**R Code**

**Uncovering and mitigating bias in large, automated MRI analyses of brain development**

Safia Elyounssi1,2\*, Keiko Kunitoki1,2\*, Jacqueline A. Clauss1,2, Eline Laurent1,2, Kristina A. Kane1,2, Dylan E. Hughes1,2,3, Casey E. Hopkinson1,2, Oren Bazer1,2, Rachel Freed Sussman1,2, Alysa E. Doyle1,4, Hang Lee5, Brenden Tervo-Clemmens1, Hamdi Eryilmaz1,2, Randy L. Gollub1,2, Deanna M. Barch6, Theodore D. Satterthwaite7,8,9, Kevin F. Dowling1,10, Joshua L. Roffman1,2

\*Equal authorship

1Department of Psychiatry, Massachusetts General Hospital and Harvard Medical School

2Martinos Center for Biomedical Imaging, Massachusetts General Hospital

3Departments of Psychiatry & Biobehavioral Sciences, University of California, Los Angeles

4Center for Genomic Medicine, Massachusetts General Hospital

5Biostatistics Center, Massachusetts General Hospital and Harvard Medical School

6Department of Psychological and Brain Sciences, Washington University in St. Louis

7Department of Psychiatry, University of Pennsylvania Perelman School of Medicine

8Penn Lifespan and Neuroimaging Center, University of Pennsylvania Perelman School of Medicine

9Penn-CHOP Lifespan Brain Institute

10Department of Psychiatry, University of Pittsburgh School of Medicine

Version date: October, 2024

Corresponding author:

Joshua L. Roffman MD, MMSc

Massachusetts General Hospital

149 13th St, Room 2616

Charlestown, MA 02129

617-724-1920

[jroffman@partners.org](mailto:jroffman@partners.org)

**########## COMPARING SMRI OUTPUTS BASED ON QC INCLUSION ##########**

.libPaths("C:/Users/keiko/Desktop/Rpackages")

library(lme4)

library(car)

library(dplyr)

library(ggplot2)

library(readr)

data <- read.csv("C:/Users/keiko/Desktop/MGH\_data/data2201.csv")

data14<-data %>% filter(RATE=="1"|RATE=="2"|RATE=="3"|RATE=="5")

data14<-data14%>%filter(EstimatedTotalIntraCranialVol!="NA")

ROIv<-data%>%select(contains('\_volume'))

ROIa<-data%>%select(contains('\_area'),-contains('White'))

ROIt<-data%>%select(contains('\_thickness'),-contains('Mean'))

ROIc<-data%>%select(contains('Left.'),contains('Right.'),-contains('.Vent'),contains('VentralDC'),-contains(".vessel"),-contains(".WM"),-contains(".choroid"))

ROIvn<-colnames(ROIv)

ROIan<-colnames(ROIa)

ROItn<-colnames(ROIt)

ROIcn<-colnames(ROIc)

for (i in 1:68){

data14[,paste('z.',ROIvn[i],sep="")]=scale(data14[,ROIvn[i]])

data14[,paste('z.',ROIan[i],sep="")]=scale(data14[,ROIan[i]])

data14[,paste('z.',ROItn[i],sep="")]=scale(data14[,ROItn[i]])

}

for (i in 1:20){

data14[,paste('z.',ROIcn[i],sep="")]=scale(data14[,ROIcn[i]])

}

data14$EstimatedTotalIntraCranialVol

head(data14$z.lh\_bankssts\_volume)

data14$newRATE<-data14$RATE

data14$newRATE[data14$newRATE=="5"]<-"4"

######

data1vs2<-data14%>%filter(newRATE=="1"|newRATE=="2")

data1vs3<-data14%>%filter(newRATE=="1"|newRATE=="3")

data1vs4<-data14%>%filter(newRATE=="1"|newRATE=="4")

### volume ###

### vol 1 vs 2

beta=c()

se=c()

p=c()

d=c()

res.df=data.frame()

for (i in 1:68){

model=lmer(data1vs2[,ROIvn[i]]~newRATE+z.age+Gender.0M.1F+z.eICV+(1|FAMILY\_ID)+(1|SCANNER)+(1|SITE\_ID),data1vs2)

p[i] =Anova(model)$'Pr(>Chisq)'[1]

se[i] =data.frame(coef(summary(model)))[2,2]

beta[i] = coef(summary(model))[2]

d[i]=coef(summary(model))[2]/sqrt(data.frame(VarCorr(model))[1,4]+data.frame(VarCorr(model))[2,4]+data.frame(VarCorr(model))[3,4]+data.frame(VarCorr(model))[4,4])

print(i)

}

res.df12v=data.frame("ROI"=ROIvn,"beta"=beta,"se"=se,"d"=d,"p"=p,"p.fdr"=p.adjust(p,method="fdr"))

write.csv(res.df12v,"C:/Users/keiko/Desktop/MGH\_data/p/compare1vs2\_vol.csv")

### vol 1 vs 3

beta=c()

se=c()

p=c()

d=c()

res.df=data.frame()

for (i in 1:68){

model=lmer(data1vs3[,ROIvn[i]]~newRATE+z.age+Gender.0M.1F+z.eICV+(1|FAMILY\_ID)+(1|SCANNER)+(1|SITE\_ID),data1vs3)

p[i] =Anova(model)$'Pr(>Chisq)'[1]

se[i] =data.frame(coef(summary(model)))[2,2]

beta[i] = coef(summary(model))[2]

d[i]=coef(summary(model))[2]/sqrt(data.frame(VarCorr(model))[1,4]+data.frame(VarCorr(model))[2,4]+data.frame(VarCorr(model))[3,4]+data.frame(VarCorr(model))[4,4])

print(i)

}

res.df13v=data.frame("ROI"=ROIvn,"beta"=beta,"se"=se,"d"=d, "p"=p,"p.fdr"=p.adjust(p,method="fdr"))

write.csv(res.df13v,"C:/Users/keiko/Desktop/MGH\_data/p/compare1vs3\_vol.csv")

### vol 1 vs 4

beta=c()

se=c()

p=c()

d=c()

res.df=data.frame()

for (i in 1:68){

model=lmer(data1vs4[,ROIvn[i]]~newRATE+z.age+Gender.0M.1F+z.eICV+(1|FAMILY\_ID)+(1|SCANNER)+(1|SITE\_ID),data1vs4)

p[i] =Anova(model)$'Pr(>Chisq)'[1]

se[i] =data.frame(coef(summary(model)))[2,2]

beta[i] = coef(summary(model))[2]

d[i]=coef(summary(model))[2]/sqrt(data.frame(VarCorr(model))[1,4]+data.frame(VarCorr(model))[2,4]+data.frame(VarCorr(model))[3,4]+data.frame(VarCorr(model))[4,4])

print(i)

}

res.df14v=data.frame("ROI"=ROIvn,"beta"=beta,"se"=se,"d"=d, "p"=p,"p.fdr"=p.adjust(p,method="fdr"))

write.csv(res.df14v,"C:/Users/keiko/Desktop/MGH\_data/p/compare1vs4\_vol.csv")

### area ###

### area 1 vs 2

beta=c()

se=c()

p=c()

d=c()

for (i in 1:68){

model=lmer(data1vs2[,ROIan[i]]~newRATE+z.age+Gender.0M.1F+z.eICV+(1|FAMILY\_ID)+(1|SCANNER)+(1|SITE\_ID),data1vs2)

p[i] =Anova(model)$'Pr(>Chisq)'[1]

se[i] =data.frame(coef(summary(model)))[2,2]

beta[i] = coef(summary(model))[2]

d[i]=coef(summary(model))[2]/sqrt(data.frame(VarCorr(model))[1,4]+data.frame(VarCorr(model))[2,4]+data.frame(VarCorr(model))[3,4]+data.frame(VarCorr(model))[4,4])

print(i)

}

res.df12a=data.frame("ROI"=ROIan,"beta"=beta,"se"=se,"d"=d, "p"=p,"p.fdr"=p.adjust(p,method="fdr"))

write.csv(res.df12a,"C:/Users/keiko/Desktop/MGH\_data/p/compare1vs2\_area.csv")

### area 1 vs 3

beta=c()

se=c()

p=c()

d=c()

for (i in 1:68){

model=lmer(data1vs3[,ROIan[i]]~newRATE+z.age+Gender.0M.1F+z.eICV+(1|FAMILY\_ID)+(1|SCANNER)+(1|SITE\_ID),data1vs3)

p[i] =Anova(model)$'Pr(>Chisq)'[1]

se[i] =data.frame(coef(summary(model)))[2,2]

beta[i] = coef(summary(model))[2]

d[i]=coef(summary(model))[2]/sqrt(data.frame(VarCorr(model))[1,4]+data.frame(VarCorr(model))[2,4]+data.frame(VarCorr(model))[3,4]+data.frame(VarCorr(model))[4,4])

print(i)

