

# Investigation of Adiposity Measures and Operational Taxonomic unit (OTU) Data Transformation Procedures in Stool Samples from a German Cohort Study Using Machine Learning Algorithms

**Martina Troll** <sup>1,2,\*</sup>, **Stefan Brandmaier** <sup>1,2</sup>, **Sandra Reitmeier** <sup>3,4</sup>, **Jonathan Adam** <sup>1,2</sup>, **Sapna Sharma** <sup>1,2</sup>, **Alice Sommer** <sup>2,5</sup>, **Marie-Abèle Bind** <sup>5</sup>, **Klaus Neuhaus** <sup>3</sup>, **Thomas Clavel** <sup>3,6</sup>, **Jerzy Adamski** <sup>7,8,9</sup>, **Dirk Haller** <sup>3,4</sup>, **Annette Peters** <sup>2,10</sup> and **Harald Grallert** <sup>1,2,11,\*</sup>

<sup>1</sup> Research Unit of Molecular Epidemiology, Helmholtz Zentrum München, 85764 Neuherberg, Germany; stefan.brandmaier@helmholtz-muenchen.de (S.B.); jonathan.adam@helmholtz-muenchen.de (J.A.); sapna.sharma@helmholtz-muenchen.de (S.S.);

<sup>2</sup> Institute of Epidemiology, Helmholtz Zentrum München, 85764 Neuherberg, Germany; ajsommer@fas.harvard.edu (A.S.); peters@helmholtz-muenchen.de (A.P.)

<sup>3</sup> ZIEL Institute for Food & Health, Technical University of Munich, 85354 Freising-Weihenstephan, Germany; sandra.reitmeier@tum.de (S.R.); neuhaus@tum.de (K.N.); tclavel@ukaachen.de (T.C.); dirk.haller@tum.de (D.H.)

<sup>4</sup> Chair of Nutrition and Immunology, Technical University of Munich, 85354 Freising-Weihenstephan, Germany

<sup>5</sup> Department of Statistics, Faculty of Arts and Sciences, Harvard University, Cambridge, MA 02138-2901, USA; ma.bind@mail.harvard.edu

<sup>6</sup> Functional Microbiome Research Group, Institute of Medical Microbiology, RWTH University Hospital, 52074 Aachen, Germany

<sup>7</sup> Research Unit Molecular Endocrinology and Metabolism, Helmholtz Zentrum München, 85764 Neuherberg, Germany; adamski@helmholtz-muenchen.de

<sup>8</sup> Chair of Experimental Genetics, Technical University of Munich, 85350 Freising-Weihenstephan, Germany

<sup>9</sup> Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, 117597 Singapore

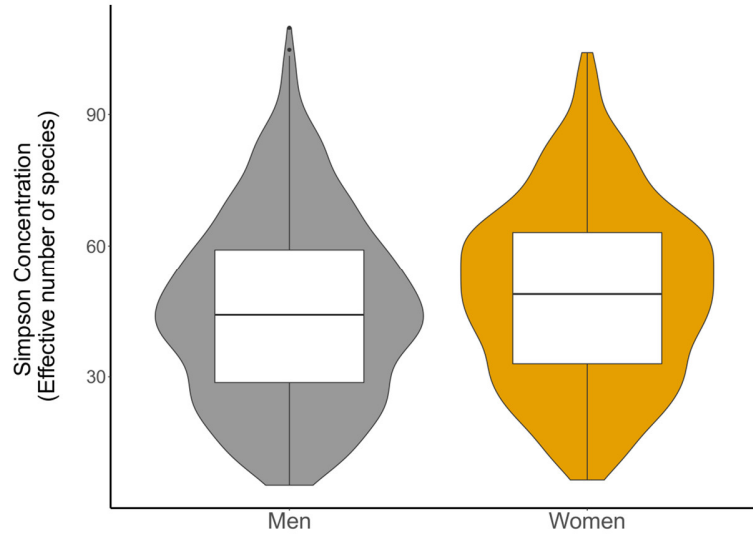
<sup>10</sup> Chair of Epidemiology, Faculty of Medicine, Ludwig-Maximilians-University München, 81377 Munich, Germany

<sup>11</sup> German Center for Diabetes Research (DZD), 85764 Neuherberg, Germany

\* Correspondence: martina.troll@helmholtz-muenchen.de (M.T.); harald.grallert@helmholtz-muenchen.de (H.G.)

