Clustering and t-test

T-test for change in Clinical Outcome with p values

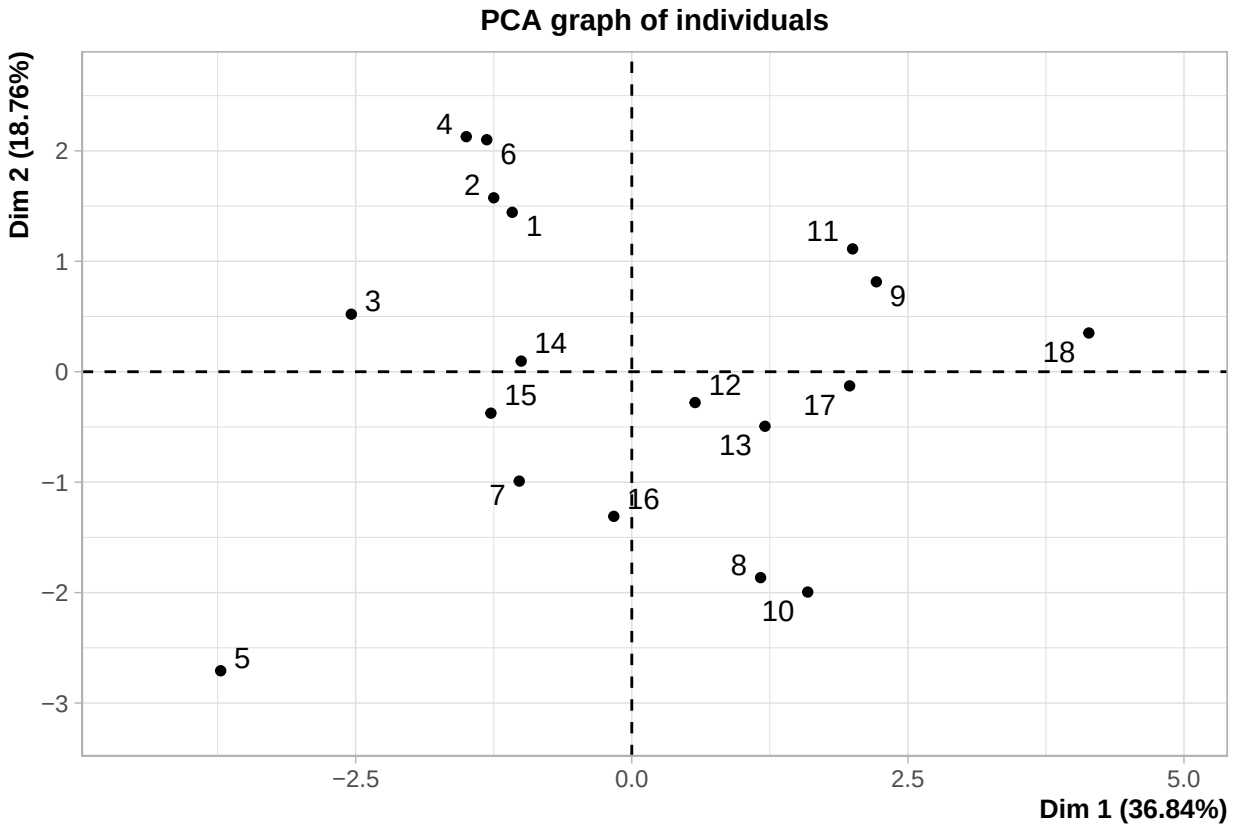
$$\% \; change \; in \; outcome = \frac{first \; read - second \; read}{first \; read}$$

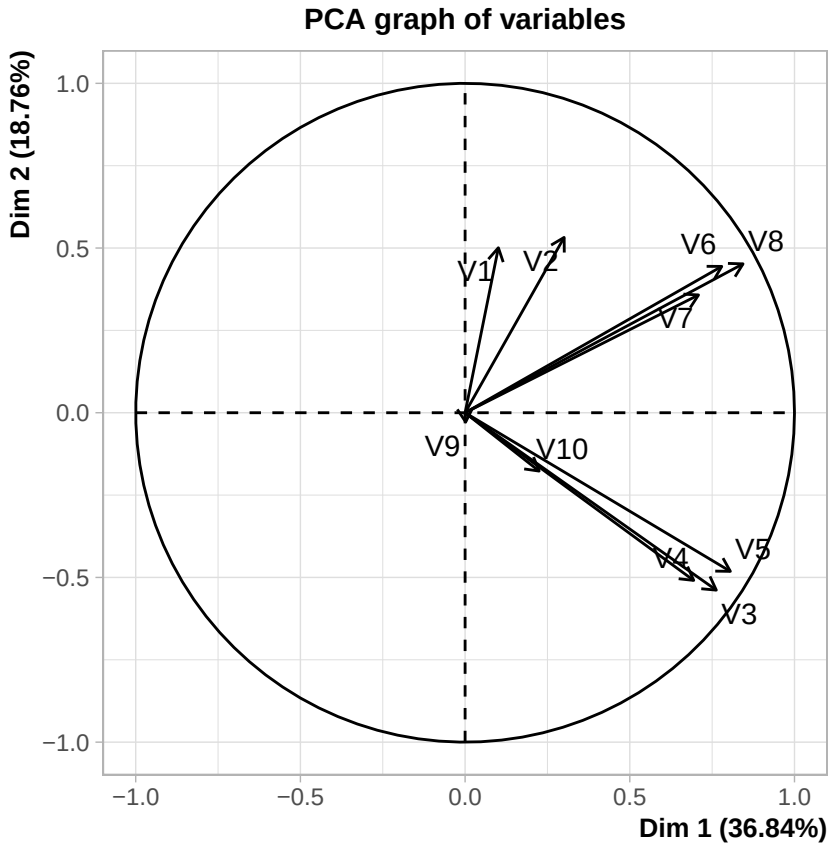
```
data <- read_excel("finaldata.xlsx", sheet = NULL, col_types = c(rep("guess", 6),
                                                                 "date", rep("guess", 2),
                                                                 rep("numeric", 10)))

status <- factor(c("low", rep("intermediate", 4),
                   "low", "intermediate", "intermediate",
                   rep("low", 2), rep("intermediate", 3),
                   "low", rep("intermediate", 2), rep("low", 2)))

status <- forcats::fct_relevel(status, c("low"))

#smoking <- as.factor(c("low", "high", "low", "high", "high", "low",
#                       rep("high", 3), rep("low", 3), "high", "NA", "high", "low"))
#smoking <- smoking[-c(8, 13, 18:20)]

smoking <-
patient <- 1:18

first_read <- data[seq(1, 54, 3), ]
second_read <- data[seq(2, 54, 3), ]

pval <- NULL
for(i in 10:19){
  change <- 1 - as.numeric(unlist(second_read[, i] / first_read[, i]))
  cat(names(second_read[, i]), "with p-value of", t.test(change ~ status)$p.value, "\n")
  pval <- c(pval, t.test(change ~ status)$p.value)
}
```

```
## Tumor Volume (mL) with p-value of 0.651912
## Largest Node Volume (mL) with p-value of 0.06978384
## SULmax Tumor with p-value of 0.2943616
## SULmedian with p-value of 0.3630501
## SULpeak with p-value of 0.3854198
## SULmax Largest Node with p-value of 0.0148899
## SULmedian node with p-value of 0.6197499
## SULpeak node with p-value of 0.03951802
## Diffusion Mean Tumor with p-value of 0.1825276
## Diffusion mean ADC Largest Node with p-value of 0.7309697
```

```
change_matrix <- array(NA, c(18, 10))

for(i in 1:10){
  change_matrix[, i] <- 1 - as.numeric(unlist(second_read[1:18, i + 9] / first_read[1:18, i + 9]))
}

change_matrix[is.na(change_matrix)] <- 0
```

```
pca <- PCA(change_matrix)
```