}

res.df13a=data.frame("ROI"=ROIan,"beta"=beta,"se"=se,"d"=d, "p"=p,"p.fdr"=p.adjust(p,method="fdr"))

write.csv(res.df13a,"C:/Users/keiko/Desktop/MGH\_data/p/compare1vs3\_area.csv")

### area 1 vs 4

beta=c()

se=c()

p=c()

d=c()

for (i in 1:68){

model=lmer(data1vs4[,ROIan[i]]~newRATE+z.age+Gender.0M.1F+z.eICV+(1|FAMILY\_ID)+(1|SCANNER)+(1|SITE\_ID),data1vs4)

p[i] =Anova(model)$'Pr(>Chisq)'[1]

se[i] =data.frame(coef(summary(model)))[2,2]

beta[i] = coef(summary(model))[2]

d[i]=coef(summary(model))[2]/sqrt(data.frame(VarCorr(model))[1,4]+data.frame(VarCorr(model))[2,4]+data.frame(VarCorr(model))[3,4]+data.frame(VarCorr(model))[4,4])

print(i)

}

res.df14a=data.frame("ROI"=ROIan,"beta"=beta,"se"=se,"d"=d, "p"=p,"p.fdr"=p.adjust(p,method="fdr"))

write.csv(res.df14a,"C:/Users/keiko/Desktop/MGH\_data/p/compare1vs4\_area.csv")

### thickness ###

### thick 1 vs 2

beta=c()

se=c()

p=c()

d=c()

for (i in 1:68){

model=lmer(data1vs2[,ROItn[i]]~newRATE+z.age+Gender.0M.1F+z.eICV+(1|FAMILY\_ID)+(1|SCANNER)+(1|SITE\_ID),data1vs2)

p[i] =Anova(model)$'Pr(>Chisq)'[1]

se[i] =data.frame(coef(summary(model)))[2,2]

beta[i] = coef(summary(model))[2]

d[i]=coef(summary(model))[2]/sqrt(data.frame(VarCorr(model))[1,4]+data.frame(VarCorr(model))[2,4]+data.frame(VarCorr(model))[3,4]+data.frame(VarCorr(model))[4,4])

print(i)

}

res.df12t=data.frame("ROI"=ROItn,"beta"=beta,"se"=se,"d"=d, "p"=p,"p.fdr"=p.adjust(p,method="fdr"))

write.csv(res.df12t,"C:/Users/keiko/Desktop/MGH\_data/p/compare1vs2\_thick.csv")

### thick 1 vs 3

beta=c()

se=c()

p=c()

d=c()

for (i in 1:68){

model=lmer(data1vs3[,ROItn[i]]~newRATE+z.age+Gender.0M.1F+z.eICV+(1|FAMILY\_ID)+(1|SCANNER)+(1|SITE\_ID),data1vs3)

p[i] =Anova(model)$'Pr(>Chisq)'[1]

se[i] =data.frame(coef(summary(model)))[2,2]

beta[i] = coef(summary(model))[2]

d[i]=coef(summary(model))[2]/sqrt(data.frame(VarCorr(model))[1,4]+data.frame(VarCorr(model))[2,4]+data.frame(VarCorr(model))[3,4]+data.frame(VarCorr(model))[4,4])

print(i)

}

res.df13t=data.frame("ROI"=ROItn,"beta"=beta,"se"=se,"d"=d, "p"=p,"p.fdr"=p.adjust(p,method="fdr"))

write.csv(res.df13t,"C:/Users/keiko/Desktop/MGH\_data/p/compare1vs3\_thick.csv")

### thick 1 vs 4

beta=c()

se=c()

p=c()

d=c()

for (i in 1:68){

model=lmer(data1vs4[,ROItn[i]]~newRATE+z.age+Gender.0M.1F+z.eICV+(1|FAMILY\_ID)+(1|SCANNER)+(1|SITE\_ID),data1vs4)

p[i] =Anova(model)$'Pr(>Chisq)'[1]

se[i] =data.frame(coef(summary(model)))[2,2]

beta[i] = coef(summary(model))[2]

d[i]=coef(summary(model))[2]/sqrt(data.frame(VarCorr(model))[1,4]+data.frame(VarCorr(model))[2,4]+data.frame(VarCorr(model))[3,4]+data.frame(VarCorr(model))[4,4])

print(i)

}

res.df14t=data.frame("ROI"=ROItn,"beta"=beta,"se"=se,"d"=d, "p"=p,"p.fdr"=p.adjust(p,method="fdr"))

write.csv(res.df14t,"C:/Users/keiko/Desktop/MGH\_data/p/compare1vs4\_thick.csv")

###

### subcortical ###

### sub.vol 1 vs 2

beta=c()

se=c()

p=c()

d=c()

res.df=data.frame()

for (i in 1:20){

model=lmer(data1vs2[,ROIcn[i]]~newRATE+z.age+Gender.0M.1F+z.eICV+(1|FAMILY\_ID)+(1|SCANNER)+(1|SITE\_ID),data1vs2)

p[i] =Anova(model)$'Pr(>Chisq)'[1]

se[i] =data.frame(coef(summary(model)))[2,2]

beta[i] = coef(summary(model))[2]

d[i]=coef(summary(model))[2]/sqrt(data.frame(VarCorr(model))[1,4]+data.frame(VarCorr(model))[2,4]+data.frame(VarCorr(model))[3,4]+data.frame(VarCorr(model))[4,4])

print(i)

}

res.df12c=data.frame("ROI"=ROIcn,"beta"=beta,"se"=se,"d"=d,"p"=p,"p.fdr"=p.adjust(p,method="fdr"))

write.csv(res.df12c,"C:/Users/keiko/Desktop/MGH\_data/p/compare1vs2\_subc.csv")

### vol 1 vs 3

beta=c()

se=c()

p=c()

d=c()

res.df=data.frame()

for (i in 1:20){

model=lmer(data1vs3[,ROIcn[i]]~newRATE+z.age+Gender.0M.1F+z.eICV+(1|FAMILY\_ID)+(1|SCANNER)+(1|SITE\_ID),data1vs3)

p[i] =Anova(model)$'Pr(>Chisq)'[1]

se[i] =data.frame(coef(summary(model)))[2,2]

beta[i] = coef(summary(model))[2]

d[i]=coef(summary(model))[2]/sqrt(data.frame(VarCorr(model))[1,4]+data.frame(VarCorr(model))[2,4]+data.frame(VarCorr(model))[3,4]+data.frame(VarCorr(model))[4,4])

print(i)

}

res.df13c=data.frame("ROI"=ROIcn,"beta"=beta,"se"=se,"d"=d, "p"=p,"p.fdr"=p.adjust(p,method="fdr"))

write.csv(res.df13c,"C:/Users/keiko/Desktop/MGH\_data/p/compare1vs3\_subc.csv")

### vol 1 vs 4

beta=c()

se=c()

p=c()

d=c()

res.df=data.frame()

for (i in 1:20){

model=lmer(data1vs4[,ROIcn[i]]~newRATE+z.age+Gender.0M.1F+z.eICV+(1|FAMILY\_ID)+(1|SCANNER)+(1|SITE\_ID),data1vs4)

p[i] =Anova(model)$'Pr(>Chisq)'[1]

se[i] =data.frame(coef(summary(model)))[2,2]

beta[i] = coef(summary(model))[2]

d[i]=coef(summary(model))[2]/sqrt(data.frame(VarCorr(model))[1,4]+data.frame(VarCorr(model))[2,4]+data.frame(VarCorr(model))[3,4]+data.frame(VarCorr(model))[4,4])

print(i)

}

res.df14c=data.frame("ROI"=ROIcn,"beta"=beta,"se"=se,"d"=d, "p"=p,"p.fdr"=p.adjust(p,method="fdr"))

write.csv(res.df14c,"C:/Users/keiko/Desktop/MGH\_data/p/compare1vs4\_subc.csv")

##

beta=c()

se=c()

p=c()

d=c()

for (i in 1:20){

model=lmer(data1vs2[,ROIcn[i]]~newRATE+z.age+Gender.0M.1F+(1|FAMILY\_ID)+(1|SCANNER)+(1|SITE\_ID),data1vs2)

p[i] =Anova(model)$'Pr(>Chisq)'[1]

se[i] =data.frame(coef(summary(model)))[2,2]

beta[i] = coef(summary(model))[2]

d[i]=coef(summary(model))[2]/sqrt(data.frame(VarCorr(model))[1,4]+data.frame(VarCorr(model))[2,4]+data.frame(VarCorr(model))[3,4]+data.frame(VarCorr(model))[4,4])

print(i)

