Breast cancer risk stratification using genetic and non-genetic risk assessment tools for 246,142 women in the UK Biobank.

hopj

2022-08-30

PREPARING DATA

Initial selection of genetic confirmed females

```
library(stringr)
bd <- read.table("pheno/ukb668808.tab",sep="\t",header=T)
bd <- bd[!is.na(bd$f.22001.0.0) & bd$f.22001.0.0==0,]  # select genetic females

bd$sex.self <- bd$f.31.0.0
bd$sex.genetic <- bd$f.22001.0.0
bd$sex.aneuploidy <- as.character(bd$f.22019.0.0)
bd$sex.aneuploidy[is.na(bd$sex.aneuploidy)] <- "No"

table(bd$sex.self,bd$sex.genetic)
table(bd$sex.aneuploidy,bd$sex.genetic)

data <- data0 <- bd[bd$sex.genetic=="Female",]

saveRDS(data,paste0("data/data_females_n",nrow(data),"_",Sys.Date(),".rds"))

data.select <- data[,c("f.eid",colnames(data)[!str_starts(colnames(data),"f.")])]
saveRDS(data.select,paste0("data/females_n",nrow(data.select),"_",Sys.Date(),".rds"))
```

Recoding

```
library(stringr)
setwd("/Volumes/research/HG/HG7/Private/Datasets/UK Biobank/project/Overlap")

data.temp <- readRDS("data/data_females_n264741_2022-09-20.rds")
dim(data.temp)
colnames(data.temp)[colnames(data.temp)=="f.eid"] <- "FID"


data.temp$age.recruitment <- data.temp$f.21022.0.0
data.temp$month.birth <- data.temp$f.52.0.0
data.temp$year.birth <- data.temp$f.34.0.0
data.temp$menopause <- data.temp$f.2724.0.0
data.temp$parity <- data.temp$f.2734.0.0
data.temp$country.birth <- data.temp$f.1647.0.0
data.temp$year.immigrated.uk <- data.temp$f.3659.0.0
data.temp$age.hrt.start <- data.temp$f.3536.0.0
data.temp$age.hrt.last <- data.temp$f.3548.0.0
data.temp$hysterectomy <- data.temp$f.3591.0.0

white <- c("White","British","Irish","Any other white background")
other <- c("Prefer not to answer","Do not know","Mixed","Other ethnic group","White a
nd Black", "White and Black African","White and Asian","Any other mixed background")
other_asian <- c("Indian","Pakistani","Bangladeshi","Asian or Asian British","Any oth
er Asian background")
african_american <- c("Caribbean","African","Black or Black British","Any other Black
background")
chinese_american <- c("Chinese")

# mixed classified as others

temp <- data.temp$f.21000.0.0
data.temp$Race <- 4
data.temp$Race[temp%in%white] <- 1
data.temp$Race[temp%in%african_american] <- 2
data.temp$Race[temp%in%other] <- 4
data.temp$Race[temp%in%chinese_american] <- 6
data.temp$Race[temp%in%other_asian] <- 11


temp <- as.character(data.temp$f.84.0.0)
temp2 <- rep("Missing",nrow(data.temp))
temp2[is.na(temp)] <- "No"
temp2[!is.na(temp) & nchar(temp)==4] <- temp[!is.na(temp) & nchar(temp)==4]
data.temp$cancer.dx.year <- temp2


temp <- as.character(data.temp$f.84.0.0)
temp2 <- rep("Missing",nrow(data.temp))
temp2[is.na(temp)] <- "No"
temp2[!is.na(temp) & nchar(temp)<4 & as.numeric(temp)>0] <- temp[!is.na(temp) & nchar
(temp)<4 & as.numeric(temp)>0]
data.temp$cancer.dx.age <- temp2

temp <- data.temp$f.40006.0.0
temp2 <- rep("Missing",nrow(data.temp))
```

```
temp2[is.na(temp) | !str_starts(temp,"C50")] <- "No"
temp2[!is.na(temp) & str_starts(temp,"C50")] <- "Yes"
data.temp$bca.icd10 <- temp2


temp <- data.temp$f.40013.0.0
temp2 <- rep("Missing",nrow(data.temp))
temp2[is.na(temp) | !str_starts(temp,"174")] <- "No"
temp2[!is.na(temp) & str_starts(temp,"174")] <- "Yes"
data.temp$bca.icd9 <- temp2

data.temp$bca <-  "No"
data.temp$bca[data.temp$bca.icd10=="Yes" | data.temp$bca.icd9=="Yes"] <- "Yes"


temp <- data.temp$f.40006.0.0
temp2 <- rep("Missing",nrow(data.temp))
temp2[is.na(temp) | !str_starts(temp,"D05")] <- "No"
temp2[!is.na(temp) & str_starts(temp,"D05")] <- "Yes"
data.temp$insitu.icd10 <- temp2

temp <- data.temp$f.40013.0.0
temp2 <- rep("Missing",nrow(data.temp))
temp2[is.na(temp) | !str_starts(temp,"2330")] <- "No"
temp2[!is.na(temp) & str_starts(temp,"2330")] <- "Yes"
data.temp$insitu.icd9 <- temp2