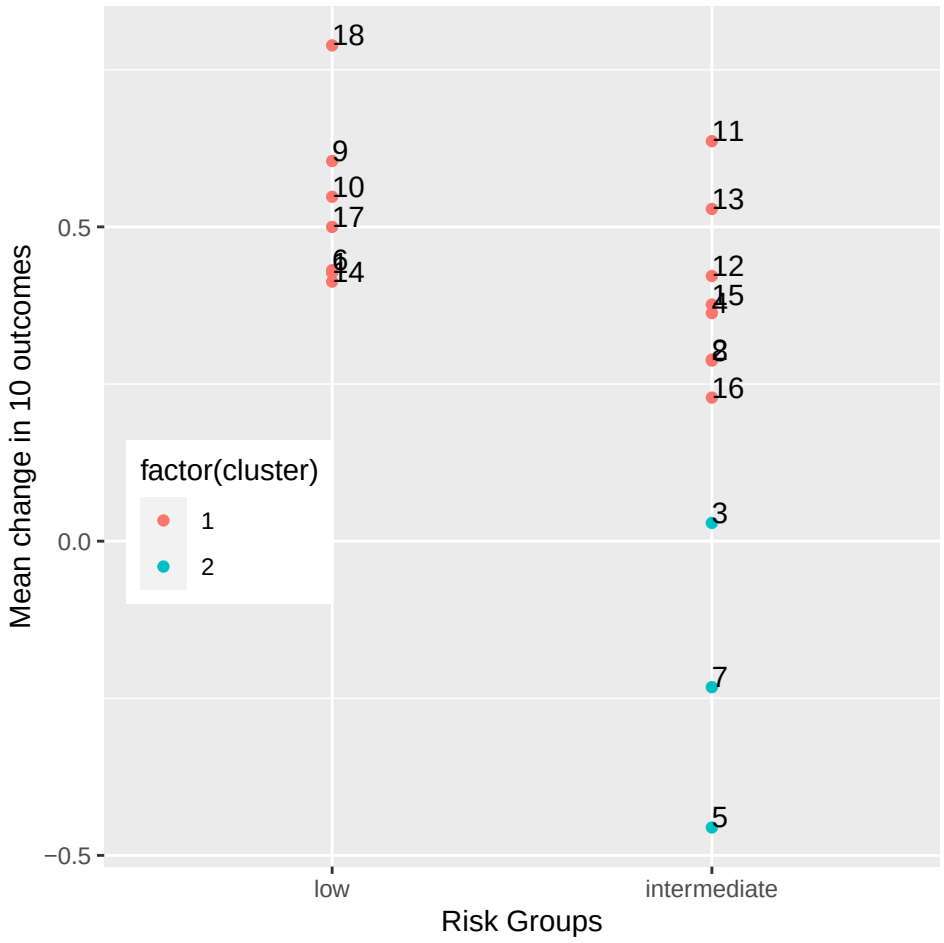



```
knitr::kable(cbind(names(data)[10:19], as.numeric(round(pca$var$coord[, 1], 3))))
```

Tumor Volume (mL)	0.101
Largest Node Volume (mL)	0.301
SULmax Tumor	0.763
SULmedian	0.694
SULpeak	0.805
SULmax Largest Node	0.779
SULmedian node	0.709
SULpeak node	0.844
Diffusion Mean Tumor	0.002
Diffusion mean ADC Largest Node	0.225

```
clust <- kmeans(change_matrix, 2, iter.max = 1000, nstart = 1000)
clust <- kmeans(change_matrix[, 6], 2, iter.max = 1000, nstart = 1000)

dat <- data.frame(risk = status, cluster = factor(clust$cluster), change = change_matrix[, 6],
                  patient = patient)
```

```
ggplot(dat, aes(risk, change, label = patient)) +
  geom_point(aes(colour = factor(cluster))) +
  geom_text(aes(label = patient), hjust = 0, vjust = 0) +
  labs(x = "Risk Groups", y = "Mean change in 10 outcomes") +
  theme(legend.position=c(0.15, 0.4),
        strip.background = element_blank())
```




```
# Patient with in cluster 1
dat$patient[dat$cluster == 1]
```

```
##  [1]  1  2  4  6  8  9 10 11 12 13 14 15 16 17 18
```

```
# Patient in cluster 2
dat$patient[dat$cluster == 2]
```

```
## [1] 3 5 7
```

```
pdf("clusterplot.pdf")
ggplot(dat, aes(risk, change, label = patient)) +
  geom_point(aes(colour = cluster)) +
  geom_text(aes(label = patient),hjust=0, vjust=0) +
  labs(x = "Risk Groups", y = "Mean change in 10 outcomes") +
  theme(legend.position=c(0.15, 0.4),
        strip.background = element_blank())
dev.off()
```

```
## pdf
##   2
```