}

res.df12ci=data.frame("ROI"=ROIcn,"beta"=beta,"se"=se,"d"=d,"p"=p,"p.fdr"=p.adjust(p,method="fdr"))

write.csv(res.df12c,"C:/Users/keiko/Desktop/MGH\_data/p/compare1vs2\_subc.csv")

##########################

### with Surface Holes ###

##########################

### volume ###

### vol 1 vs 2

beta=c()

se=c()

p=c()

d=c()

for (i in 1:68){

model=lmer(data1vs2[,ROIvn[i]]~newRATE+z.age+Gender.0M.1F+z.eICV+SurfaceHoles+(1|FAMILY\_ID)+(1|SCANNER)+(1|SITE\_ID),data1vs2)

p[i] =Anova(model)$'Pr(>Chisq)'[1]

se[i] =data.frame(coef(summary(model)))[2,2]

beta[i] = coef(summary(model))[2]

d[i]=coef(summary(model))[2]/sqrt(data.frame(VarCorr(model))[1,4]+data.frame(VarCorr(model))[2,4]+data.frame(VarCorr(model))[3,4]+data.frame(VarCorr(model))[4,4])

print(i)

}

res.dfs12v=data.frame("ROI"=ROIvn,"beta"=beta,"se"=se,"d"=d, "p"=p,"p.fdr"=p.adjust(p,method="fdr"))

write.csv(res.dfs12v,"C:/Users/keiko/Desktop/MGH\_data/p/compare\_s\_1vs2\_vol.csv")

### vol 1 vs 3

beta=c()

se=c()

p=c()

d=c()

res.dfs=data.frame()

for (i in 1:68){

model=lmer(data1vs3[,ROIvn[i]]~newRATE+z.age+Gender.0M.1F+z.eICV+SurfaceHoles+(1|FAMILY\_ID)+(1|SCANNER)+(1|SITE\_ID),data1vs3)

p[i] =Anova(model)$'Pr(>Chisq)'[1]

se[i] =data.frame(coef(summary(model)))[2,2]

beta[i] = coef(summary(model))[2]

d[i]=coef(summary(model))[2]/sqrt(data.frame(VarCorr(model))[1,4]+data.frame(VarCorr(model))[2,4]+data.frame(VarCorr(model))[3,4]+data.frame(VarCorr(model))[4,4])

print(i)

}

res.dfs13v=data.frame("ROI"=ROIvn,"beta"=beta,"se"=se,"d"=d, "p"=p,"p.fdr"=p.adjust(p,method="fdr"))

write.csv(res.dfs13v,"C:/Users/keiko/Desktop/MGH\_data/p/compare\_s\_1vs3\_vol.csv")

### vol 1 vs 4

beta=c()

se=c()

p=c()

d=c()

res.dfs=data.frame()

for (i in 1:68){

model=lmer(data1vs4[,ROIvn[i]]~newRATE+z.age+Gender.0M.1F+z.eICV+SurfaceHoles+(1|FAMILY\_ID)+(1|SCANNER)+(1|SITE\_ID),data1vs4)

p[i] =Anova(model)$'Pr(>Chisq)'[1]

se[i] =data.frame(coef(summary(model)))[2,2]

beta[i] = coef(summary(model))[2]

d[i]=coef(summary(model))[2]/sqrt(data.frame(VarCorr(model))[1,4]+data.frame(VarCorr(model))[2,4]+data.frame(VarCorr(model))[3,4]+data.frame(VarCorr(model))[4,4])

print(i)

}

res.dfs14v=data.frame("ROI"=ROIvn,"beta"=beta,"se"=se,"d"=d, "p"=p,"p.fdr"=p.adjust(p,method="fdr"))

write.csv(res.dfs14v,"C:/Users/keiko/Desktop/MGH\_data/p/compare\_s\_1vs4\_vol.csv")

### area ###

### area 1 vs 2

beta=c()

se=c()

p=c()

d=c()

res.dfs=data.frame()

for (i in 1:68){

model=lmer(data1vs2[,ROIan[i]]~newRATE+z.age+Gender.0M.1F+z.eICV+SurfaceHoles+(1|FAMILY\_ID)+(1|SCANNER)+(1|SITE\_ID),data1vs2)

p[i] =Anova(model)$'Pr(>Chisq)'[1]

se[i] =data.frame(coef(summary(model)))[2,2]

beta[i] = coef(summary(model))[2]

d[i]=coef(summary(model))[2]/sqrt(data.frame(VarCorr(model))[1,4]+data.frame(VarCorr(model))[2,4]+data.frame(VarCorr(model))[3,4]+data.frame(VarCorr(model))[4,4])

print(i)

}

res.dfs12a=data.frame("ROI"=ROIan,"beta"=beta,"se"=se,"d"=d, "p"=p,"p.fdr"=p.adjust(p,method="fdr"))

write.csv(res.dfs12a,"C:/Users/keiko/Desktop/MGH\_data/p/compare\_s\_1vs2\_area.csv")

### area 1 vs 3

beta=c()

se=c()

p=c()

d=c()

for (i in 1:68){

model=lmer(data1vs3[,ROIan[i]]~newRATE+z.age+Gender.0M.1F+z.eICV+SurfaceHoles+(1|FAMILY\_ID)+(1|SCANNER)+(1|SITE\_ID),data1vs3)

p[i] =Anova(model)$'Pr(>Chisq)'[1]

se[i] =data.frame(coef(summary(model)))[2,2]

beta[i] = coef(summary(model))[2]

d[i]=coef(summary(model))[2]/sqrt(data.frame(VarCorr(model))[1,4]+data.frame(VarCorr(model))[2,4]+data.frame(VarCorr(model))[3,4]+data.frame(VarCorr(model))[4,4])

print(i)

}

res.dfs13a=data.frame("ROI"=ROIan,"beta"=beta,"se"=se,"d"=d, "p"=p,"p.fdr"=p.adjust(p,method="fdr"))

write.csv(res.dfs13a,"C:/Users/keiko/Desktop/MGH\_data/p/compare\_s\_1vs3\_area.csv")

### area 1 vs 4

beta=c()

se=c()

p=c()

d=c()

for (i in 1:68){

model=lmer(data1vs4[,ROIan[i]]~newRATE+z.age+Gender.0M.1F+z.eICV+SurfaceHoles+(1|FAMILY\_ID)+(1|SCANNER)+(1|SITE\_ID),data1vs4)

p[i] =Anova(model)$'Pr(>Chisq)'[1]

se[i] =data.frame(coef(summary(model)))[2,2]

beta[i] = coef(summary(model))[2]

d[i]=coef(summary(model))[2]/sqrt(data.frame(VarCorr(model))[1,4]+data.frame(VarCorr(model))[2,4]+data.frame(VarCorr(model))[3,4]+data.frame(VarCorr(model))[4,4])

print(i)

}

res.dfs14a=data.frame("ROI"=ROIan,"beta"=beta,"se"=se, "d"=d,"p"=p,"p.fdr"=p.adjust(p,method="fdr"))

write.csv(res.dfs14a,"C:/Users/keiko/Desktop/MGH\_data/p/compare\_s\_1vs4\_area.csv")

### thickness ###

### thick 1 vs 2

beta=c()

se=c()

p=c()

d=c()

for (i in 1:68){

model=lmer(data1vs2[,ROItn[i]]~newRATE+z.age+Gender.0M.1F+z.eICV+SurfaceHoles+(1|FAMILY\_ID)+(1|SCANNER)+(1|SITE\_ID),data1vs2)

p[i] =Anova(model)$'Pr(>Chisq)'[1]

se[i] =data.frame(coef(summary(model)))[2,2]

beta[i] = coef(summary(model))[2]

d[i]=coef(summary(model))[2]/sqrt(data.frame(VarCorr(model))[1,4]+data.frame(VarCorr(model))[2,4]+data.frame(VarCorr(model))[3,4]+data.frame(VarCorr(model))[4,4])

print(i)

}

res.dfs12t=data.frame("ROI"=ROItn,"beta"=beta,"se"=se,"d"=d, "p"=p,"p.fdr"=p.adjust(p,method="fdr"))

write.csv(res.dfs12t,"C:/Users/keiko/Desktop/MGH\_data/p/compare\_s\_1vs2\_thick.csv")

### thickness 1 vs 3

beta=c()

se=c()

p=c()

d=c()

for (i in 1:68){

model=lmer(data1vs3[,ROItn[i]]~newRATE+z.age+Gender.0M.1F+z.eICV+SurfaceHoles+(1|FAMILY\_ID)+(1|SCANNER)+(1|SITE\_ID),data1vs3)

p[i] =Anova(model)$'Pr(>Chisq)'[1]

se[i] =data.frame(coef(summary(model)))[2,2]

beta[i] = coef(summary(model))[2]

d[i]=coef(summary(model))[2]/sqrt(data.frame(VarCorr(model))[1,4]+data.frame(VarCorr(model))[2,4]+data.frame(VarCorr(model))[3,4]+data.frame(VarCorr(model))[4,4])

print(i)

