Multi-site Data Harmonization with ComBat

Dana Tudorascu (BDM Core)

Work was done by Chen Luo (PET Center, Pittsburgh)

Objective

- To understand how the ComBat adjust the data
  - on different batch (site).
  - if/not including biological covariates.
- To understand how the the effect of small outlier batch when performing ComBat.

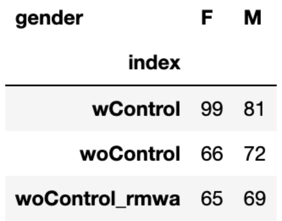
Data

- Data:
  - Mayo_Aparc:
    - unit: mm
    - definition: thickness
  - Mayo_Aseg:
    - unit: mm3
    - definition: volume
  - Mayo_Aseg_Norm:
    - unit: % (mm3/mm3*100%)
    - definition: percentage of volumn based on EstimatedTotalIntraCranialVol
- Sample size:
  - Aparc and Aseg have the same sample size total 198.
  - Because 18 sample has missing the AgeAtScan and diagnosis, only 180 are used,
  - By removing the control, 138 remain.
  - By removing the WashU datapoint, 134 remained.

- Level: dataset
  - wControl:
    - with control datapoint and include WashU's data
  - woCotrol
    - without control datapoint and include WashU's data
  - woCotrol_rmwa
    - without control datapoint and exclude WashU's data

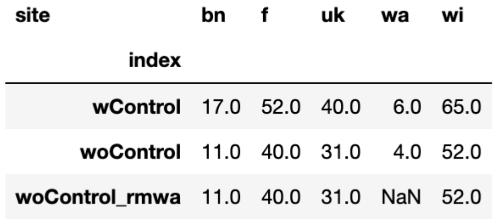
- Level: adjust
  - Original:
    - the raw data
  - ComBat:
    - without considering other biological covariance
  - ComBat_mod:
    - with considering other biological covariance

Demographic information



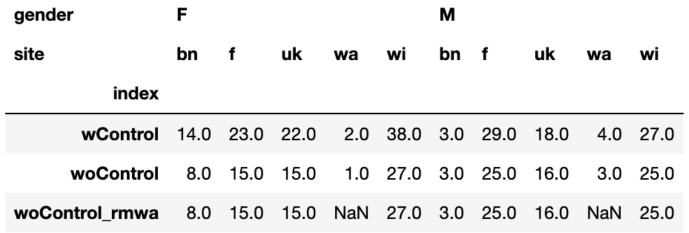
gender	F	M
index		
wControl	99	81
woControl	66	72
woControl_rmwa	65	69

Sample size by gender



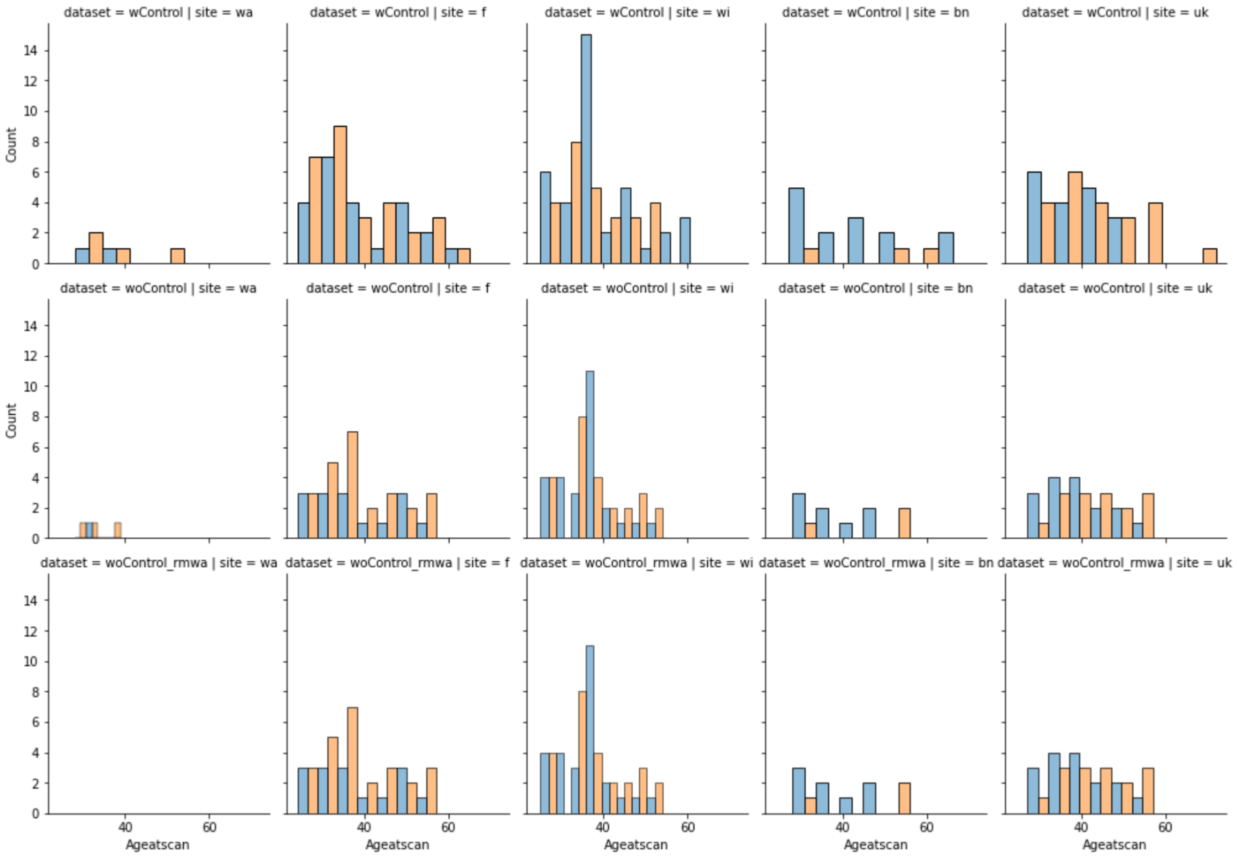
site	bn	f	uk	wa	wi
index					
wControl	17.0	52.0	40.0	6.0	65.0
woControl	11.0	40.0	31.0	4.0	52.0
woControl_rmwa	11.0	40.0	31.0	NaN	52.0

Sample size by site



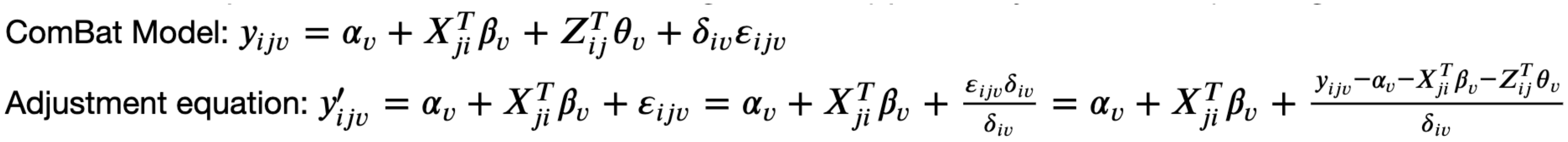
gender	F					M				
site	bn	f	uk	wa	wi	bn	f	uk	wa	wi
index										
wControl	14.0	23.0	22.0	2.0	38.0	3.0	29.0	18.0	4.0	27.0
woControl	8.0	15.0	15.0	1.0	27.0	3.0	25.0	16.0	3.0	25.0
woControl_rmwa	8.0	15.0	15.0	NaN	27.0	3.0	25.0	16.0	NaN	25.0