Received: 9 March 2020; Accepted: 7 April 2020; Published: 10 April 2020

## Supplementary Materials:



**Figure S1.** Alpha diversity as effective number of species by Simpson index in men and women.

**Table S1:** Extended benchmark data on total study population (n = 1,923).

Characteristics	Mean	SD	Median	Min	Max
Age [years]	60.0	12.1	60.0	38.0	88.0
Body mass index [kg/m <sup>2</sup> ]	27.9	5.0	27.1	18.6	62.2
Waist circumference [cm]	97.0	14.2	97.2	64.7	172.3
Waist-hip ratio	0.91	0.09	0.91	0.65	1.20
Waist-height ratio	0.58	0.08	0.57	0.39	0.95
Body adiposity index	31.0	6.0	29.7	19.2	61.4
Fat mass index [kg/m <sup>2</sup> ]	9.4	3.4	8.8	2.6	29.2
Lean body mass index [kg/m <sup>2</sup> ]	18.5	2.6	18.4	12.8	33.0
Appendicular muscle mass index [kg/m <sup>2</sup> ]	7.7	1.3	7.6	4.8	14.6
Body fat [%]	32.9	7.1	32.8	11.9	50.9

**Table S2a. Comparison of levels of prevalence in waist-height ratio.** In each step, OTUs present in less than x % of the samples were excluded. For each subset separately, CLR relative abundance (RA) and log-transformation with one pseudocount was used as prior data transformation. Support vector machine regression (SVMReg) with normalized poly kernel (NPK) and Partial Least Squares (PLS) with 10-fold cross-validation was used. CC: correlation coefficient; RMSE: root mean squared error.

Excluded percentage missingness in samples	Number of OTUs	SVMReg NPK & CLR		PLS4 & RA+ Log	
		CC	RMSE	CC	RMSE
0 %	2,089 (100 %)	0.37	0.94	0.36	0.94
10 %	958 (46 %)	0.37	0.94	0.36	0.94
20 %	766 (37 %)	0.36	0.94	0.36	0.94
30 %	591 (28 %)	0.34	0.95	0.36	0.94
40 %	451 (22 %)	0.34	0.95	0.35	0.94
50 %	359 (17 %)	0.32	0.96	0.34	0.95
60 %	276 (13 %)	0.30	0.97	0.33	0.95
70 %	200 (10 %)	0.26	0.99	0.30	0.96
80 %	122 (6 %)	0.21	1.01	0.28	0.96
90 %	52 (2 %)	0.19	1.01	0.24	0.97

**Table S2b. Comparison of levels of relative abundance.** In each step, OTUs with an across-sample relative abundance of x % were excluded. For each subset separately, CLR or relative abundance (RA) and log-transformation with one pseudocount was used as prior data transformation. Support vector machine regression (SVMReg) with normalized poly kernel (NPK) and Partial Least Squares (PLS) with 10-fold cross-validation was used. CC: correlation coefficient; RMSE: root mean squared error.

Min. relative abundance in samples	Number of OTUs	SVMReg NPK & CLR		PLS4 & RA+ Log	
		CC	RMSE	CC	RMSE
0.0 %	2,089 (100 %)	0.37	0.94	0.36	0.94
0.001 %	1,917 (92 %)	0.37	0.94	0.36	0.94
0.005 %	1,353 (65 %)	0.37	0.94	0.36	0.94
0.01 %	1,064 (51 %)	0.36	0.94	0.36	0.94
0.05 %	384 (18 %)	0.30	0.97	0.34	0.95
0.1 %	204 (10 %)	0.27	0.98	0.31	0.96
0.5 %	28 (1 %)	0.14	1.02	0.22	0.98

#### Minimal effects of zero replacement via pseudocounts

The handling of zeros is crucial in an analytical approach that uses log transformations. One approach is to add a certain value to all measurements in the data set and therefore offset all zero values, so called pseudocounts. Since there is no common answer to the question of optimal pseudocount value choice [1] we evaluated the performance of support vector machine in waist-height ratio prediction according to varying pseudocount values. Previous studies have used very small values to offset zeros in the data set,

e.g.  $10^{-6}$  [2,3]. In our study, the value of the pseudocount was only minimally relevant to the prediction performance within the same machine learning algorithm (Table S3). However, using the value one leads to the retaining of zeros after the log transformation, therefore preserving the original data structure.

**Table S3. Comparison of pseudocounts.** CLR with varying pseudocounts was used as prior data transformation. Support vector machine regression (SVMReg) with normalized poly kernel (NPK) and 10-fold cross-validation was used. Waist-height ratio was used as dependent variable. CC: correlation coefficient; RMSE: root mean squared error.

Value of pseudocount	SVMReg NPK	
	CC	RMSE
1	0.37	0.94
0.1	0.38	0.93
0.01	0.38	0.93
0.000001	0.38	0.93

## References

1. Weiss, S.; Xu, Z.Z.; Peddada, S.; Amir, A.; Bittinger, K.; Gonzalez, A.; Lozupone, C.; Zaneveld, J.R.; Vázquez-Baeza, Y.; Birmingham, A.; et al. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* **2017**, *5*, 27.
2. Beaumont, M.; Goodrich, J.K.; Jackson, M.A.; Yet, I.; Davenport, E.R.; Vieira-Silva, S.; Debelius, J.; Pallister, T.; Mangino, M.; Raes, J.; et al. Heritable components of the human fecal microbiome are associated with visceral fat. *Genome Biol.* **2016**, *17*, 189.
3. Zierer, J.; Jackson, M.A.; Kastenmüller, G.; Mangino, M.; Long, T.; Telenti, A.; Mohn, R.P.; Small, K.S.; Bell, J.T.; Steves, C.J.; et al. The fecal metabolome as a functional readout of the gut microbiome. *Nat. Genet.* **2018**, *50*, 790–795.