}

res.dfs13t=data.frame("ROI"=ROItn,"beta"=beta,"se"=se,"d"=d, "p"=p,"p.fdr"=p.adjust(p,method="fdr"))

write.csv(res.dfs13t,"C:/Users/keiko/Desktop/MGH\_data/p/compare\_s\_1vs3\_thick.csv")

### thickness 1 vs 4

beta=c()

se=c()

p=c()

d=c()

for (i in 1:68){

model=lmer(data1vs4[,ROItn[i]]~newRATE+z.age+Gender.0M.1F+z.eICV+SurfaceHoles+(1|FAMILY\_ID)+(1|SCANNER)+(1|SITE\_ID),data1vs4)

p[i] =Anova(model)$'Pr(>Chisq)'[1]

se[i] =data.frame(coef(summary(model)))[2,2]

beta[i] = coef(summary(model))[2]

d[i]=coef(summary(model))[2]/sqrt(data.frame(VarCorr(model))[1,4]+data.frame(VarCorr(model))[2,4]+data.frame(VarCorr(model))[3,4]+data.frame(VarCorr(model))[4,4])

print(i)

}

res.dfs14t=data.frame("ROI"=ROItn,"beta"=beta,"se"=se,"d"=d, "p"=p,"p.fdr"=p.adjust(p,method="fdr"))

write.csv(res.dfs14t,"C:/Users/keiko/Desktop/MGH\_data/p/compare\_s\_1vs4\_thick.csv")

#################

### cohen's d ###

#################

model=lmer(data1vs2[,ROIvn[1]]~newRATE+z.age+Gender.0M.1F+z.eICV+(1|FAMILY\_ID)+(1|SCANNER)+(1|SITE\_ID),data1vs2)

summary(model)

#

### volume ###

beta=c()

se=c()

p=c()

d=c()

beta2=c()

se2=c()

p2=c()

d2=c()

for (i in 1:68){

model=lmer(data14[,ROIvn[i]]~BL.CBCL\_EXTERNAL\_T+z.age+Gender.0M.1F+z.eICV+(1|FAMILY\_ID)+(1|SCANNER)+(1|SITE\_ID),data14)

p[i] =Anova(model)$'Pr(>Chisq)'[1]

se[i] =data.frame(coef(summary(model)))[2,2]

beta[i] = coef(summary(model))[2]

d[i]=coef(summary(model))[2]/sqrt(data.frame(VarCorr(model))[1,4]+data.frame(VarCorr(model))[2,4]+data.frame(VarCorr(model))[3,4]+data.frame(VarCorr(model))[4,4])

model2=lmer(data14[,ROIvn[i]]~BL.CBCL\_EXTERNAL\_T+newRATE+z.age+Gender.0M.1F+z.eICV+(1|FAMILY\_ID)+(1|SCANNER)+(1|SITE\_ID),data14)

p2[i] =Anova(model2)$'Pr(>Chisq)'[1]

se2[i] =data.frame(coef(summary(model2)))[2,2]

beta2[i] = coef(summary(model2))[2]

d2[i]=coef(summary(model2))[2]/sqrt(data.frame(VarCorr(model))[1,4]+data.frame(VarCorr(model))[2,4]+data.frame(VarCorr(model))[3,4]+data.frame(VarCorr(model))[4,4])

print(i)

}

comp.v.rate=data.frame("ROI"=ROIvn,"beta"=beta,"se"=se,"d"=d,"p"=p,"p.fdr"=p.adjust(p,method="fdr"),"beta2"=beta2,"se2"=se2,"d2"=d2,"p2"=p2,"p2.fdr"=p.adjust(p2,method="fdr"),"b.dif"=beta2-beta,"d.dif"=d2-d)

write.csv(comp.v.rate,"C:/Users/keiko/Desktop/MGH\_data/p/diff\_w\_wo\_rate\_vol.csv")

### subc.volume ###

beta=c()

se=c()

p=c()

d=c()

beta2=c()

se2=c()

p2=c()

d2=c()

for (i in 1:20){

model=lmer(data14[,ROIcn[i]]~BL.CBCL\_EXTERNAL\_T+z.age+Gender.0M.1F+z.eICV+(1|FAMILY\_ID)+(1|SCANNER)+(1|SITE\_ID),data14)

p[i] =Anova(model)$'Pr(>Chisq)'[1]

se[i] =data.frame(coef(summary(model)))[2,2]

beta[i] = coef(summary(model))[2]

d[i]=coef(summary(model))[2]/sqrt(data.frame(VarCorr(model))[1,4]+data.frame(VarCorr(model))[2,4]+data.frame(VarCorr(model))[3,4]+data.frame(VarCorr(model))[4,4])

model2=lmer(data14[,ROIcn[i]]~BL.CBCL\_EXTERNAL\_T+newRATE+z.age+Gender.0M.1F+z.eICV+(1|FAMILY\_ID)+(1|SCANNER)+(1|SITE\_ID),data14)

p2[i] =Anova(model2)$'Pr(>Chisq)'[1]

se2[i] =data.frame(coef(summary(model2)))[2,2]

beta2[i] = coef(summary(model2))[2]

d2[i]=coef(summary(model2))[2]/sqrt(data.frame(VarCorr(model))[1,4]+data.frame(VarCorr(model))[2,4]+data.frame(VarCorr(model))[3,4]+data.frame(VarCorr(model))[4,4])

print(i)

}

comp.c.rate=data.frame("ROI"=ROIcn,"beta"=beta,"se"=se,"d"=d,"p"=p,"p.fdr"=p.adjust(p,method="fdr"),"beta2"=beta2,"se2"=se2,"d2"=d2,"p2"=p2,"p2.fdr"=p.adjust(p2,method="fdr"),"b.dif"=beta2-beta,"d.dif"=d2-d)

write.csv(comp.c.rate,"C:/Users/keiko/Desktop/MGH\_data/p/diff\_w\_wo\_rate\_subc.csv")

**########## AUC ANALYSES FOR SHN CUTOFFS ##########**

.libPaths("C:/Users/keiko/Desktop/Rpackages")

library(dplyr)

library(ggplot2)

library(readr)

library(tidyr)

library(pROC)

##############

## baseline ##

##############

data <- read.csv("C:/Users/keiko/Desktop/MGH\_data/data2201.csv")

table(data$RATE)

#remove defective ones

data14<-data %>% filter(RATE=="1"|RATE=="2"|RATE=="3"|RATE=="5")

data14c<-data14%>%filter(SurfaceHoles!="NA")

data1c<-data14c %>% filter(RATE=="1")

data1<-data14 %>% filter(RATE=="1")

library("tableone")

vars.bl <- c("INTERVIEW\_AGE","Gender.0M.1F", "SCANNER","BL.CBCL\_INTERNAL\_T","BL.CBCL\_EXTERNAL\_T","BL.CBCL\_TOT\_T")

factorVars.bl<-c("Gender.0M.1F", "SCANNER")

ra <- CreateTableOne(vars = vars.bl,factorVars = factorVars.bl ,strata = c("RATE"), data =data14)

ra

tab\_csv <- print(ra,

#nonnormal = "INTERVIEW\_AGE",

printToggle = FALSE)

write.csv(tab\_csv,"C:/Users/keiko/Desktop/MGH\_data/p/bl\_tableone.csv")

data14\_f<-data14%>%filter(SCANNER!="NA")%>%filter(SurfaceHoles!="NA")

data14\_f%>% group\_by(RATE,SCANNER) %>%

summarize(mean.outcome = mean(SurfaceHoles))

data14\_f%>% group\_by(RATE,SCANNER) %>%

summarize(se.outcome = se(SurfaceHoles))

data14$newRATE<-data14$RATE

data14$newRATE[data14$newRATE=="5"]<-"4"

data14\_f<-data14%>%filter(SCANNER!="NA")%>%filter(SurfaceHoles!="NA")

data14\_f%>% group\_by(newRATE,SCANNER) %>%

summarize(mean.outcome = mean(SurfaceHoles))

se <- function(x) sqrt(var(x) / length(x))

data14\_f%>% group\_by(newRATE,SCANNER) %>%

summarize(se.outcome = se(SurfaceHoles))

data14\_f%>%ggplot(aes(x=SurfaceHoles,fill=factor(newRATE)))+geom\_histogram(binwidth=5)+

labs(x ="Surface Hole Number", y ="Count" , fill="MCQ")+ theme\_bw()