table(data.temp$bca.icd9,data.temp$insitu.icd9)
table(data.temp$bca.icd10,data.temp$insitu.icd10)

temp <- data.temp$f.20007.0.0
temp2 <- rep("Missing",nrow(data.temp))
temp2[is.na(temp)] <- "Unclear"
temp2[!is.na(temp) & as.numeric(temp)<120 & as.numeric(temp)>0] <- temp[!is.na(temp)
 & as.numeric(temp)<120 & as.numeric(temp)>0]
data.temp$cancer.dx.age.interpolated <- temp2

temp <- data.temp$f.40008.0.0
temp2 <- rep("Missing",nrow(data.temp))
temp2[is.na(temp)] <- "Unclear"
temp2[!is.na(temp) & as.numeric(temp)<120 & as.numeric(temp)>0] <- temp[!is.na(temp)
 & as.numeric(temp)<120 & as.numeric(temp)>0]
data.temp$cancer.dx.age.registry <- temp2

data.temp$cancer.dx.age <- data.temp$cancer.dx.age.registry
data.temp$cancer.dx.age[data.temp$cancer.dx.age.registry%in%c("Missing","Unclear") &
 !data.temp$cancer.dx.age.interpolated%in%c("Missing","Unclear")] <- data.temp$cance
r.dx.age.interpolated[data.temp$cancer.dx.age.registry%in%c("Missing","Unclear") & !d
ata.temp$cancer.dx.age.interpolated%in%c("Missing","Unclear")]


temp <- data.temp$cancer.dx.age
temp[data.temp$bca == "No"] <- "No"
temp[data.temp$bca== "Yes" & data.temp$cancer.dx.age%in%c("Missing","No","Unclear")]
 <- "Missing"
data.temp$bca.dx.age  <- temp
```

```
id.prevalent <- which(data.temp$bca=="Yes" & !is.na(as.numeric(data.temp$cancer.dx.ag
e)) & data.temp$age.recruitment>=as.numeric(data.temp$cancer.dx.age))
id.incident <- which(data.temp$bca=="Yes" & !is.na(as.numeric(data.temp$cancer.dx.ag
e)) & data.temp$age.recruitment<as.numeric(data.temp$cancer.dx.age))
id.undetermined <- which(data.temp$bca=="Yes" & data.temp$cancer.dx.age%in%c("Missin
g","No","Unclear"))
data.temp$bca.incidence <- "Missing"
data.temp$bca.incidence[id.incident] <- "Incident"
data.temp$bca.incidence[id.prevalent] <- "Prevalent"
data.temp$bca.incidence[id.undetermined] <- "Yes.undetermined"
data.temp$bca.incidence[data.temp$bca=="No"] <- "No"

data.temp$cancer.year <- as.numeric(data.temp$year.birth) + floor(as.numeric(data.tem
p$cancer.dx.age))
data.temp$cancer.year[data.temp$cancer.dx.age%in%c("Missing","No")] <- data.temp$canc
er.dx.age[data.temp$cancer.dx.age%in%c("Missing","No")]

data.temp$bca.year <- as.numeric(data.temp$year.birth) + floor(as.numeric(data.temp$b
ca.dx.age))
data.temp$bca.year[data.temp$bca.dx.age%in%c("Missing","No")] <- data.temp$bca.dx.age
[data.temp$bca.dx.age%in%c("Missing","No")]


temp1 <- data.temp$f.2754.0.0
temp2 <- data.temp$f.3872.0.0
data.temp$Age1st <- 99
data.temp$Age1st[!is.na(temp1) & temp1>0 &data.temp$parity>1] <- temp1[!is.na(temp1)
 & temp1>0 & data.temp$parity>1]
data.temp$Age1st[!is.na(temp2) & temp2>0 &data.temp$parity==1] <- temp2[!is.na(temp2)
& temp2>0 & data.temp$parity==1]
data.temp$Age1st[data.temp$parity==0] <- 98

temp <- data.temp$f.2714.0.0
data.temp$AgeMen <- 99
data.temp$AgeMen[!is.na(temp) & temp>0] <- temp[!is.na(temp) & temp>0]


temp <- data.temp$f.20110.0.0
data.temp$fh.mother <- 0
data.temp$fh.mother[str_detect(temp,"Breast cancer")] <- 1

temp <- data.temp$f.20111.0.0
data.temp$fh.siblings <- 0
data.temp$fh.siblings[str_detect(temp,"Breast cancer")] <- 1

data.temp$N_Rels <- data.temp$fh.mother + data.temp$fh.siblings

data.temp$T1 <- data.temp$age.recruitment


temp <- data.temp$f.2674.0.0
temp1 <- rep("missing",nrow(data.temp))
temp1[temp%in%c("No","Yes")] <- as.character(temp[temp%in%c("No","Yes")])
data.temp$BCaScreeningEver <- temp1
```

```
data.temp$age.death <- data.temp$f.40007.0.0
data.temp$death.status <- "Alive"
data.temp$death.status[!is.na(data.temp$age.death)] <- "Dead"
data.temp$death.status[!is.na(data.temp$age.death) & data.temp$bca.incidence!="No"] <
- "DeadAftBCa"

data.select <- data.temp[,c(colnames(data.temp)[!str_starts(colnames(data.temp),"f
\\.")])]


saveRDS(data.select,paste0("data/gail_females_n",nrow(data.select),"_selected_",Sys.D
ate(),".rds"))
```