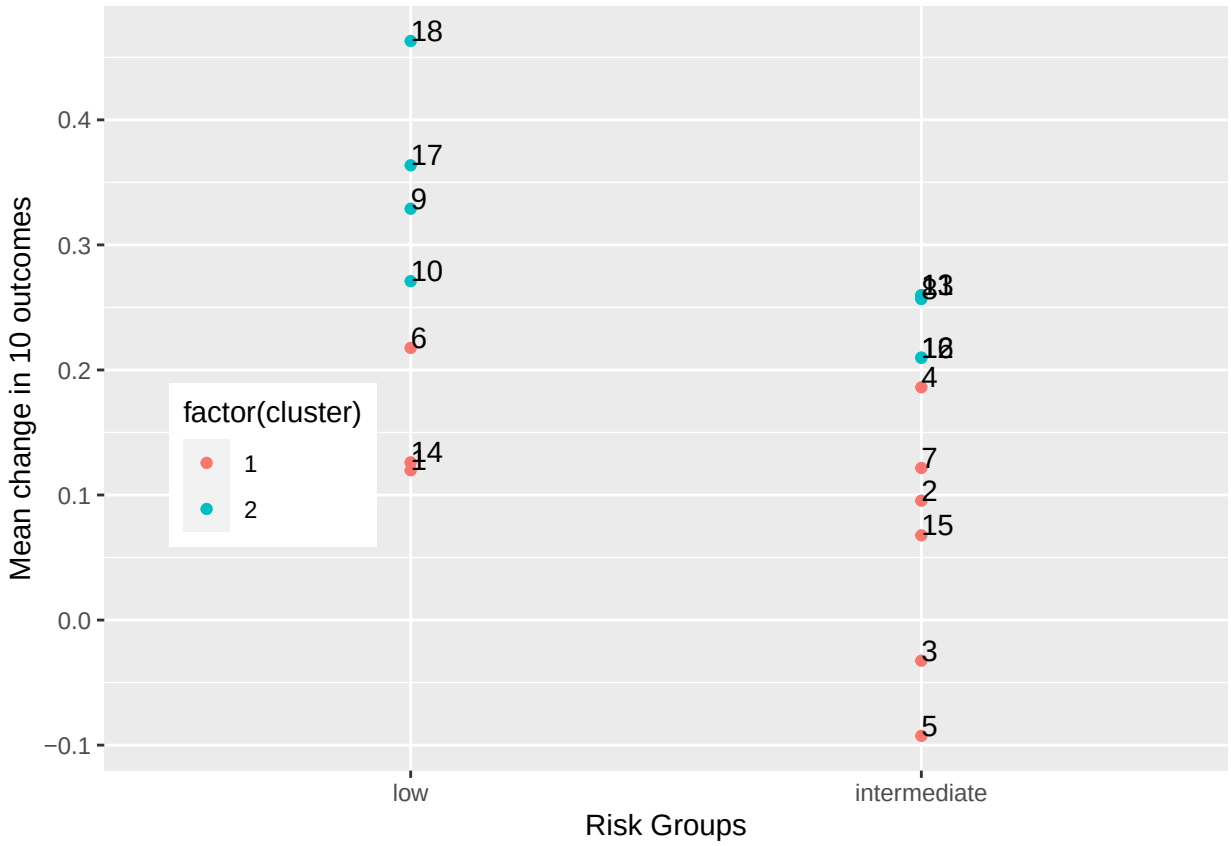
```
knitr::kable(data.frame(Patient = dat$patient, Cluster = dat$cluster, value = round(change_matrix[, 6],
```

Patient	Cluster	value
1	1	0.427
2	1	0.289
3	2	0.029
4	1	0.363
5	2	-0.456
6	1	0.431
7	2	-0.232
8	1	0.287
9	1	0.605
10	1	0.548
11	1	0.636
12	1	0.422
13	1	0.529
14	1	0.413
15	1	0.376
16	1	0.228
17	1	0.500
18	1	0.789

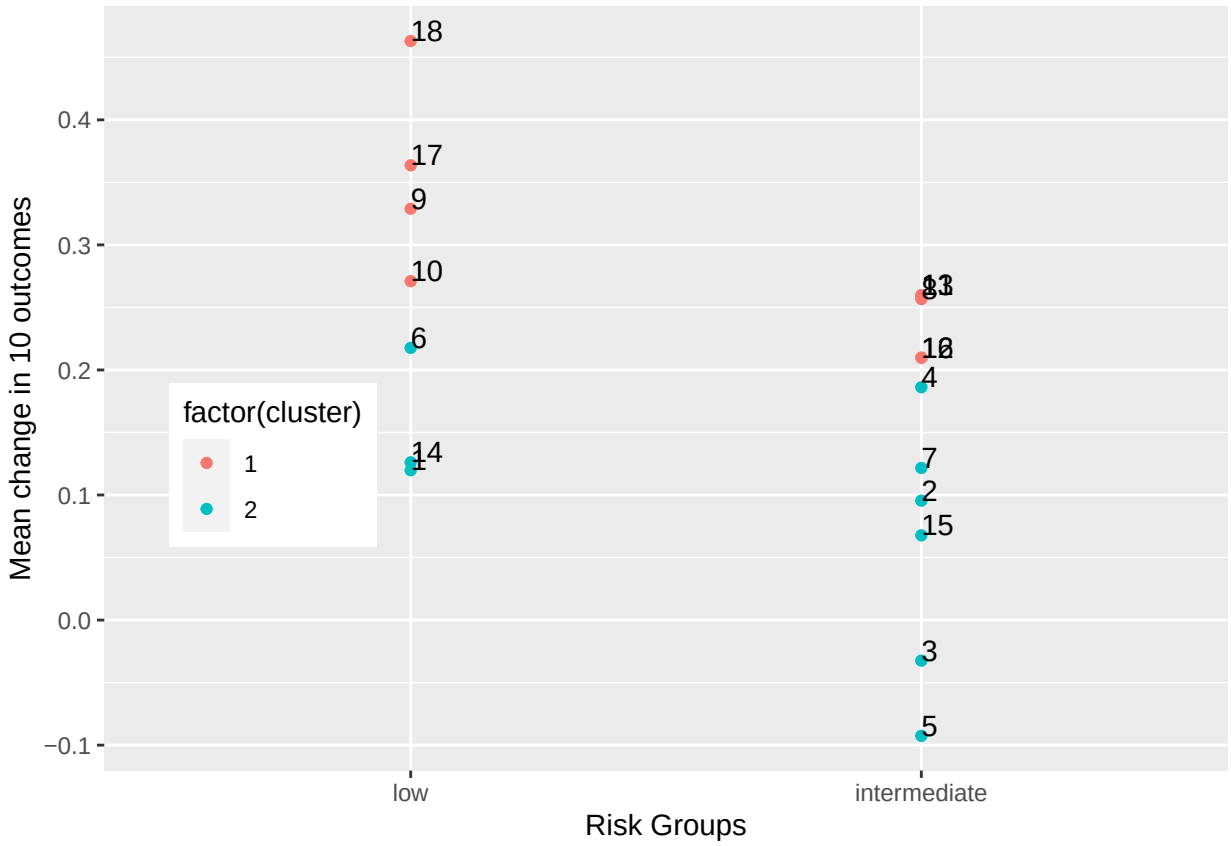
Cross validation

```
grouping  <- array(NA, c(18, 10))

meanchange <- apply(change_matrix, 1, function(x){mean(x, na.rm = T)})

for(i in 1:10){
  clust <- kmeans(change_matrix[, -i], 2, iter.max = 1000, nstart = 1000)
  dat <- data.frame(risk = status, cluster = factor(clust$cluster),
                    change = meanchange, patient = patient)
  cat("Removing ", names(second_read[i + 9]), "\n")
  print(ggplot(dat, aes(risk, change, label = patient)) +
  geom_point(aes(colour = factor(cluster))) +
  geom_text(aes(label = patient), hjust = 0, vjust = 0) +
  labs(x = "Risk Groups", y = "Mean change in 10 outcomes") +
  theme(legend.position=c(0.15, 0.4),
        strip.background = element_blank()))

  grouping[, i] <- clust$cluster
}
```