Sample size by gender and site



Age distribution by site and gender (F: Orange, M: Blue)

Method: ComBat Model



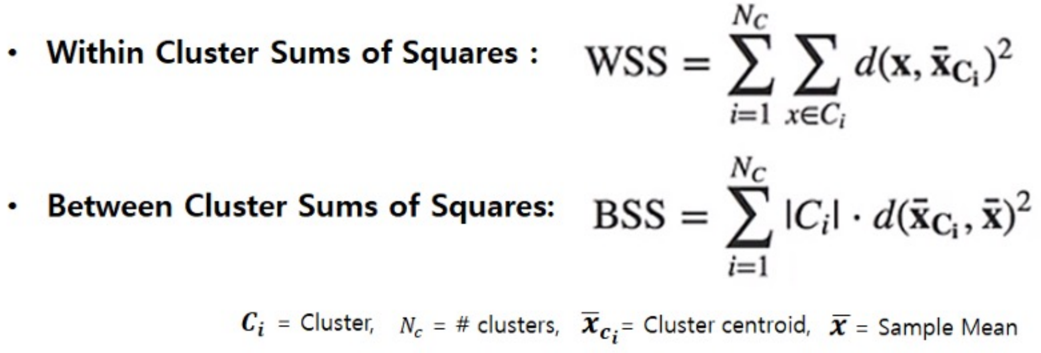
ComBat Model: $y_{ijv} = \alpha_v + X_{ji}^T \beta_v + Z_{ij}^T \theta_v + \delta_{iv} \varepsilon_{ijv}$

Adjustment equation: $y'_{ijv} = \alpha_v + X_{ji}^T \beta_v + \varepsilon_{ijv} = \alpha_v + X_{ji}^T \beta_v + \frac{\varepsilon_{ijv} \delta_{iv}}{\delta_{iv}} = \alpha_v + X_{ji}^T \beta_v + \frac{y_{ijv} - \alpha_v - X_{ji}^T \beta_v - Z_{ij}^T \theta_v}{\delta_{iv}}$

- The ComBat reconstructs the data by separating:
  - mean biological effect,
  - non-biological effect (batch),
  - Other biological effect (mod).
- input:
  - data: VOI measurement
  - mod: with/without age and gender and diagnosis
  - batch effect: site
- setup argument :
  - \`mean.only=FALSE\`: change mean and adjust scale.
  - \`par.prior=FALSE\`: nonparametric estimation

Method: Cluster analysis

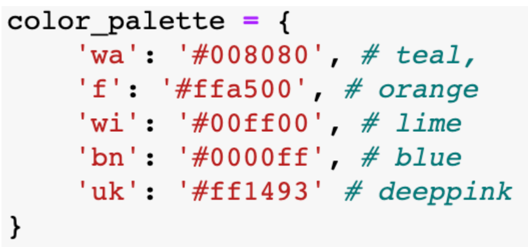
- By considering data from one site as one cluster, evaluate the batch (site) effect using separation metric, which commonly used in clustering analysis.
- The following statistics are evaluated to compare the dataset before and after the ComBat Adjustment



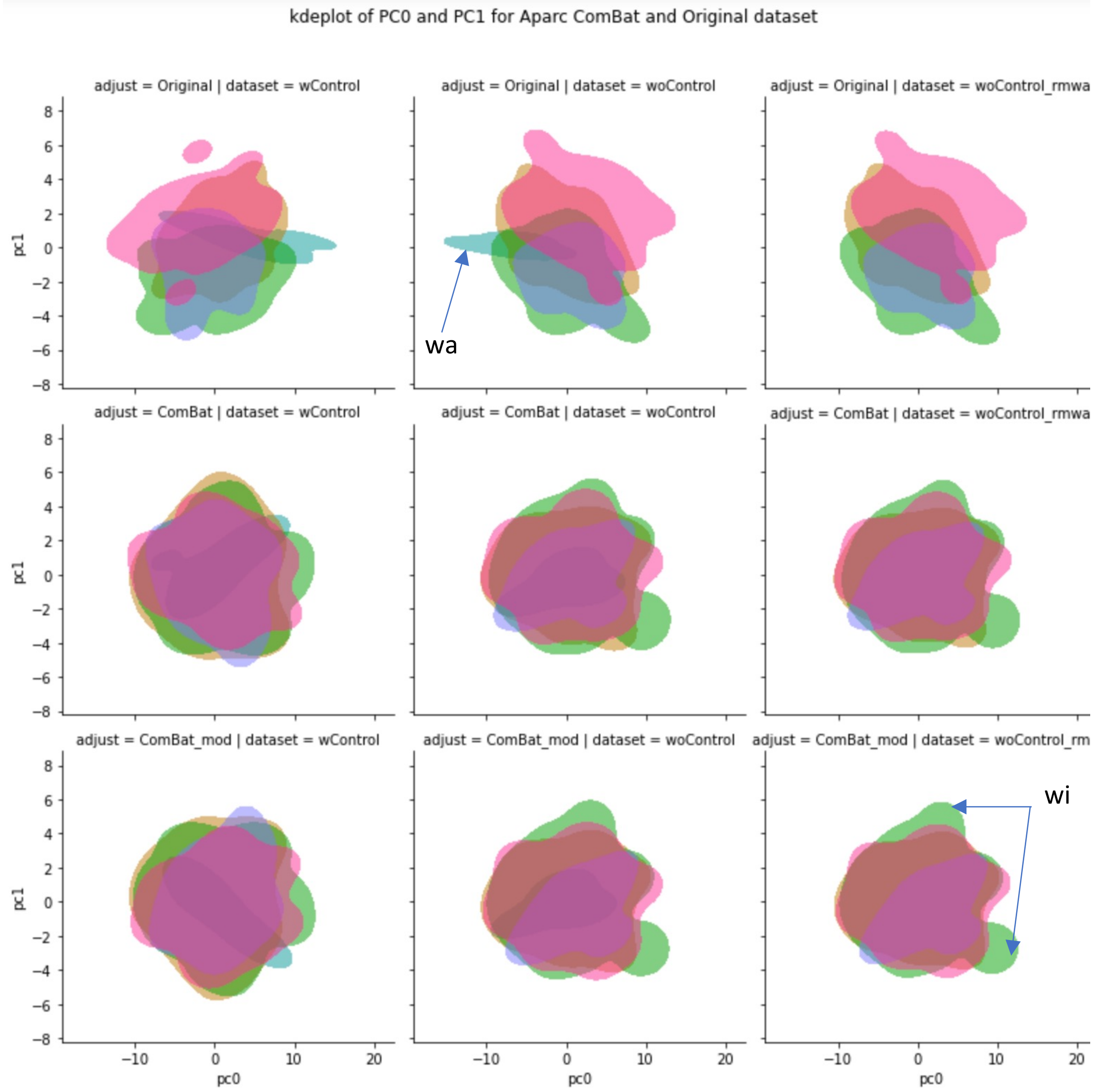
- **Within Cluster Sums of Squares :** $$\mathrm{WSS} = \sum_{i=1}^{N_C} \sum_{x \in C_i} d(\mathbf{x}, \bar{\mathbf{x}}_{\mathrm{C_i}})^2$$
- **Between Cluster Sums of Squares:** $$\mathrm{BSS} = \sum_{i=1}^{N_C} |C_i| \cdot d(\bar{\mathbf{x}}_{\mathrm{C_i}}, \bar{\mathbf{x}})^2$$