#+geom\_vline(xintercept=c(29.5,36.5,62.5))

plt\_bl<-data14\_f%>%ggplot(aes(x=SurfaceHoles,fill=factor(newRATE)))+geom\_density(alpha=0.6)+

labs(x ="Surface Hole Number", y ="Density" , fill="MQC")+ theme\_bw()

ggsave("C:/Users/keiko/Desktop/MGH\_data/img/bl\_mqc\_densityplot.tiff", plot = plt\_bl, width = 5, height = 6.5, units = "in", dpi = 300)

# 1 vs 234

data14$RATE12<-ifelse(data14$RATE=="1",1,0)

pROC\_obj <- roc(data14$RATE12, data14$SurfaceHoles,

smoothed = TRUE,

# arguments for ci

ci=TRUE, ci.alpha=0.9, stratified=FALSE,

# arguments for plot

plot=TRUE, auc.polygon=TRUE, max.auc.polygon=TRUE, grid=TRUE,

print.auc=F, show.thres=TRUE)

sens.ci <- ci.se(pROC\_obj)

plot(sens.ci, type="shape", col="lightblue")

coords(pROC\_obj, "best",best.method="youden",ret="all")

coords(pROC\_obj, "best",best.method="closest.topleft")

#

data14$RATE23<-ifelse(data14$RATE=="1"|data14$RATE=="2",1,0)

pROC\_obj <- roc(data14$RATE23, data14$SurfaceHoles,

smoothed = TRUE,

# arguments for ci

ci=TRUE, ci.alpha=0.9, stratified=FALSE,

# arguments for plot

plot=TRUE, auc.polygon=TRUE, max.auc.polygon=TRUE, grid=TRUE,

print.auc=F, show.thres=TRUE)

sens.ci <- ci.se(pROC\_obj)

plot(sens.ci, type="shape", col="lightblue")

coords(pROC\_obj, "best",best.method="youden",ret="all")

coords(pROC\_obj, "best",best.method="closest.topleft")

#

data14$RATE34<-ifelse(data14$RATE=="1"|data14$RATE=="2"|data14$RATE=="3",1,0)

pROC\_obj <- roc(data14$RATE34, data14$SurfaceHoles,

smoothed = TRUE,

# arguments for ci

ci=TRUE, ci.alpha=0.9, stratified=FALSE,

# arguments for plot

plot=TRUE, auc.polygon=TRUE, max.auc.polygon=TRUE, grid=TRUE,

print.auc=F, show.thres=TRUE)

sens.ci <- ci.se(pROC\_obj)

plot(sens.ci, type="shape", col="lightblue")

coords(pROC\_obj, "best",best.method="youden",ret="all")

coords(pROC\_obj, "best",best.method="closest.topleft")

#######################

### with revised 4s ###

#######################

rerate<-read.csv("C:/Users/keiko/Desktop/MGH\_data/QCrerate\_old4s.csv")

table(rerate$r\_rate)

redata<-data %>% left\_join(rerate,by="SUBJECTKEY")

redata15<-redata %>% filter(RATE=="1"|RATE=="2"|RATE=="3"|RATE=="4"|RATE=="5")

redata15<-redata15%>%mutate(newRATE=case\_when(redata15$RATE=="1" ~"1",

redata15$RATE=="2" ~"2",

redata15$RATE=="3" ~"3",

redata15$RATE=="4" ~"5",

redata15$RATE=="5" ~"4"))

table(redata15$r\_rate)

table(redata15$newRATE)

redata15$com\_rate<-ifelse(redata15$newRATE!="5",redata15$newRATE,redata15$r\_rate)

table(redata15$com\_rate)

redata5<-redata%>%filter(new\_rate=="5")

redata5$com\_rate

summary(redata5$SurfaceHoles)

redata$RATE

redata5$new\_rate

redata5$com\_rate

redata15%>%filter(com\_rate!="0")%>%ggplot(aes(x=SurfaceHoles,fill=factor(com\_rate)))+geom\_histogram(binwidth=5)+

labs(x ="Surface Hole Number", y ="Count" , fill="MCQ")+geom\_vline(xintercept=c(29.5,36.5,62.5))+ theme\_bw()

plt\_bl\_wsd<-redata15%>%filter(com\_rate!="0")%>%ggplot(aes(x=SurfaceHoles,fill=factor(com\_rate)))+geom\_density(alpha=.6)+

labs(x ="Surface Hole Number", y ="Density" , fill="MQC")+ theme\_bw()#+geom\_vline(xintercept=c(29.5,36.5,62.5))

ggsave("C:/Users/keiko/Desktop/MGH\_data/img/bl\_wsd\_mqc\_densityplot.tiff", plot = plt\_bl\_wsd, width = 5, height = 6.5, units = "in", dpi = 300)

redata15\_0<-redata15%>%filter(com\_rate!="0")

library("tableone")

vars.bl <- c("INTERVIEW\_AGE","Gender.0M.1F", "SCANNER","BL.CBCL\_INTERNAL\_T","BL.CBCL\_EXTERNAL\_T","BL.CBCL\_TOT\_T")

factorVars.bl<-c("Gender.0M.1F", "SCANNER")

ra <- CreateTableOne(vars = vars.bl,factorVars = factorVars.bl ,strata = c("com\_rate"), data =redata15\_0)

ra

tab\_csv <- print(ra,

#nonnormal = "INTERVIEW\_AGE",

printToggle = FALSE)

write.csv(tab\_csv,"C:/Users/keiko/Desktop/MGH\_data/p/bl\_tableone\_w\_sdo.csv")

###

m <- lm(INTERVIEW\_AGE ~ com\_rate, data = redata15\_0)

summary(m)

anova(m)

m <- lm(Gender.0M.1F ~ com\_rate, data = redata15\_0)

summary(m)

anova(m)

chisq.test(redata15\_0$SCANNER, redata15\_0$com\_rate, correct=FALSE)

m <- lm(BL.CBCL\_EXTERNAL\_T ~ com\_rate, data = redata15\_0)

summary(m)

anova(m)

m <- lm(BL.CBCL\_INTERNAL\_T ~ com\_rate, data = redata15\_0)

summary(m)

anova(m)

m <- lm(BL.CBCL\_TOT\_T ~ com\_rate, data = redata15\_0)

summary(m)

anova(m)

############

# 1 vs 234

redata15\_0$RATE12<-ifelse(redata15\_0$com\_rate=="1",1,0)

pROC\_obj <- roc(redata15\_0$RATE12, redata15\_0$SurfaceHoles,

smoothed = TRUE,

# arguments for ci

ci=TRUE, ci.alpha=0.9, stratified=FALSE,

# arguments for plot

plot=TRUE, auc.polygon=TRUE, max.auc.polygon=TRUE, grid=TRUE,

print.auc=T, show.thres=TRUE)

sens.ci <- ci.se(pROC\_obj)

plot(sens.ci, type="shape", col="lightblue")

coords(pROC\_obj, "best",best.method="youden",ret="all")

coords(pROC\_obj, "best",best.method="closest.topleft")

#

redata15\_0$RATE23<-ifelse(redata15\_0$com\_rate=="1"|redata15\_0$com\_rate=="2",1,0)

pROC\_obj <- roc(redata15\_0$RATE23, redata15\_0$SurfaceHoles,

smoothed = TRUE,

# arguments for ci

ci=TRUE, ci.alpha=0.9, stratified=FALSE,

# arguments for plot

plot=TRUE, auc.polygon=TRUE, max.auc.polygon=TRUE, grid=TRUE,

print.auc=T, show.thres=TRUE)

sens.ci <- ci.se(pROC\_obj)

plot(sens.ci, type="shape", col="lightblue")

coords(pROC\_obj, "best",best.method="youden",ret="all")

coords(pROC\_obj, "best",best.method="closest.topleft")

#

redata15\_0$RATE34<-ifelse(redata15\_0$com\_rate=="1"|redata15\_0$com\_rate=="2"|redata15\_0$com\_rate=="3",1,0)

pROC\_obj <- roc(redata15\_0$RATE34, redata15\_0$SurfaceHoles,

smoothed = TRUE,

# arguments for ci

ci=TRUE, ci.alpha=0.9, stratified=FALSE,

# arguments for plot

plot=TRUE, auc.polygon=TRUE, max.auc.polygon=TRUE, grid=TRUE,

print.auc=F, show.thres=TRUE)

sens.ci <- ci.se(pROC\_obj)

plot(sens.ci, type="shape", col="lightblue")

coords(pROC\_obj, "best",best.method="youden",ret="all")

coords(pROC\_obj, "best",best.method="closest.topleft")