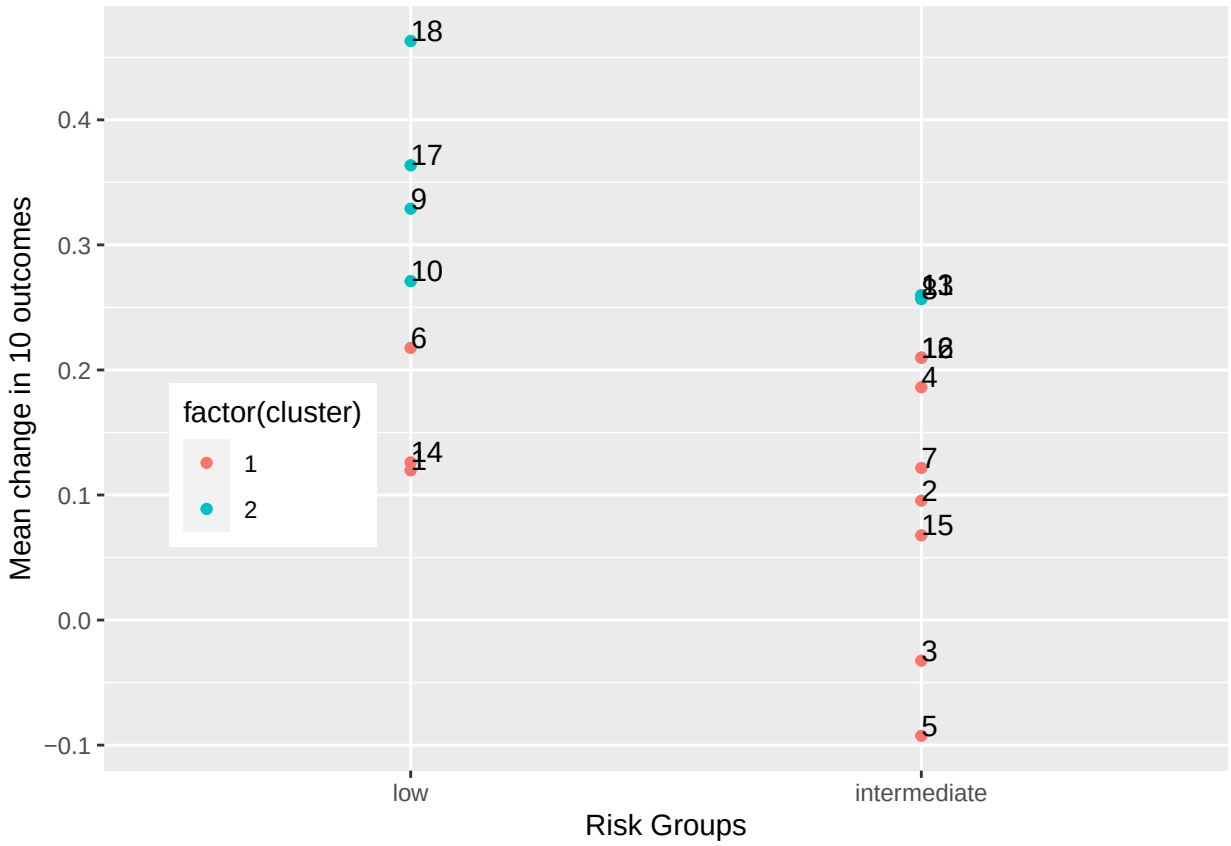
```
## Removing  Tumor Volume (mL)
```



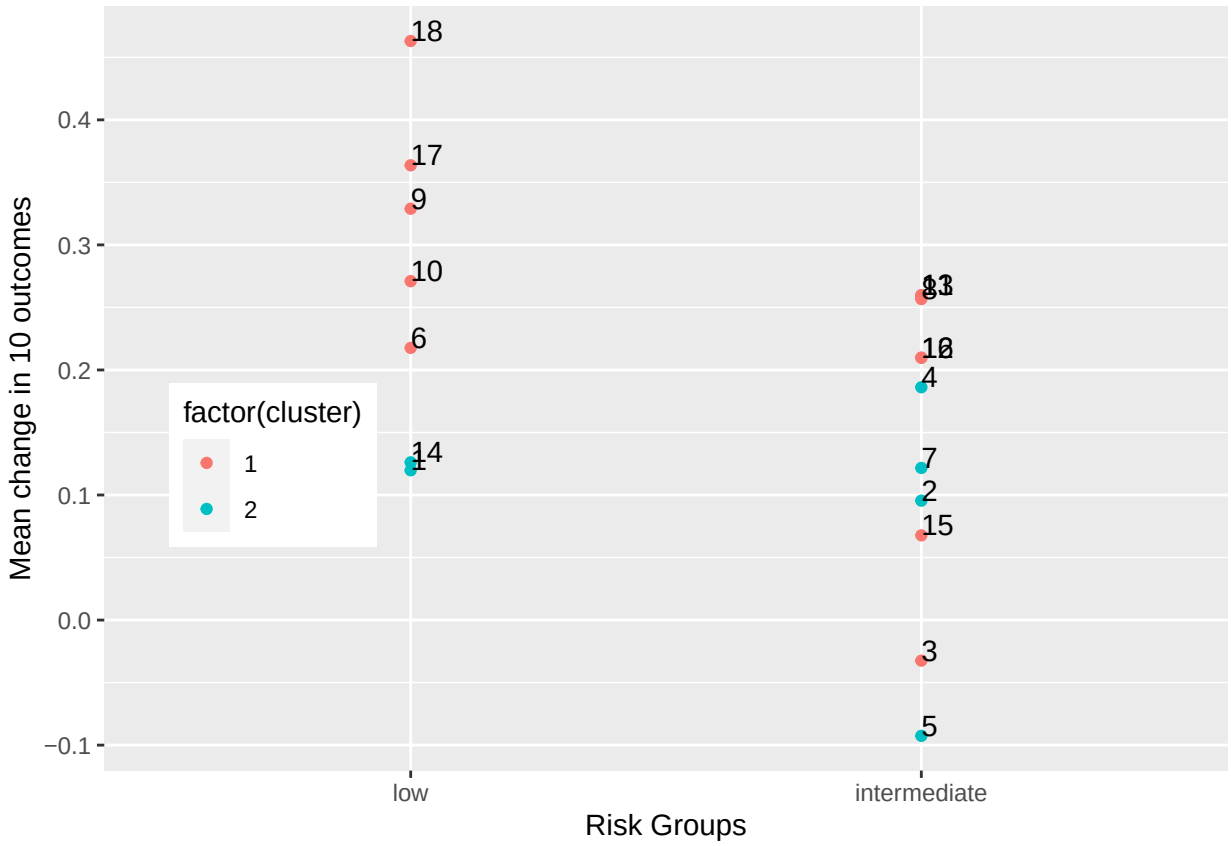

```
## Removing  Largest Node Volume (mL)
```