$\boldsymbol{C_i}$ = Cluster, N_c = # clusters, $\overline{\boldsymbol{x}}_{\boldsymbol{c_i}}$= Cluster centroid, $\overline{\boldsymbol{x}}$ = Sample Mean

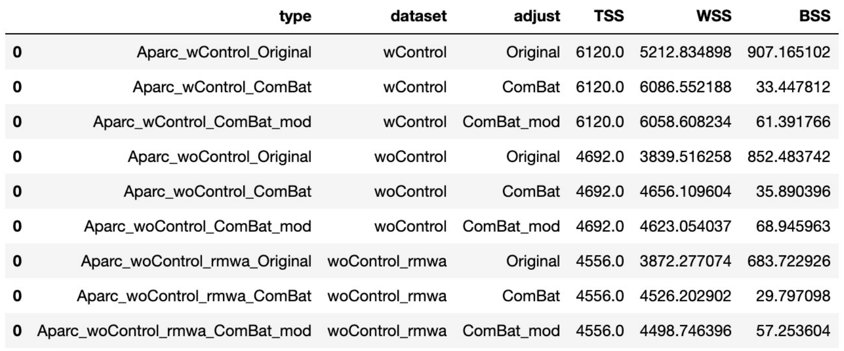
Grand comparison: Aparc



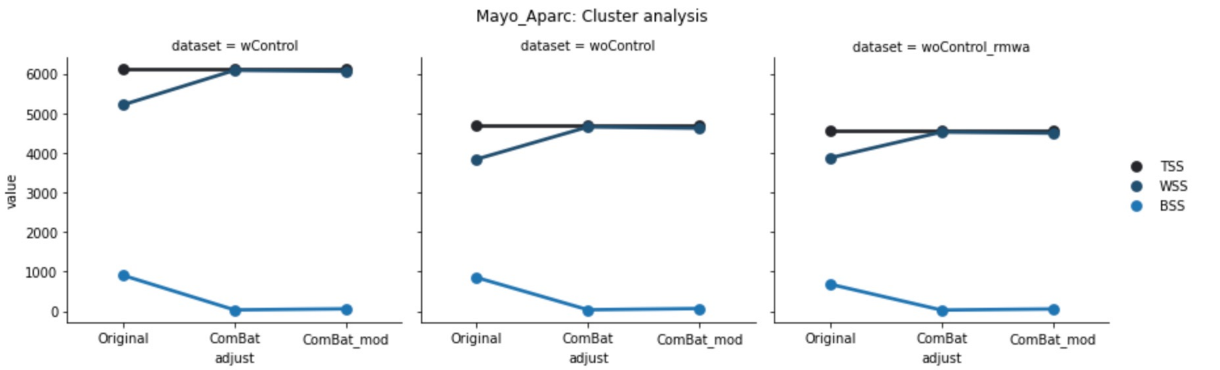
```
color_palette = {
    'wa': '#008080', # teal,
    'f': '#ffa500', # orange
    'wi': '#00ff00', # lime
    'bn': '#0000ff', # blue
    'uk': '#ff1493' # deeppink
}
```



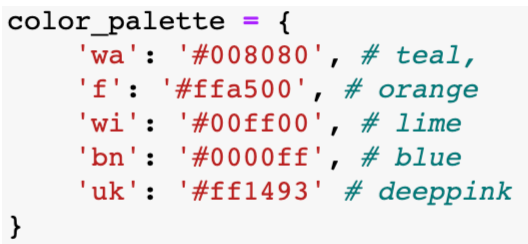
kdeplot of PC0 and PC1 for Aparc ComBat and Original dataset



	type	dataset	adjust	TSS	WSS	BSS
0	Aparc_wControl_Original	wControl	Original	6120.0	5212.834898	907.165102
0	Aparc_wControl_ComBat	wControl	ComBat	6120.0	6086.552188	33.447812
0	Aparc_wControl_ComBat_mod	wControl	ComBat_mod	6120.0	6058.608234	61.391766
0	Aparc_woControl_Original	woControl	Original	4692.0	3839.516258	852.483742
0	Aparc_woControl_ComBat	woControl	ComBat	4692.0	4656.109604	35.890396
0	Aparc_woControl_ComBat_mod	woControl	ComBat_mod	4692.0	4623.054037	68.945963
0	Aparc_woControl_rmwa_Original	woControl_rmwa	Original	4556.0	3872.277074	683.722926
0	Aparc_woControl_rmwa_ComBat	woControl_rmwa	ComBat	4556.0	4526.202902	29.797098
0	Aparc_woControl_rmwa_ComBat_mod	woControl_rmwa	ComBat_mod	4556.0	4498.746396	57.253604