##############

### year 2 ###

##############

rate\_y2<- read.csv("C:/Users/keiko/Desktop/MGH\_data/rate/EK\_y2.csv")

aseg\_y2 <- read.table("C:/Users/keiko/Desktop/MGH\_data/rate/aseg\_stats\_y2.txt",header=T)

data\_bl <- read.csv("C:/Users/keiko/Desktop/MGH\_data/data2201.csv")

scan\_bl<-data\_bl%>% select("SUBJECTKEY",SCANNER)

aseg\_y2$key<-substr(aseg\_y2$Measure.volume,10,20)

aseg\_y2$SUBJECTKEY<-paste("NDAR\_",aseg\_y2$key,sep="")

aseg\_y2\_f<-aseg\_y2%>%select(SUBJECTKEY,SurfaceHoles)

rate\_y2$rater<-substr(rate\_y2$ï..id,1,1)

data\_y2<-rate\_y2%>%left\_join(aseg\_y2\_f,by="SUBJECTKEY")%>%left\_join(scan\_bl,by="SUBJECTKEY")

data\_y2<-data\_y2%>%filter(SUBJECTKEY!="")

data\_y2%>% ggplot(aes(x=SurfaceHoles,fill=factor(rating)))+geom\_histogram(binwidth = 10)+geom\_vline(xintercept=c(29.5,36.5,62.5))+ theme\_bw()

K<-data\_y2%>%filter(rater=="K")

E<-data\_y2%>%filter(rater=="E")

data\_y2%>%filter(SurfaceHoles!="NA")%>%filter(rating!="0")%>%group\_by(rating)%>%

summarize(mean = mean(SurfaceHoles),

median = median(SurfaceHoles),

sd = sd(SurfaceHoles),

min=min(SurfaceHoles),

q1 = quantile(SurfaceHoles, 0.25),

q3 = quantile(SurfaceHoles, 0.75),

max=max(SurfaceHoles))

data\_y2%>%filter(rating!="0")%>% ggplot(aes(x=SurfaceHoles,fill=factor(rating)))+geom\_histogram(binwidth = 10)+geom\_vline(xintercept=c(29.5,36.5,62.5))+facet\_wrap(vars(rater),nrow=2)

data\_y2%>%filter(rating!="0")%>% ggplot(aes(x=SurfaceHoles,fill=factor(rating)))+geom\_histogram(binwidth = 10)+geom\_vline(xintercept=c(29.5,36.5,62.5))+ theme\_bw()

data\_y2%>%filter(rating!="0")%>% ggplot(aes(x=SurfaceHoles,fill=factor(rating)))+geom\_density(binwidth = 10,alpha=0.6)+geom\_vline(xintercept=c(29.5,36.5,62.5))+ theme\_bw()

data\_y2%>%filter(rating!="0")%>%ggplot(aes(x=SurfaceHoles,fill=factor(rating)))+geom\_histogram(binwidth=5)+geom\_vline(xintercept=c(29.5,36.5,62.5))+

labs(x ="Surface Hole Number", y ="Count" , fill="MCQ")

plt<-data\_y2%>%filter(rating!="0")%>%ggplot(aes(x=SurfaceHoles,fill=factor(rating)))+geom\_density(alpha=0.6)+#geom\_vline(xintercept=c(29.5,36.5,62.5))+

labs(x ="Surface Hole Number", y ="Density" , fill="MQC")+ theme\_bw()

ggsave("C:/Users/keiko/Desktop/MGH\_data/img/y2\_mqc\_densityplot.tiff", plot = plt, width = 5, height = 6.5, units = "in", dpi = 300)

data\_y2<-data\_y2%>%mutate(SHgroup\_y2=case\_when(SurfaceHoles<29.5 ~"1",

(SurfaceHoles<36.5&SurfaceHoles>29.5) ~"2",

(SurfaceHoles<62.5 &SurfaceHoles>36.5) ~"3",

SurfaceHoles>62.5 ~"4"))

table(data\_y2$rating,data\_y2$SHgroup)

data\_y2%>%filter(rating!="0")%>% ggplot(aes(x=SurfaceHoles,fill=factor(SHgroup\_y2)))+geom\_histogram(binwidth = 10)+geom\_vline(xintercept=c(29.5,36.5,62.5))+facet\_wrap(vars(rater),nrow=2)

data\_y2%>%filter(rating!="0")%>% ggplot(aes(x=SurfaceHoles,fill=factor(SHgroup)))+geom\_histogram(binwidth = 10)+geom\_vline(xintercept=c(29.5,36.5,62.5))

data\_y2%>%filter(rating!="0")%>% ggplot(aes(x=SurfaceHoles,fill=factor(rating)))+geom\_histogram(binwidth = 10)+facet\_wrap(vars(SHgroup\_y2),nrow=2)

### tableone ###

CBCL<-read.csv("C:/Users/keiko/Desktop/MGH\_data/lcbcl\_y1\_3summary.csv")

scanner<-read.csv("C:/Users/keiko/Desktop/MGH\_data/abcd\_mri01.csv")

scanner$SUBJECTKEY<-scanner$subjectkey

scanner\_y2<-scanner%>%filter(eventname=="2\_year\_follow\_up\_y\_arm\_1")%>%dplyr::select("SUBJECTKEY","mri\_info\_manufacturer","interview\_age","sex" )

CBCL\_y2<-CBCL%>%dplyr::select(contains(".y2"),"SUBJECTKEY")

tb1\_y2<-data\_y2%>%left\_join(scanner\_y2,by="SUBJECTKEY")%>%left\_join(CBCL\_y2,by="SUBJECTKEY")

tb1\_y2$scannermatch<-ifelse(tb1\_y2$SCANNER==tb1\_y2$mri\_info\_manufacturer,"1","0")

table(tb1\_y2$scannermatch)

view(tb1\_y2[tb1\_y2$scannermatch=="0",])

tb1\_y2\_n0<-tb1\_y2%>%filter(rating!="0")

library("tableone")

vars <- c("interview\_age","sex", "mri\_info\_manufacturer","cbcl\_syn\_internal\_t.y2","cbcl\_syn\_external\_t.y2","cbcl\_syn\_totprob\_t.y2","rating")

factorVars<-c("sex", "mri\_info\_manufacturer","rating")

ra <- CreateTableOne(vars = vars,factorVars = factorVars ,strata = c("SHgroup\_y2"), data =tb1\_y2\_n0)

ra

tab\_csv <- print(ra,

nonnormal = "Age",

printToggle = FALSE)

write.csv(tab\_csv,"C:/Users/keiko/Desktop/MGH\_data/p/rate\_y2\_tableone.csv")

###

m <- lm(interview\_age ~ SHgroup\_y2, data = tb1\_y2\_n0)

summary(m)

anova(m)

tb1\_y2\_n0$sex<-as.numeric(tb1\_y2\_n0$sex)

m <- lm(sex ~ SHgroup\_y2, data = tb1\_y2\_n0)

summary(m)

anova(m)

chisq.test(tb1\_y2\_n0$SCANNER, tb1\_y2\_n0$SHgroup\_y2, correct=FALSE)

m <- lm(cbcl\_syn\_internal\_t.y2 ~ SHgroup\_y2, data = tb1\_y2\_n0)

summary(m)

anova(m)

m <- lm(cbcl\_syn\_external\_t.y2 ~ SHgroup\_y2, data = tb1\_y2\_n0)

summary(m)

anova(m)

m <- lm(cbcl\_syn\_totprob\_t.y2 ~ SHgroup\_y2, data = tb1\_y2\_n0)

summary(m)

anova(m)

chisq.test(tb1\_y2\_n0$rating, tb1\_y2\_n0$SHgroup\_y2, correct=FALSE)

**########## APPLIED ANALYSES: EFFECT OF CORTICAL VOLUME ON CBCL EXTERNALIZING, TESTING WITH VARIOUS QC THRESHOLDS ##########**

**##### NB: EFFECT OF AGE ON CORTICAL VOLUME FOLLOWS SAME MODEL**

library(readr)

library(writexl)

library(dplyr)

library(tidyr)

library(readxl)

library(lme4)

library(car)

## ORGANIZING DATA

input\_data <- read\_xlsx("/Users/sofiaperdomo/Desktop/ABCD/QC\_Paper/QC\_R1\_CBCL\_102024.xlsx")

input\_data <- input\_data %>%

mutate(

INTERVIEW\_AGE = z\_score(INTERVIEW\_AGE),

SHN = z\_score(SHN),

ICV = z\_score(ICV)

)

input\_data$SHN\_Tiers <- as.numeric(input\_data$SHN\_Tiers)

Data\_MQC\_1 <- input\_data %>% filter(MQC == 1)

Data\_MQC\_2 <- input\_data %>% filter(MQC == 2)

