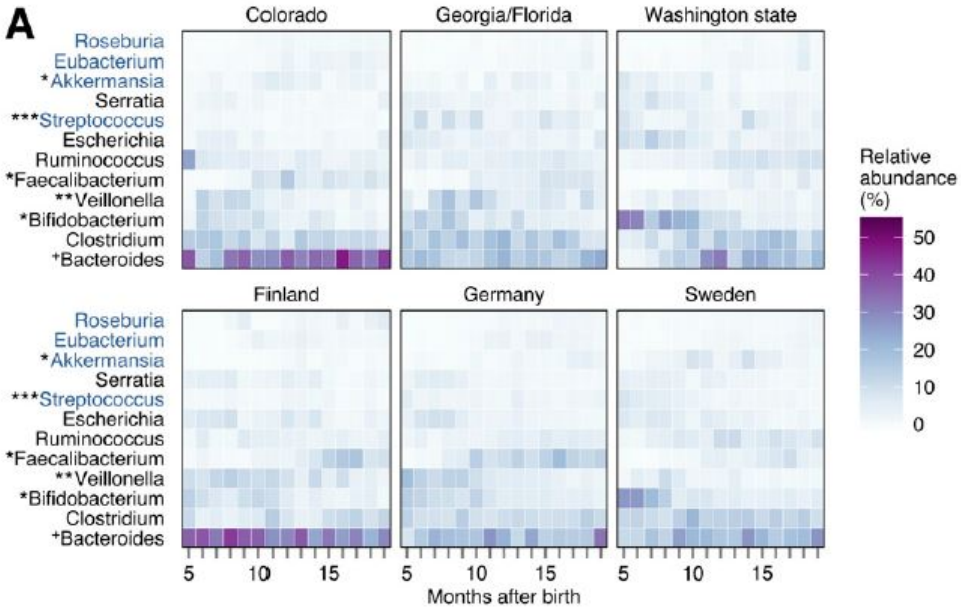

Figure 1—A: A heat map of the relative abundance of the most abundant bacterial genera shows a distinct pattern of development at each study site. 16S rRNA read values were grouped according to age of subject (in months) at the time of sample collection. If a subject had more than one sample within 1 month, the read values were averaged to prevent over-representation of a single individual.

DSIC

A heat map created using values in the archived data provided by the DCC. It was created using conditional formatting in Microsoft Excel.