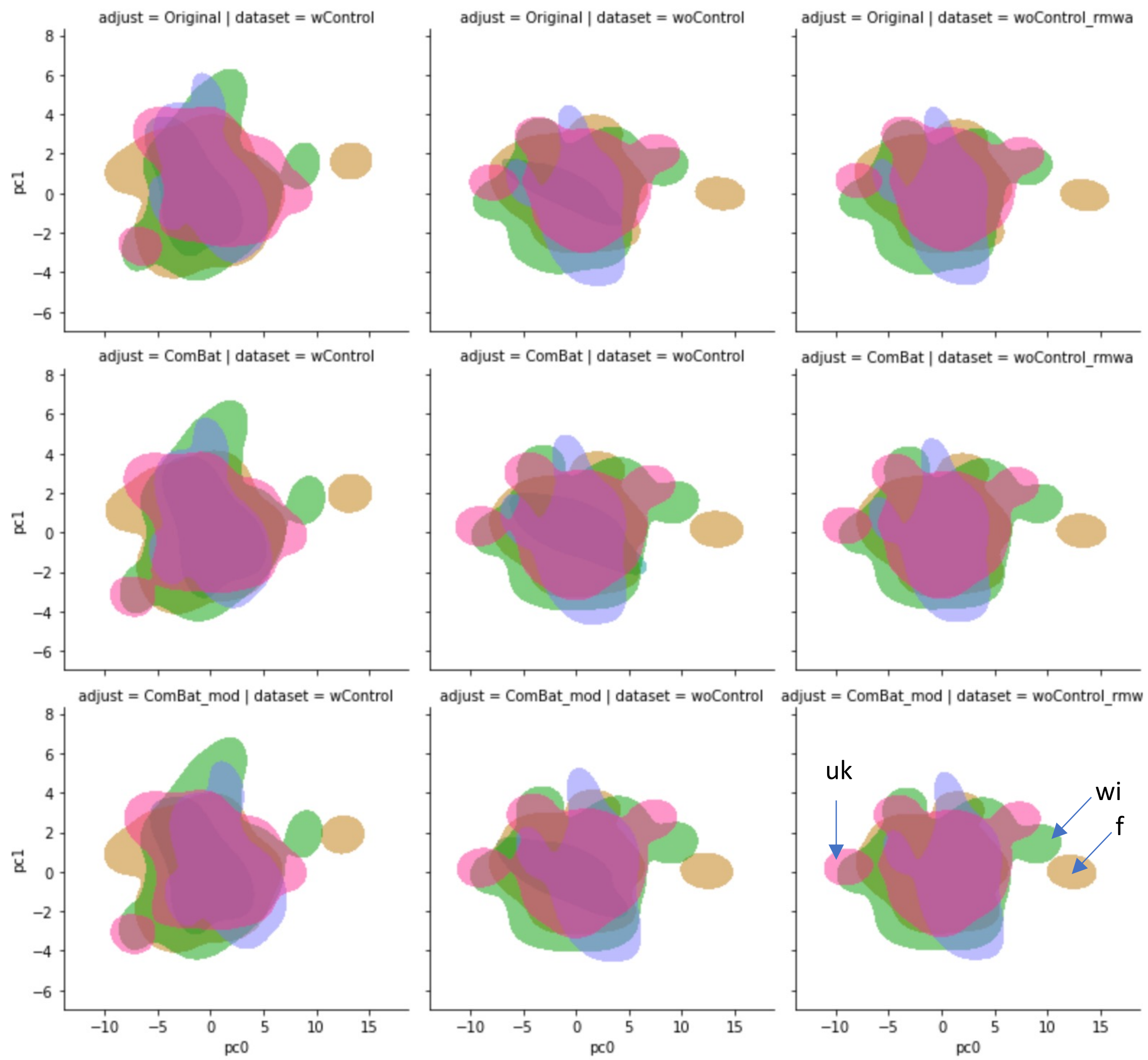


Mayo_Aparc: Cluster analysis

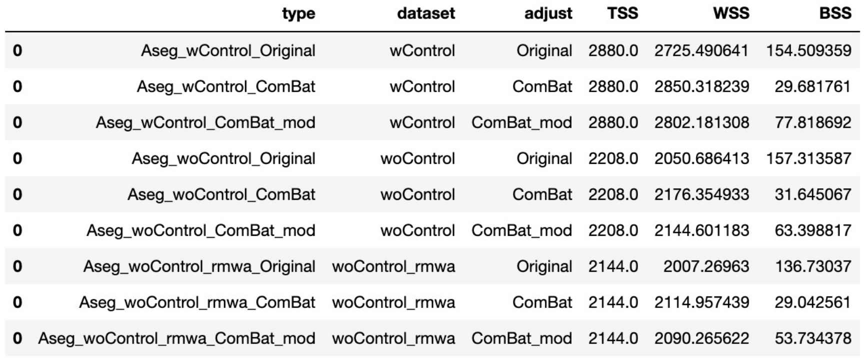


```
color_palette = {
    'wa': '#008080', # teal,
    'f': '#ffa500', # orange
    'wi': '#00ff00', # lime
    'bn': '#0000ff', # blue
    'uk': '#ff1493' # deeppink
}
```

Grand comparison: Aseg

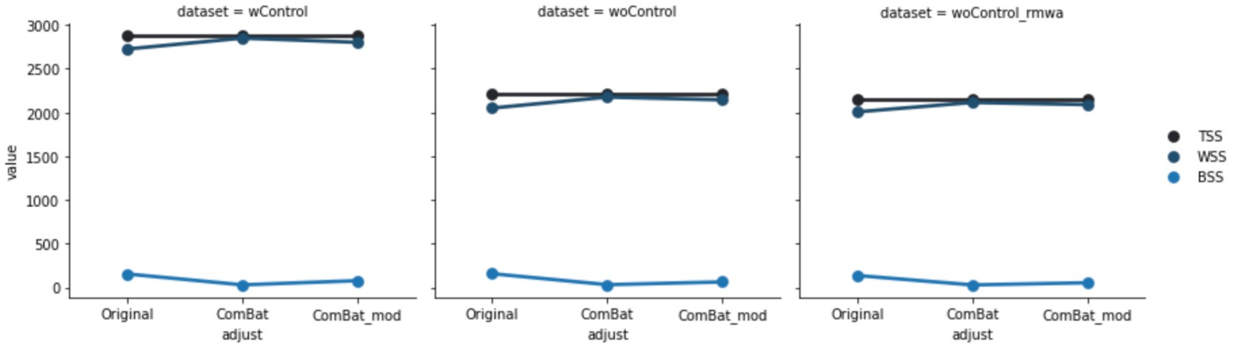


kdeplot of PC0 and PC1 for Aseg ComBat and Original dataset

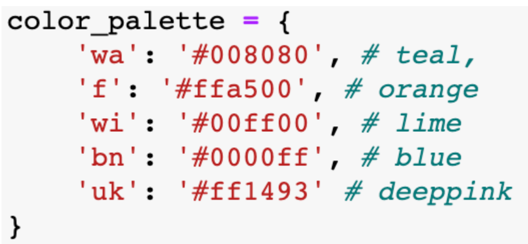


	type	dataset	adjust	TSS	WSS	BSS
0	Aseg_wControl_Original	wControl	Original	2880.0	2725.490641	154.509359
0	Aseg_wControl_ComBat	wControl	ComBat	2880.0	2850.318239	29.681761
0	Aseg_wControl_ComBat_mod	wControl	ComBat_mod	2880.0	2802.181308	77.818692
0	Aseg_woControl_Original	woControl	Original	2208.0	2050.686413	157.313587
0	Aseg_woControl_ComBat	woControl	ComBat	2208.0	2176.354933	31.645067
0	Aseg_woControl_ComBat_mod	woControl	ComBat_mod	2208.0	2144.601183	63.398817
0	Aseg_woControl_rmwa_Original	woControl_rmwa	Original	2144.0	2007.26963	136.73037
0	Aseg_woControl_rmwa_ComBat	woControl_rmwa	ComBat	2144.0	2114.957439	29.042561
0	Aseg_woControl_rmwa_ComBat_mod	woControl_rmwa	ComBat_mod	2144.0	2090.265622	53.734378