Data\_MQC\_2\_3 <- input\_data %>% filter(MQC %in% c(2, 3))

Data\_MQC\_2\_3\_4 <- input\_data %>% filter(MQC %in% c(2, 3, 4))

Data\_SHN\_Tier\_1 <- input\_data %>% filter(SHN\_Tiers == 1)

Data\_SHN\_Tier\_1\_2 <- input\_data %>% filter(SHN\_Tiers %in% c(1, 2))

Data\_SHN\_Tier\_1\_2\_3 <- input\_data %>% filter(SHN\_Tiers %in% c(1, 2, 3))

Data\_SHN\_Tier\_1\_2\_3\_4 <- input\_data %>% filter(SHN\_Tiers %in% c(1, 2, 3, 4))

print(length(unique(Data\_MQC\_1$SUBJECTKEY)))

print(length(unique(Data\_MQC\_2$SUBJECTKEY)))

print(length(unique(Data\_MQC\_2\_3$SUBJECTKEY)))

print(length(unique(Data\_MQC\_2\_3\_4$SUBJECTKEY)))

print(length(unique(Data\_SHN\_Tier\_1$SUBJECTKEY)))

print(length(unique(Data\_SHN\_Tier\_1\_2$SUBJECTKEY)))

print(length(unique(Data\_SHN\_Tier\_1\_2\_3$SUBJECTKEY)))

print(length(unique(Data\_SHN\_Tier\_1\_2\_3\_4$SUBJECTKEY)))

# Function to calculate p-value from t-value

calculate\_pvalue <- function(t\_value, df) {

2 \* pt(abs(t\_value), df, lower.tail = FALSE)

}

# Manual Z-Score Standardization

z\_score <- function(x) {

(x - mean(x, na.rm = TRUE)) / sd(x, na.rm = TRUE)

}

## LINEAR MIXED MODEL FITTING

roi\_columns <- grep("volume", names(input\_data), value = TRUE)

# Data frames for results

results\_MQC\_1 <- data.frame(ROI = character(), beta\_volume = numeric(), pvalue\_volume = numeric(),

SE\_volume = numeric(), stringsAsFactors = FALSE)

results\_MQC\_2 <- data.frame(ROI = character(), beta\_volume = numeric(), pvalue\_volume = numeric(),

SE\_volume = numeric(), stringsAsFactors = FALSE)

results\_MQC\_2\_3 <- data.frame(ROI = character(), beta\_volume = numeric(), pvalue\_volume = numeric(),

SE\_volume = numeric(), stringsAsFactors = FALSE)

results\_MQC\_2\_3\_4 <- data.frame(ROI = character(), beta\_volume = numeric(), pvalue\_volume = numeric(),

SE\_volume = numeric(), stringsAsFactors = FALSE)

results\_SHN\_Tier\_1 <- data.frame(ROI = character(), beta\_volume = numeric(), pvalue\_volume = numeric(),

SE\_volume = numeric(), stringsAsFactors = FALSE)

results\_SHN\_Tier\_1\_2 <- data.frame(ROI = character(), beta\_volume = numeric(), pvalue\_volume = numeric(),

SE\_volume = numeric(), stringsAsFactors = FALSE)

results\_SHN\_Tier\_1\_2\_3 <- data.frame(ROI = character(), beta\_volume = numeric(), pvalue\_volume = numeric(),

SE\_volume = numeric(), stringsAsFactors = FALSE)

results\_SHN\_Tier\_1\_2\_3\_4 <- data.frame(ROI = character(), beta\_volume = numeric(), pvalue\_volume = numeric(),

SE\_volume = numeric(), stringsAsFactors = FALSE)

results\_SHN\_Tier\_1\_2\_3\_4\_woSHNcov <- data.frame(ROI = character(), beta\_volume = numeric(), pvalue\_volume = numeric(),

SE\_volume = numeric(), stringsAsFactors = FALSE)

# Loop through each ROI column

for (roi in roi\_columns) {

# Ground Analysis for MQC 1

cat("Analyzing ROI:", roi, "MQC 1 subjects\n")

model\_MQC\_1 <- lmer(as.formula(paste("ZBL.CBCL\_EXTERNAL\_T ~ INTERVIEW\_AGE + Sex.0M.1F + SHN + ICV +", roi, "+

(1 | SITE\_ID) + (1 | FAMILY\_ID) + (1 | SCANNER)")),

data = Data\_MQC\_1)

summary\_MQC\_1 <- summary(model\_MQC\_1)$coefficients

beta\_volume <- summary\_MQC\_1[roi, "Estimate"]

t\_value\_volume <- summary\_MQC\_1[roi, "t value"]

pvalue\_volume <- calculate\_pvalue(t\_value\_volume, model\_MQC\_1@devcomp$dims[["nmp"]])

SE\_volume <- summary\_MQC\_1[roi, "Std. Error"]

results\_MQC\_1 <- rbind(results\_MQC\_1, data.frame(ROI = roi, beta\_volume, pvalue\_volume, SE\_volume))

print(summary(model\_MQC\_1))

print(Anova(model\_MQC\_1))

# Analysis 1 for MQC 2

cat("Analyzing ROI:", roi, "MQC 2 subjects\n")

model\_MQC\_2 <- lmer(as.formula(paste("ZBL.CBCL\_EXTERNAL\_T ~ INTERVIEW\_AGE + Sex.0M.1F + SHN + ICV +", roi, "+

(1 | SITE\_ID) + (1 | FAMILY\_ID) + (1 | SCANNER)")),

data = Data\_MQC\_2)

summary\_MQC\_2 <- summary(model\_MQC\_2)$coefficients

beta\_volume <- summary\_MQC\_2[roi, "Estimate"]

t\_value\_volume <- summary\_MQC\_2[roi, "t value"]

pvalue\_volume <- calculate\_pvalue(t\_value\_volume, model\_MQC\_2@devcomp$dims[["nmp"]])

SE\_volume <- summary\_MQC\_2[roi, "Std. Error"]

results\_MQC\_2 <- rbind(results\_MQC\_2, data.frame(ROI = roi, beta\_volume, pvalue\_volume, SE\_volume))

print(summary(model\_MQC\_2))

print(Anova(model\_MQC\_2))

# Analysis 2 for MQC 2, 3

cat("Analyzing ROI:", roi, "MQC 2, 3 subjects\n")

model\_MQC\_2\_3 <- lmer(as.formula(paste("ZBL.CBCL\_EXTERNAL\_T ~ INTERVIEW\_AGE + Sex.0M.1F + SHN + ICV +", roi, "+ (1 | SITE\_ID) + (1 | FAMILY\_ID) + (1 | SCANNER)")),

data = Data\_MQC\_2\_3)

summary\_MQC\_2\_3 <- summary(model\_MQC\_2\_3)$coefficients

beta\_volume <- summary\_MQC\_2\_3[roi, "Estimate"]

t\_value\_volume <- summary\_MQC\_2\_3[roi, "t value"]

pvalue\_volume <- calculate\_pvalue(t\_value\_volume, model\_MQC\_2\_3@devcomp$dims[["nmp"]])

SE\_volume <- summary\_MQC\_2\_3[roi, "Std. Error"]

results\_MQC\_2\_3 <- rbind(results\_MQC\_2\_3, data.frame(ROI = roi, beta\_volume, pvalue\_volume, SE\_volume))

print(summary(model\_MQC\_2\_3))

print(Anova(model\_MQC\_2\_3))

# Analysis 3 for MQC 2, 3, 4

cat("Analyzing ROI:", roi, "MQC 2, 3, 4 subjects\n")

model\_MQC\_2\_3\_4 <- lmer(as.formula(paste("ZBL.CBCL\_EXTERNAL\_T ~ INTERVIEW\_AGE + Sex.0M.1F + SHN + ICV +", roi,

"+ (1 | SITE\_ID) + (1 | FAMILY\_ID) + (1 | SCANNER)")),

data = Data\_MQC\_2\_3\_4)

summary\_MQC\_2\_3\_4 <- summary(model\_MQC\_2\_3\_4)$coefficients

beta\_volume <- summary\_MQC\_2\_3\_4[roi, "Estimate"]

t\_value\_volume <- summary\_MQC\_2\_3\_4[roi, "t value"]

pvalue\_volume <- calculate\_pvalue(t\_value\_volume, model\_MQC\_2\_3\_4@devcomp$dims[["nmp"]])

SE\_volume <- summary\_MQC\_2\_3\_4[roi, "Std. Error"]

results\_MQC\_2\_3\_4 <- rbind(results\_MQC\_2\_3\_4, data.frame(ROI = roi, beta\_volume, pvalue\_volume, SE\_volume))

print(summary(model\_MQC\_2\_3\_4))

print(Anova(model\_MQC\_2\_3\_4))