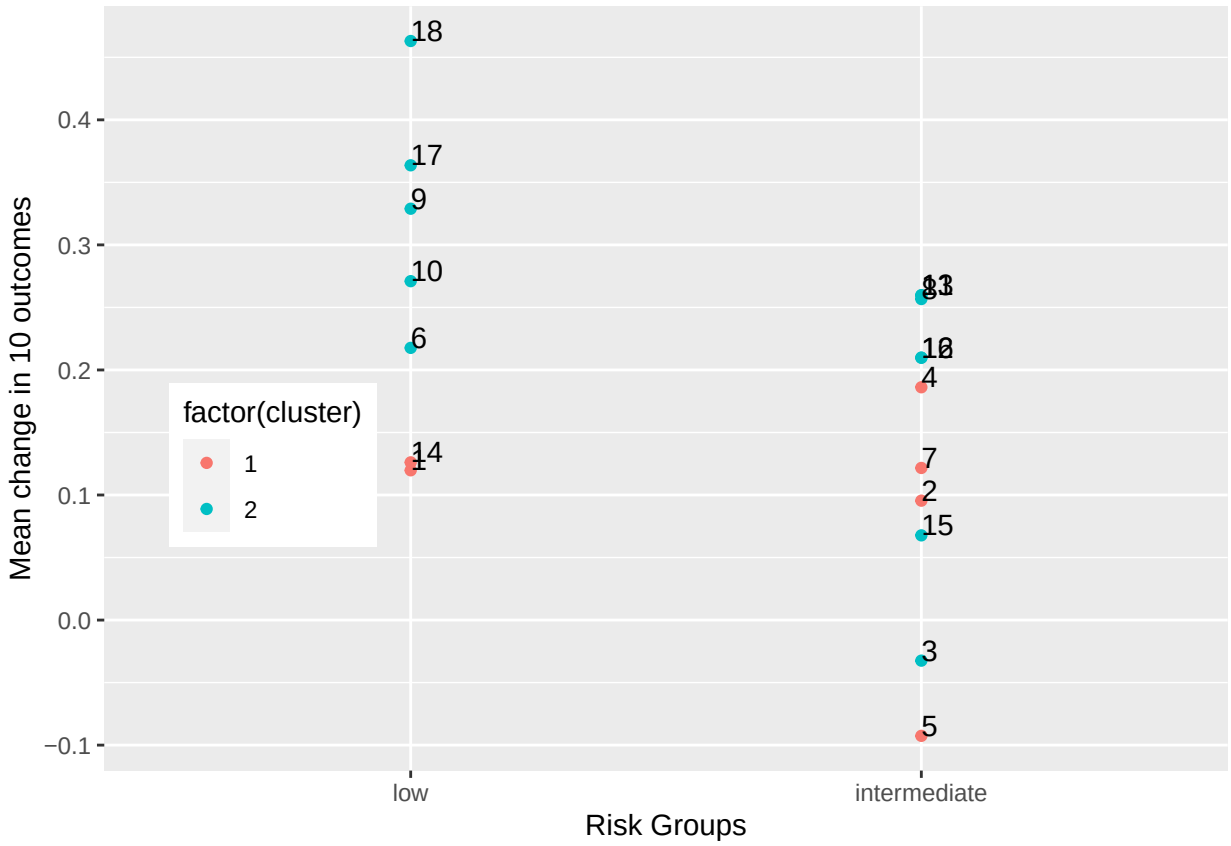

```
## Removing  SULmax Tumor
```



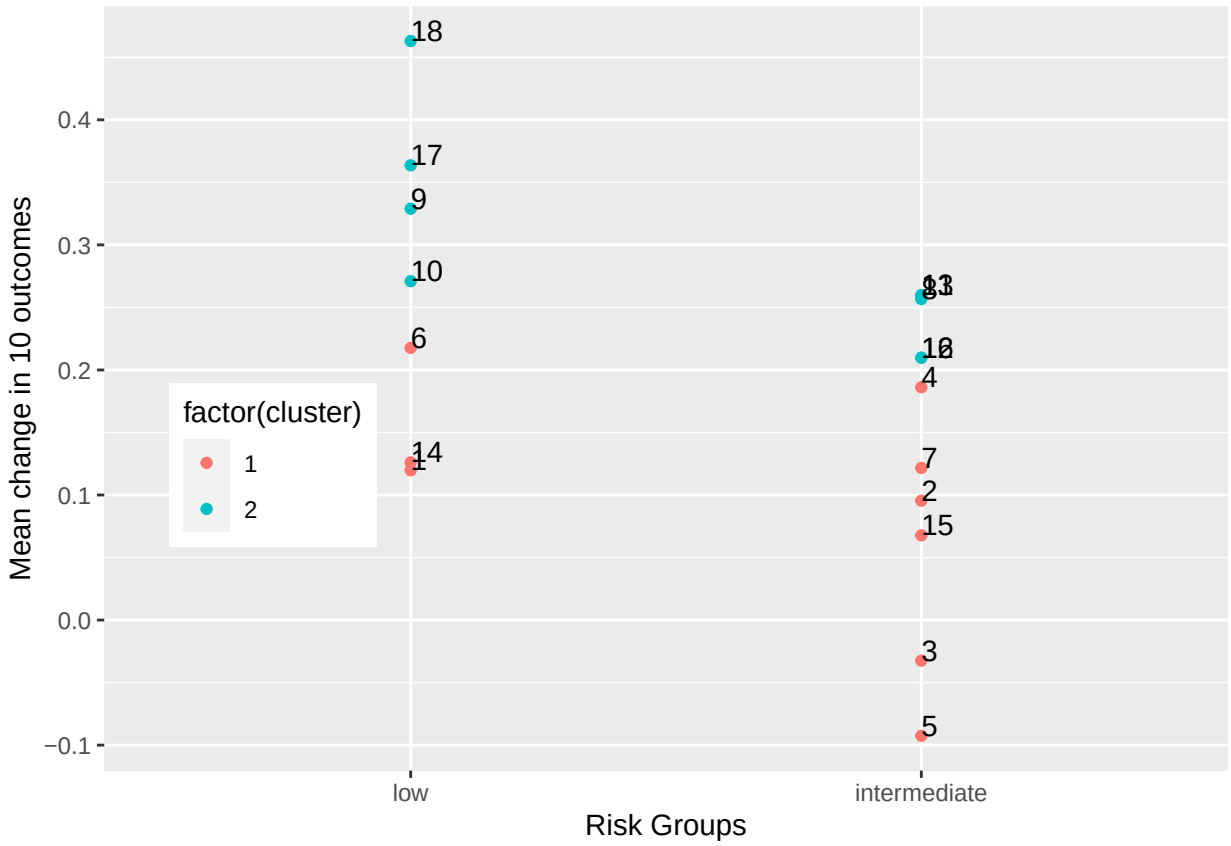

```
## Removing  SULmedian
```