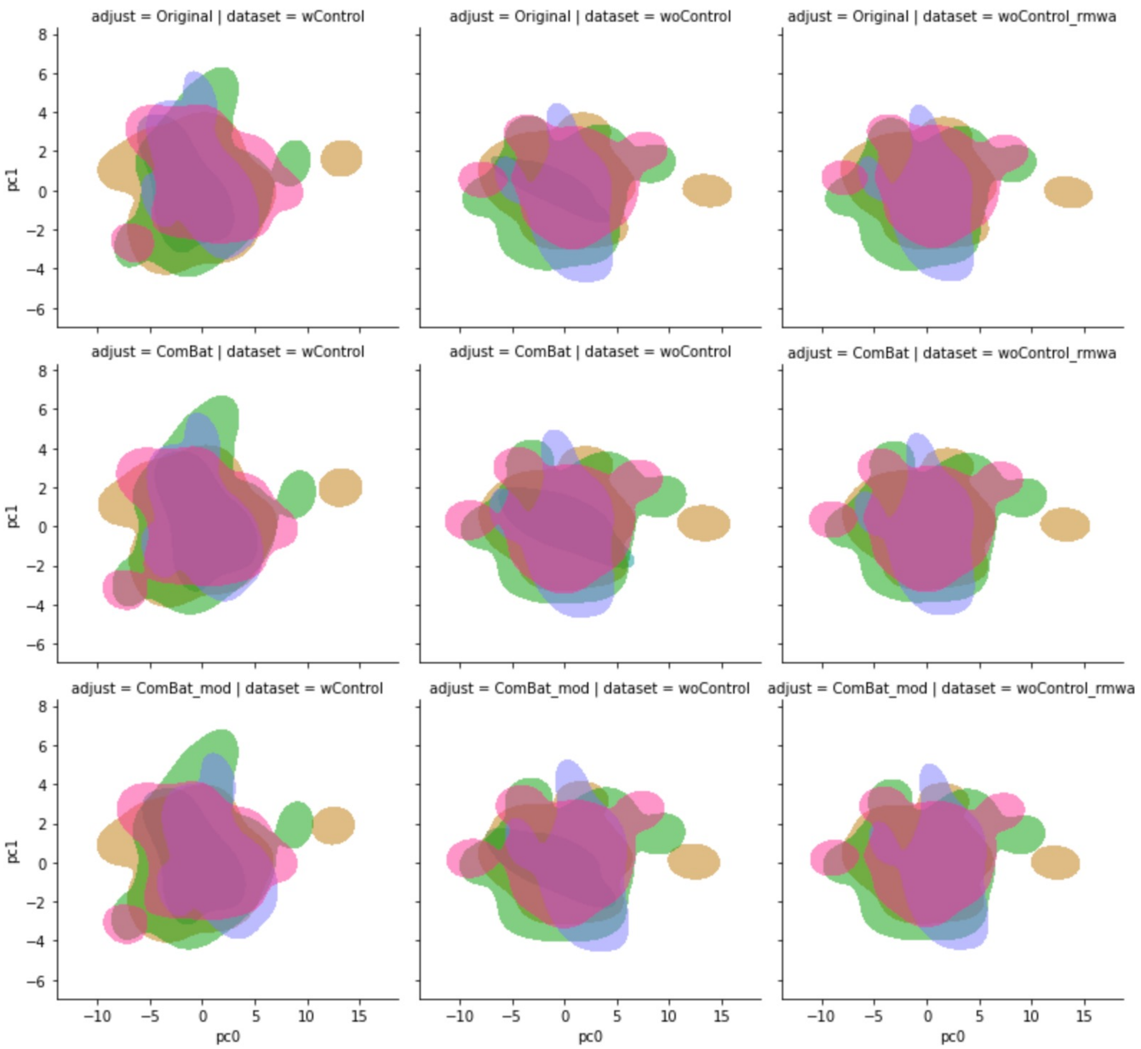
Mayo_Aseg: Cluster analysis



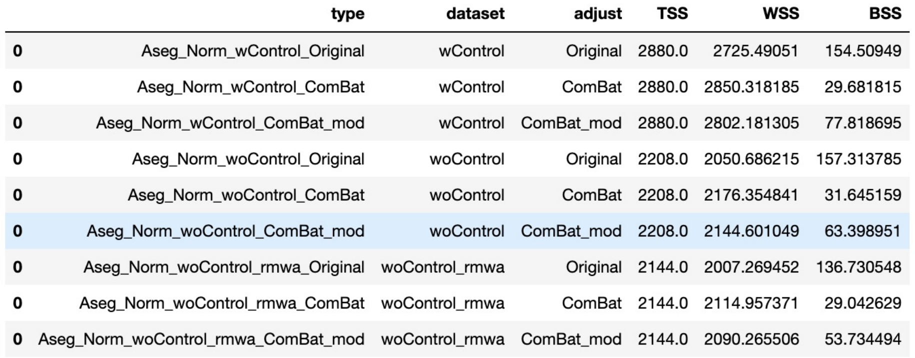
Grand comparison: Aseg_Norm



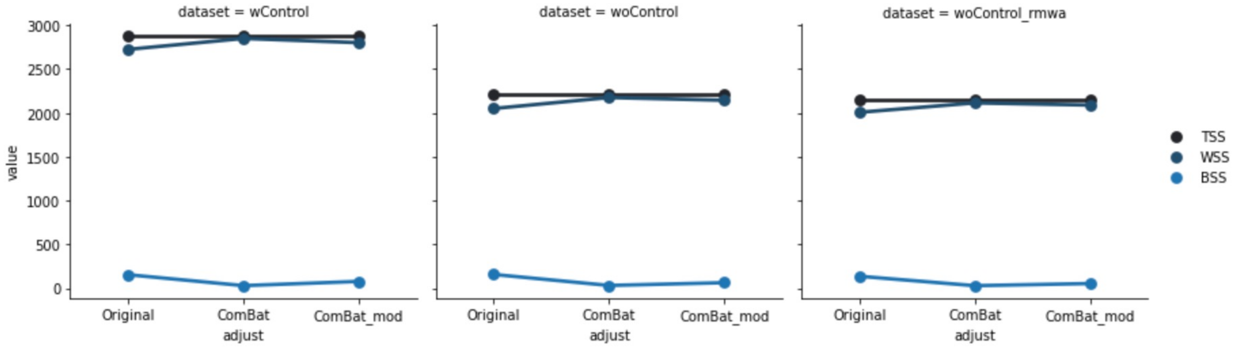
```
color_palette = {
    'wa': '#008080', # teal,
    'f': '#ffa500', # orange
    'wi': '#00ff00', # lime
    'bn': '#0000ff', # blue
    'uk': '#ff1493' # deeppink
}
```



kdeplot of PC0 and PC1 for Aseg_Norm ComBat and Original dataset



	type	dataset	adjust	TSS	WSS	BSS
0	Aseg_Norm_wControl_Original	wControl	Original	2880.0	2725.49051	154.50949
0	Aseg_Norm_wControl_ComBat	wControl	ComBat	2880.0	2850.318185	29.681815
0	Aseg_Norm_wControl_ComBat_mod	wControl	ComBat_mod	2880.0	2802.181305	77.818695
0	Aseg_Norm_woControl_Original	woControl	Original	2208.0	2050.686215	157.313785
0	Aseg_Norm_woControl_ComBat	woControl	ComBat	2208.0	2176.354841	31.645159
0	Aseg_Norm_woControl_ComBat_mod	woControl	ComBat_mod	2208.0	2144.601049	63.398951
0	Aseg_Norm_woControl_rmwa_Original	woControl_rmwa	Original	2144.0	2007.269452	136.730548
0	Aseg_Norm_woControl_rmwa_ComBat	woControl_rmwa	ComBat	2144.0	2114.957371	29.042629
0	Aseg_Norm_woControl_rmwa_ComBat_mod	woControl_rmwa	ComBat_mod	2144.0	2090.265506	53.734494