# Analysis 4 for SHN Tier 1

cat("Analyzing ROI:", roi, "SHN Tier 1 subjects\n")

model\_SHN\_Tier\_1 <- lmer(as.formula(paste("ZBL.CBCL\_EXTERNAL\_T ~ INTERVIEW\_AGE + Sex.0M.1F + SHN + ICV +", roi,

"+ (1 | SITE\_ID) + (1 | FAMILY\_ID) + (1 | SCANNER)")),

data = Data\_SHN\_Tier\_1)

summary\_SHN\_Tier\_1 <- summary(model\_SHN\_Tier\_1)$coefficients

beta\_volume <- summary\_SHN\_Tier\_1[roi, "Estimate"]

t\_value\_volume <- summary\_SHN\_Tier\_1[roi, "t value"]

pvalue\_volume <- calculate\_pvalue(t\_value\_volume, model\_SHN\_Tier\_1@devcomp$dims[["nmp"]])

SE\_volume <- summary\_SHN\_Tier\_1[roi, "Std. Error"]

results\_SHN\_Tier\_1 <- rbind(results\_SHN\_Tier\_1, data.frame(ROI = roi, beta\_volume, pvalue\_volume, SE\_volume))

print(summary(model\_SHN\_Tier\_1))

print(Anova(model\_SHN\_Tier\_1))

# Analysis 5 for SHN Tier 1, 2

cat("Analyzing ROI:", roi, "SHN Tier 1, 2 subjects\n")

model\_SHN\_Tier\_1\_2 <- lmer(as.formula(paste("ZBL.CBCL\_EXTERNAL\_T ~ INTERVIEW\_AGE + Sex.0M.1F + SHN + ICV +", roi,

"+ (1 | SITE\_ID) + (1 | FAMILY\_ID) + (1 | SCANNER)")),

data = Data\_SHN\_Tier\_1\_2)

summary\_SHN\_Tier\_1\_2 <- summary(model\_SHN\_Tier\_1\_2)$coefficients

beta\_volume <- summary\_SHN\_Tier\_1\_2[roi, "Estimate"]

t\_value\_volume <- summary\_SHN\_Tier\_1\_2[roi, "t value"]

pvalue\_volume <- calculate\_pvalue(t\_value\_volume, model\_SHN\_Tier\_1\_2@devcomp$dims[["nmp"]])

SE\_volume <- summary\_SHN\_Tier\_1\_2[roi, "Std. Error"]

results\_SHN\_Tier\_1\_2 <- rbind(results\_SHN\_Tier\_1\_2, data.frame(ROI = roi, beta\_volume, pvalue\_volume, SE\_volume))

print(summary(model\_SHN\_Tier\_1\_2))

print(Anova(model\_SHN\_Tier\_1\_2))

# Analysis 6 for SHN Tier 1, 2, 3

cat("Analyzing ROI:", roi, "SHN Tier 1, 2, 3 subjects\n")

model\_SHN\_Tier\_1\_2\_3 <- lmer(as.formula(paste("ZBL.CBCL\_EXTERNAL\_T ~ INTERVIEW\_AGE + Sex.0M.1F + SHN + ICV +", roi,

"+ (1 | SITE\_ID) + (1 | FAMILY\_ID) + (1 | SCANNER)")),

data = Data\_SHN\_Tier\_1\_2\_3)

summary\_SHN\_Tier\_1\_2\_3 <- summary(model\_SHN\_Tier\_1\_2\_3)$coefficients

beta\_volume <- summary\_SHN\_Tier\_1\_2\_3[roi, "Estimate"]

t\_value\_volume <- summary\_SHN\_Tier\_1\_2\_3[roi, "t value"]

pvalue\_volume <- calculate\_pvalue(t\_value\_volume, model\_SHN\_Tier\_1\_2\_3@devcomp$dims[["nmp"]])

SE\_volume <- summary\_SHN\_Tier\_1\_2\_3[roi, "Std. Error"]

results\_SHN\_Tier\_1\_2\_3 <- rbind(results\_SHN\_Tier\_1\_2\_3, data.frame(ROI = roi, beta\_volume, pvalue\_volume, SE\_volume))

print(summary(model\_SHN\_Tier\_1\_2\_3))

print(Anova(model\_SHN\_Tier\_1\_2\_3))

# Analysis 7 for SHN Tier 1, 2, 3, 4

cat("Analyzing ROI:", roi, "SHN Tier 1, 2, 3, 4 subjects\n")

model\_SHN\_Tier\_1\_2\_3\_4 <- lmer(as.formula(paste("ZBL.CBCL\_EXTERNAL\_T ~ INTERVIEW\_AGE + Sex.0M.1F + SHN + ICV +", roi,

"+ (1 | SITE\_ID) + (1 | FAMILY\_ID) + (1 | SCANNER)")),

data = Data\_SHN\_Tier\_1\_2\_3\_4)

summary\_SHN\_Tier\_1\_2\_3\_4 <- summary(model\_SHN\_Tier\_1\_2\_3\_4)$coefficients

beta\_volume <- summary\_SHN\_Tier\_1\_2\_3\_4[roi, "Estimate"]

t\_value\_volume <- summary\_SHN\_Tier\_1\_2\_3\_4[roi, "t value"]

pvalue\_volume <- calculate\_pvalue(t\_value\_volume, model\_SHN\_Tier\_1\_2\_3\_4@devcomp$dims[["nmp"]])

SE\_volume <- summary\_SHN\_Tier\_1\_2\_3\_4[roi, "Std. Error"]

results\_SHN\_Tier\_1\_2\_3\_4 <- rbind(results\_SHN\_Tier\_1\_2\_3\_4, data.frame(ROI = roi, beta\_volume, pvalue\_volume, SE\_volume))

print(summary(model\_SHN\_Tier\_1\_2\_3\_4))

print(Anova(model\_SHN\_Tier\_1\_2\_3\_4))

# Analysis 8 for SHN Tier 1, 2, 3, 4 ... wo SHN as covariate

cat("Analyzing ROI:", roi, "SHN Tier 1, 2, 3, 4 subjects wo SHN covariate\n")

model\_SHN\_Tier\_1\_2\_3\_4\_woSHNcov <- lmer(as.formula(paste("ZBL.CBCL\_EXTERNAL\_T ~ INTERVIEW\_AGE + Sex.0M.1F + ICV +", roi,

"+ (1 | SITE\_ID) + (1 | FAMILY\_ID) + (1 | SCANNER)")),

data = Data\_SHN\_Tier\_1\_2\_3\_4)

summary\_SHN\_Tier\_1\_2\_3\_4\_woSHNcov <- summary(model\_SHN\_Tier\_1\_2\_3\_4\_woSHNcov)$coefficients

beta\_volume <- summary\_SHN\_Tier\_1\_2\_3\_4\_woSHNcov[roi, "Estimate"]

t\_value\_volume <- summary\_SHN\_Tier\_1\_2\_3\_4\_woSHNcov[roi, "t value"]

pvalue\_volume <- calculate\_pvalue(t\_value\_volume, model\_SHN\_Tier\_1\_2\_3\_4\_woSHNcov@devcomp$dims[["nmp"]])

SE\_volume <- summary\_SHN\_Tier\_1\_2\_3\_4\_woSHNcov[roi, "Std. Error"]

results\_SHN\_Tier\_1\_2\_3\_4\_woSHNcov <- rbind(results\_SHN\_Tier\_1\_2\_3\_4\_woSHNcov, data.frame(ROI = roi, beta\_volume, pvalue\_volume, SE\_volume))

print(summary(model\_SHN\_Tier\_1\_2\_3\_4\_woSHNcov))

print(Anova(model\_SHN\_Tier\_1\_2\_3\_4\_woSHNcov))

}

write\_xlsx(list(

MQC\_1 = results\_MQC\_1,

MQC\_2 = results\_MQC\_2,

MQC\_2\_3 = results\_MQC\_2\_3,

MQC\_2\_3\_4 = results\_MQC\_2\_3\_4,

SHN\_Tier\_1 = results\_SHN\_Tier\_1,

SHN\_Tier\_1\_2 = results\_SHN\_Tier\_1\_2,

SHN\_Tier\_1\_2\_3 = results\_SHN\_Tier\_1\_2\_3,

SHN\_Tier\_1\_2\_3\_4 = results\_SHN\_Tier\_1\_2\_3\_4,

SHN\_Tier\_1\_2\_3\_4\_woSHNcov = results\_SHN\_Tier\_1\_2\_3\_4\_woSHNcov

), "/Users/sofiaperdomo/Desktop/ABCD/QC\_Paper/QC\_SHN\_Tier\_LMM\_Analyses.xlsx")