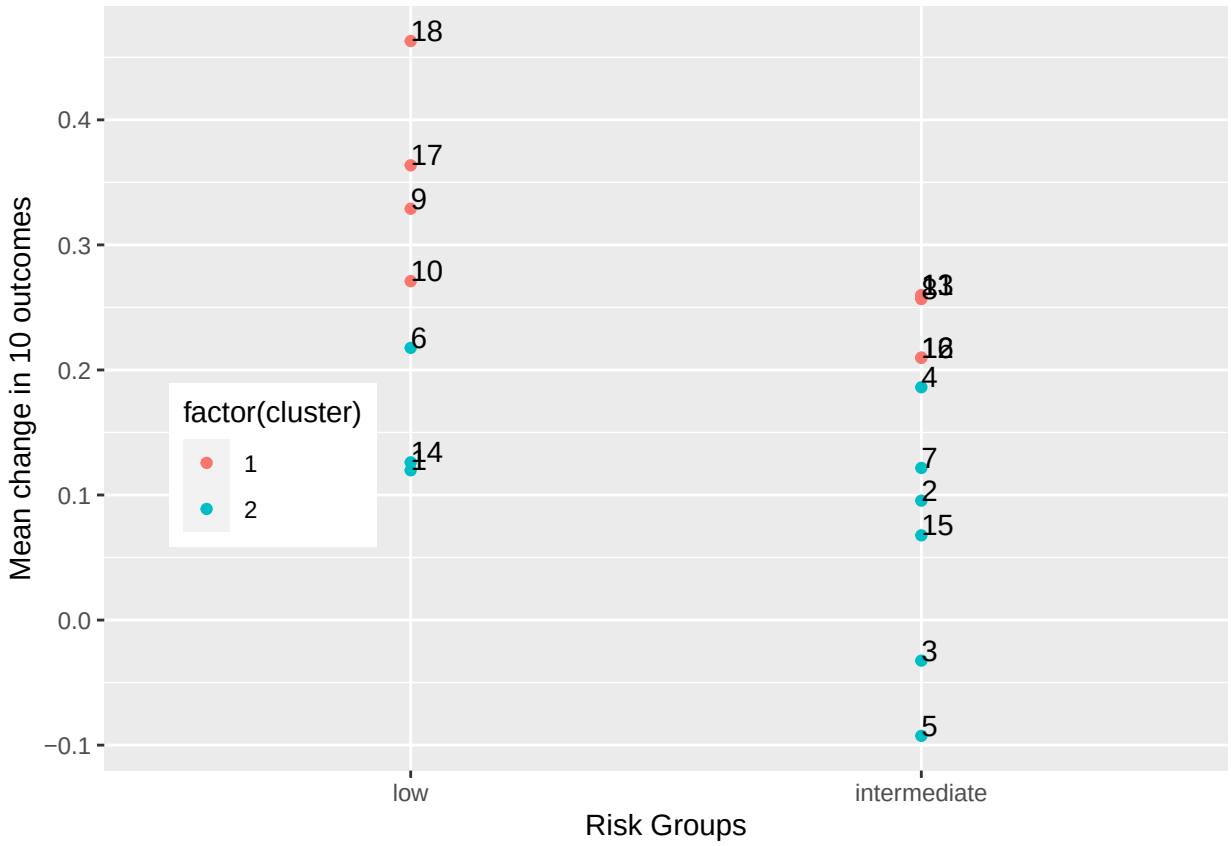

```
## Removing  SULpeak
```



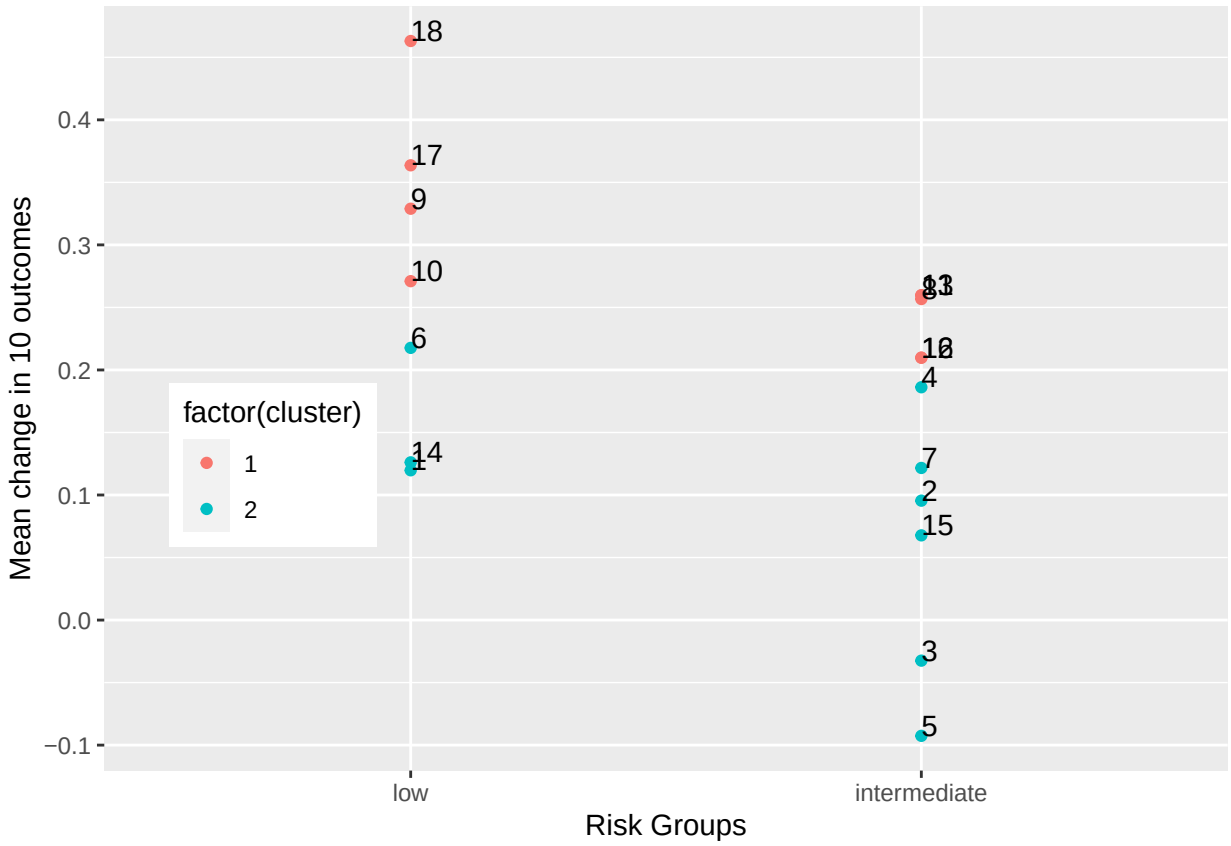

```
## Removing  SULmax Largest Node
```