Mayo_Aseg_Norm: Cluster analysis

Grand comparison

- The ComBat largely adjusted the Mayo_Aparc subset, but not so much on the Mayo_Aseg/Mayo_Aseg_Norm.
- ComBat **decreases separation** (BSS) between batch (site), while **increases compactness** (WSS).
    - Note: K-Means algorithm tries to get the optimized points of the centroid, which minimize the value of WSS and maximize the value of BSS.
- Additionally, excluding WashU's data only have small effect on the cluster analysis statistics compare to that of including WashU's data.
- In the following detail analysis, only Mayo_Aparc and Mayo_Aseg_Norm will be comparing between the level of **Original vs ComBat_mod**, and between the level of **include vs exclude WashU 's samples**. The reason are:
    - Drop Control: Control should be excluded from this scope of study. Note: By removing the control subject, the total sum of squred reduce as expected.
    - Drop Mayo_Aseg: There is no different between Mayo_Aseg and Mayo_Aseg_Norm as expected, because their difference are only re-scaling.
    - Drop ComBat: the ComBat without counting other biological covariance are not reflecting the reality.

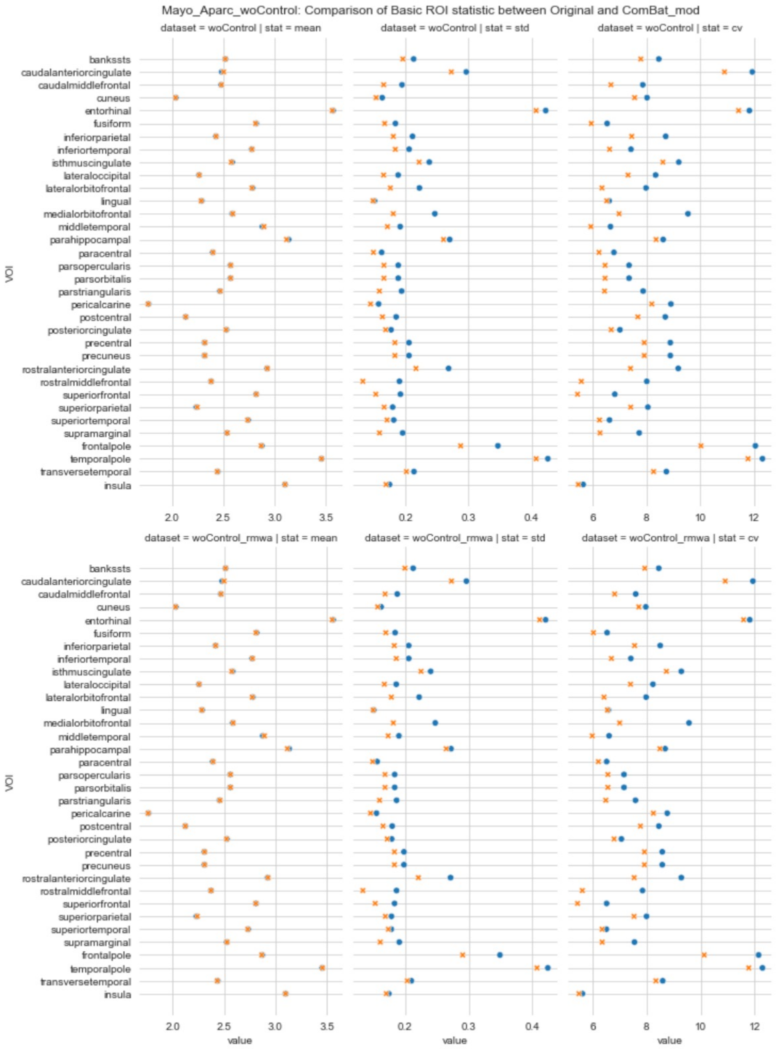
Method: Statistical analysis

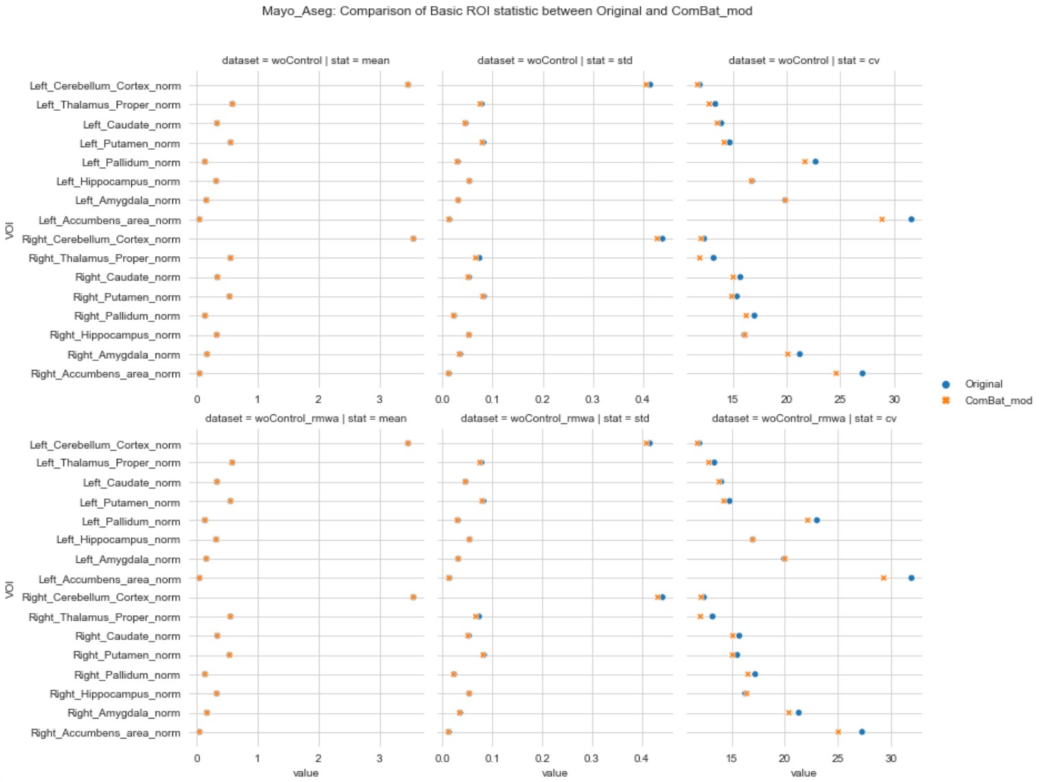
- To understand the Combat in detail of effect on biological covariance and each individual VOI, the following statistics are evaluated:
  - Mean
  - standard deviation (std)
  - coefficient of variation (cv = std/mean)
- Notation:
  - increase (+),
  - decrease (-),
  - mix of increase and decrease (+-),
  - undistinguisable change (~)

Statistics of VOI variables

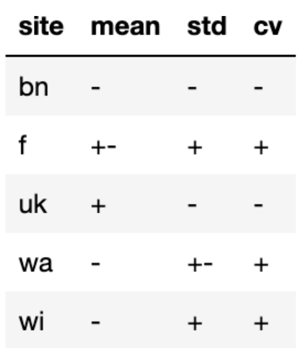
- As expected, ComBat models consistently reduce the std and CV of each VOI (with some pull back compared to ComBat model without considering biological, which is not shown here)
- Mayo_Aparc and Mayo_Aseg_Norm both have the same pattern as the following (Adjust - Original):

dataset	mean	std	cv
woControl	~	-	-
woControl_rmwa	~	-	-

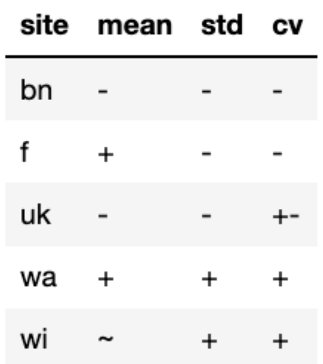

Mayo_Aparc_woControl: Comparison of Basic ROI statistic between Original and ComBat_mod


Mayo_Aseg: Comparison of Basic ROI statistic between Original and ComBat_mod

Statistics of VOI variables by site

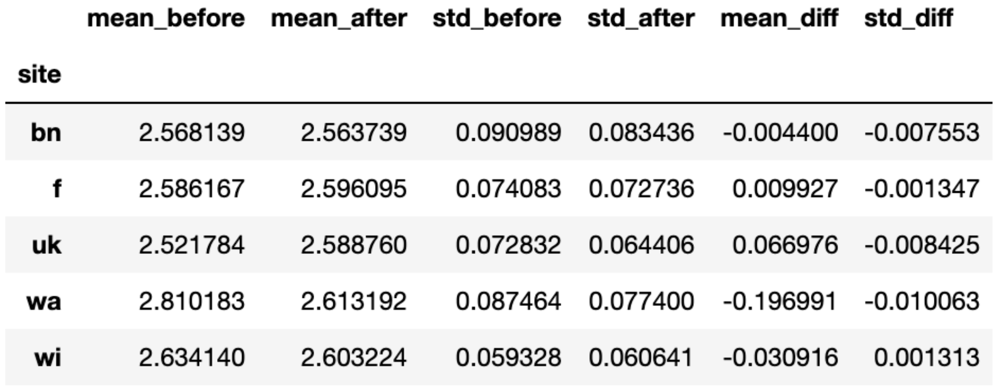


site	mean	std	cv
bn	-	-	-
f	+-	+	+
uk	+	-	-
wa	-	+-	+
wi	-	+	+



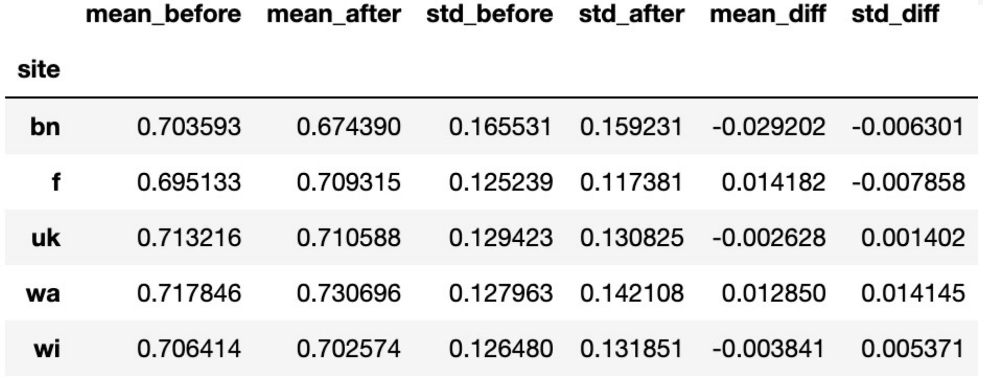
site	mean	std	cv
bn	-	-	-
f	+	-	-
uk	-	-	+-
wa	+	+	+
wi	~	+	+

- As expected, The ComBat location/scale adjustment vary by batches (site).
- The result for two data are summarized at the following (Adjust - Original):



site	mean_before	mean_after	std_before	std_after	mean_diff	std_diff
bn	2.568139	2.563739	0.090989	0.083436	-0.004400	-0.007553
f	2.586167	2.596095	0.074083	0.072736	0.009927	-0.001347
uk	2.521784	2.588760	0.072832	0.064406	0.066976	-0.008425
wa	2.810183	2.613192	0.087464	0.077400	-0.196991	-0.010063
wi	2.634140	2.603224	0.059328	0.060641	-0.030916	0.001313