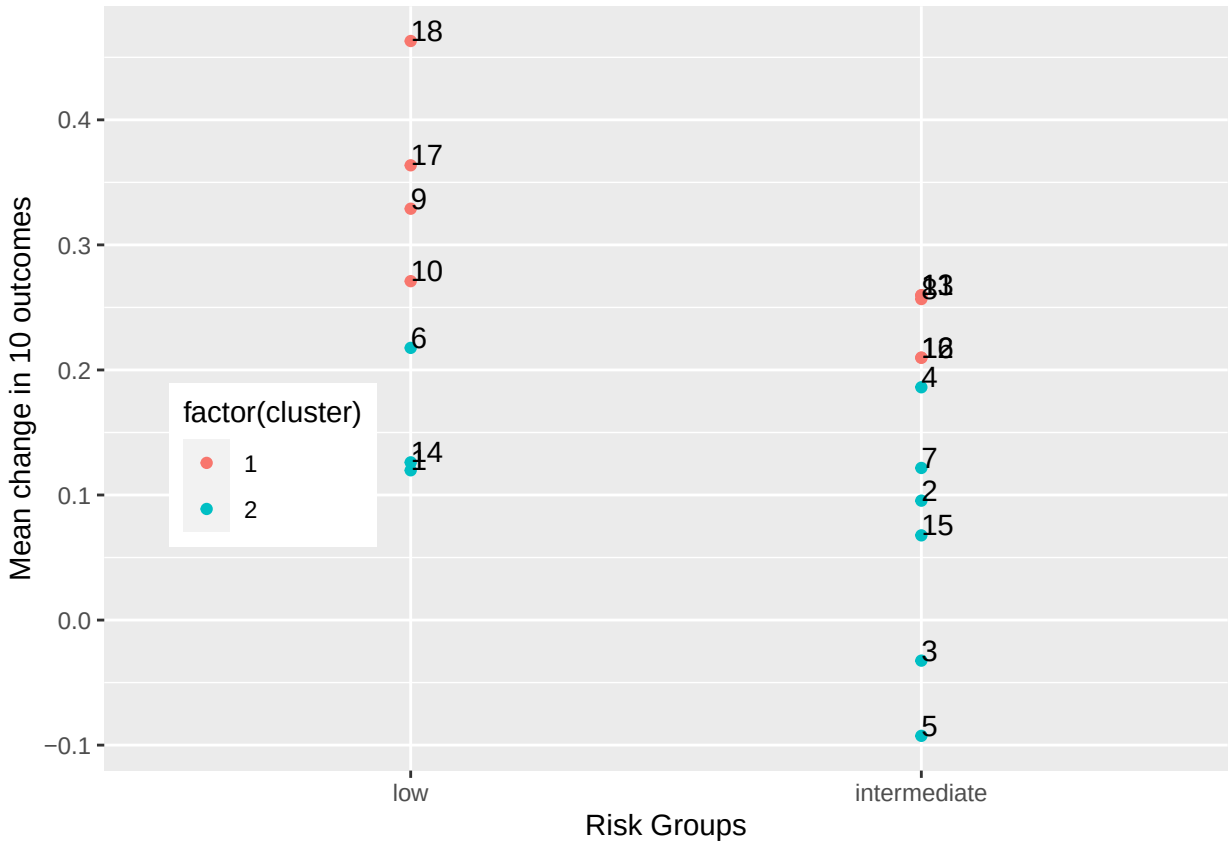

```
## Removing  SULmedian node
```



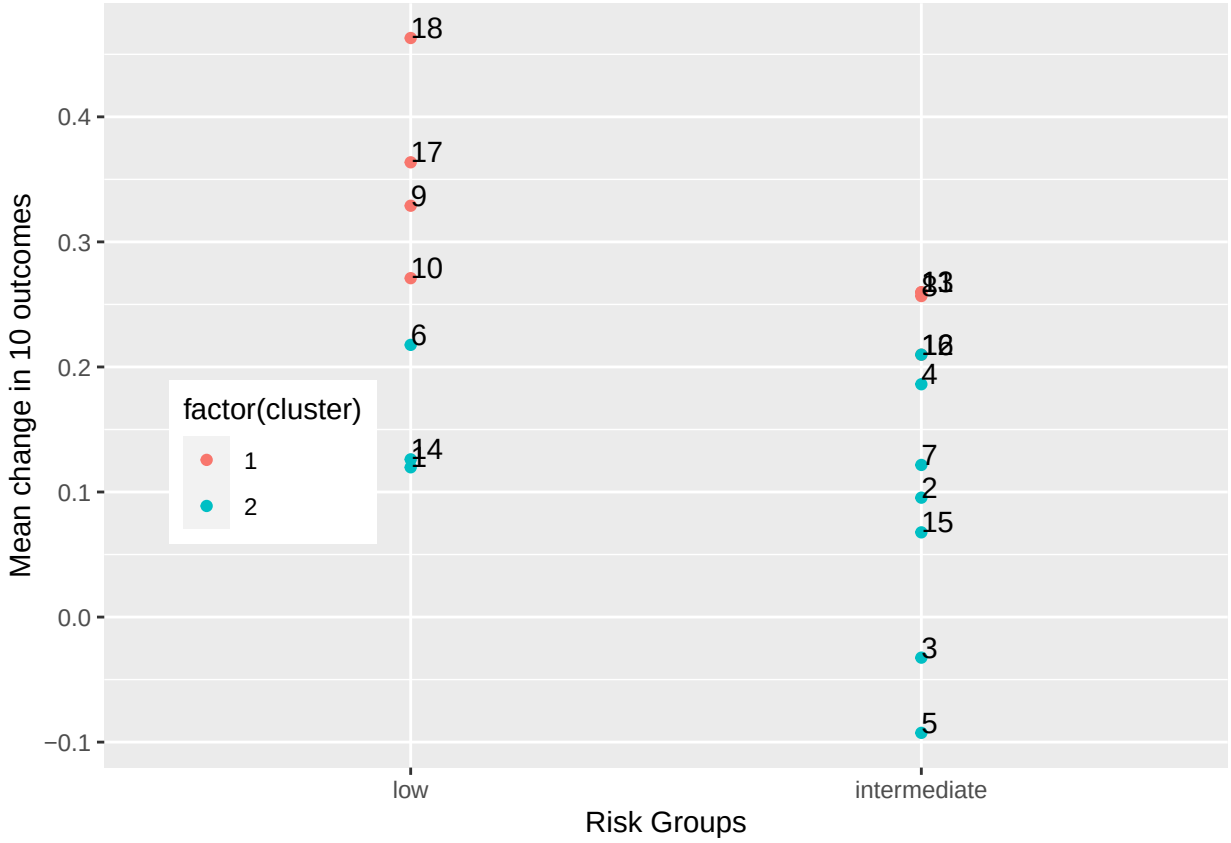

```
## Removing  SULpeak node
```




```
## Removing  Diffusion Mean Tumor
```




```
## Removing  Diffusion mean ADC Largest Node
```



Clustering table after removing each column

```
colnames(grouping) <- names(second_read[10:19])

knitr::kable(grouping)
```

Tumor Volume (mL)	Largest Node Volume (mL)	SULmax Tumor	SULmedian	SULpeak	SULmax Largest Node	SULmedian node	SULpeak node	Diffusion Mean Tumor	Diffusion mean ADC Largest Node
1	2	1	2	1	1	2	2	2	2
1	2	1	2	1	1	2	2	2	2
1	2	1	1	2	1	2	2	2	2
1	2	1	2	1	1	2	2	2	2
1	2	1	2	1	1	2	2	2	2
1	2	1	1	2	1	2	2	2	2
1	2	1	2	1	1	2	2	2	2
2	1	2	1	2	2	1	1	1	1
2	1	2	1	2	2	1	1	1	1
2	1	2	1	2	2	1	1	1	1
2	1	2	1	2	2	1	1	1	1
2	1	2	1	2	2	1	1	1	1
2	1	2	1	2	2	1	1	1	1
1	2	1	2	1	1	2	2	2	2
1	2	1	1	2	1	2	2	2	2
2	1	1	1	2	2	1	1	1	2
2	1	2	1	2	2	1	1	1	1

Tumor Volume (mL)	Largest Node Volume (mL)	SULmax Tumor	SULmedian Tumor	SULpeak Tumor	SULmax Largest Node	SULmedian node	SULpeak node	Diffusion Mean Tumor	Diffusion mean ADC Largest Node
2	1	2	1	2	2	1	1	1	1