Mayo_Aparc_woControl

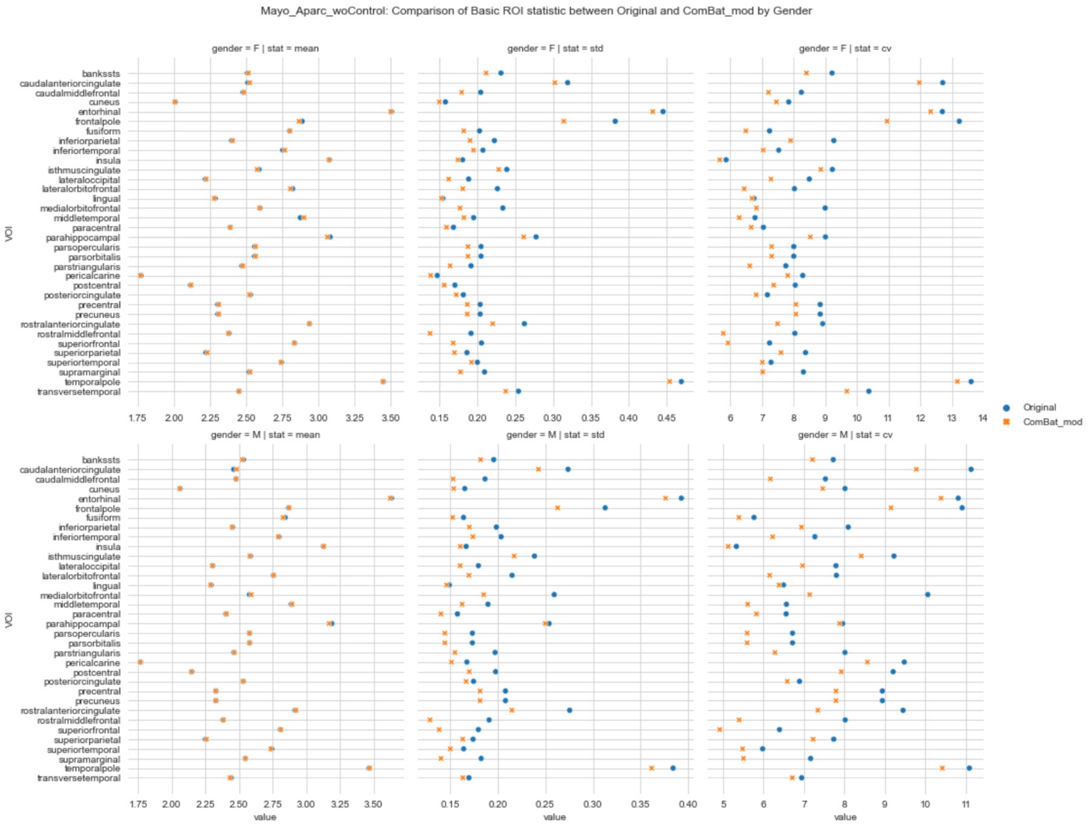


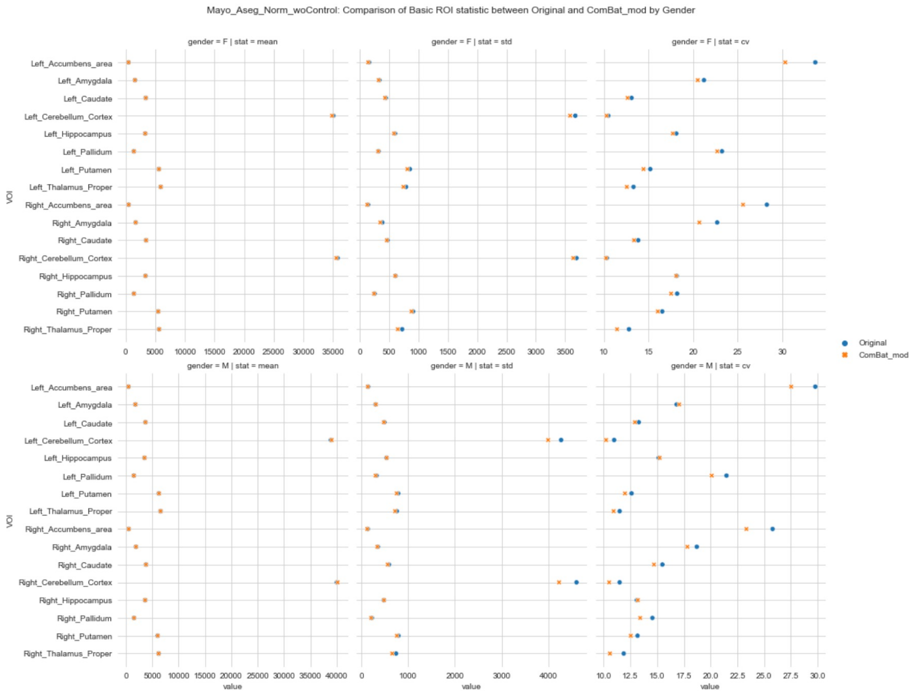
site	mean_before	mean_after	std_before	std_after	mean_diff	std_diff
bn	0.703593	0.674390	0.165531	0.159231	-0.029202	-0.006301
f	0.695133	0.709315	0.125239	0.117381	0.014182	-0.007858
uk	0.713216	0.710588	0.129423	0.130825	-0.002628	0.001402
wa	0.717846	0.730696	0.127963	0.142108	0.012850	0.014145
wi	0.706414	0.702574	0.126480	0.131851	-0.003841	0.005371

Mayo_Aseg_Norm_woControl

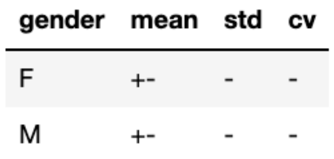
- No Need to exclude WashU' data.
  - Removing WashU' small data slightly change the mean effect in the model, slightly reduce the variance of the total sum of square in the adjusted data.
  - The other batch were not so different after ComBat when excluding WashU's data.

Statistics of VOI variables by Gender





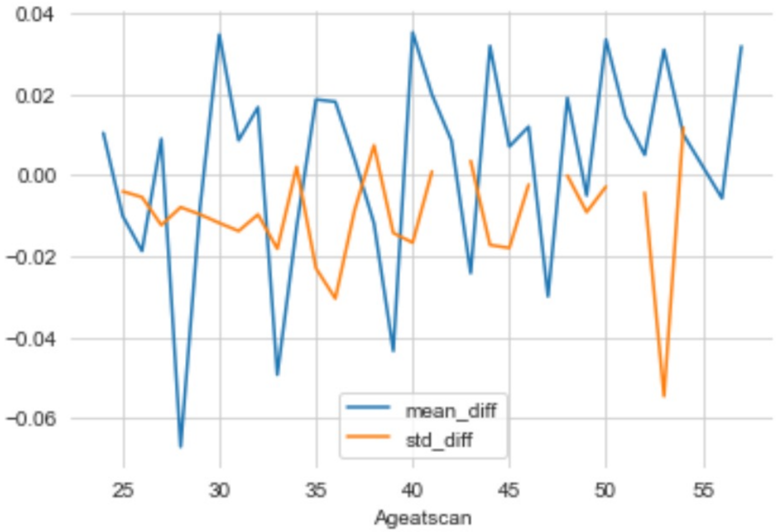
- For both male and female, ComBat adjusts the data by both up and down to reduce the std and cv of each VOI within each gender group.
- Mayo_Aparc and Mayo_Aseg_Norm both have the same pattern as the following (Adjust - Original):



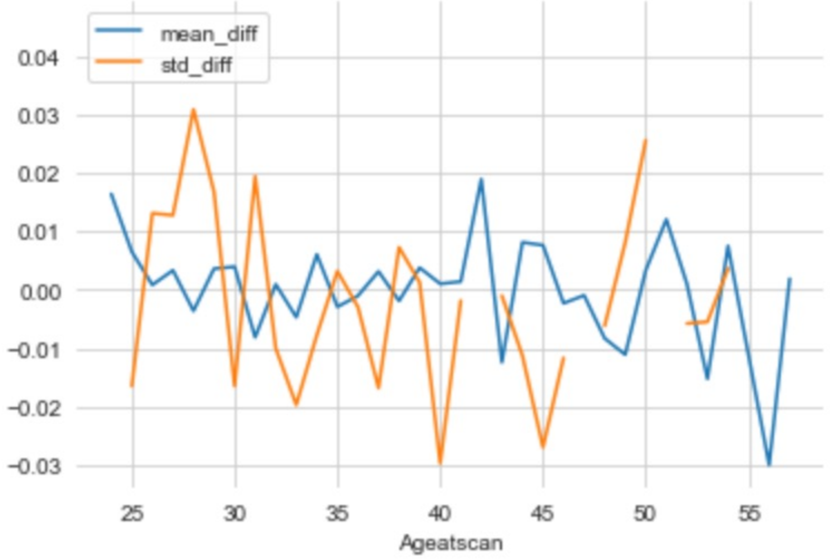
gender	mean	std	cv
F	+-	-	-
M	+-	-	-

Statistics of VOI variables by Age

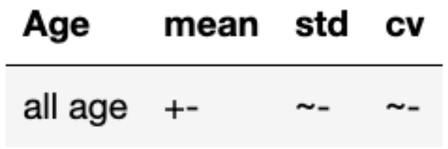
- across age, ComBat adjusts the data by both up and down to reduce the std and cv of each VOI within each age group.
- Mayo_Aparc and Mayo_Aseg_Norm Both has the same pattern as the following:



Mayo_Aparc



Mayo_Aseg_Norm



Age	mean	std	cv
all age	+-	~-	~-