Gail model

```
library(BCRA)

data.select <- readRDS("data/gail_females_n264741_selected_2022-10-06.rds")
dim(data.select)
data.select$Age1st[data.select$parity==0] <- 98
gail <- data.select[,c("T1","Race","Age1st","AgeMen","N_Rels")]
gail$N_Biop <- 99
gail$HypPlas <- 99


data.out <- as.data.frame(data.select$FID)
colnames(data.out) <- "FID"
for(abs.year in c(5,10,15,2)){
  gail$T2 <- gail$T1 + abs.year
  gail.check <- recode.check(gail,Raw_Ind=1)
  print(gail.check[gail.check $Error_Ind==1,])
  gail.abs <- absolute.risk(gail, Raw_Ind=1)
  gail.rr <- relative.risk(gail, Raw_Ind=1)
  gail.out <- cbind(gail.abs,gail.rr)
  colnames(gail.out) <- paste0(colnames(gail.out),".interval_",str_pad(abs.year,width
=2,side="left",pad="0"))
  data.out <- cbind(data.out,gail.out)
}


for(max.age in c(60,70,80)){
  gail$T2 <- max.age
  gail.check <- recode.check(gail,Raw_Ind=1)
  print(gail.check[gail.check $Error_Ind==1,])
  gail.abs <- absolute.risk(gail, Raw_Ind=1)
  gail.rr <- relative.risk(gail, Raw_Ind=1)
  gail.out <- cbind(gail.abs,gail.rr)
  colnames(gail.out) <- paste0(colnames(gail.out),".lifetime_",str_pad(max.age,width=
2,side="left",pad="0"))
  data.out <- cbind(data.out,gail.out)
}

temp <- cbind(data.out,gail.check)
temp$gail.abs.interval_15[data.select$age.recruitment+15>80] <- NA
temp$gail.abs.interval_10[data.select$age.recruitment+10>80] <- NA
data.out <- temp
colnames(data.out) <- str_replace(colnames(data.out),"\\.abs","")

saveRDS(data.out,paste0("data/gail_females_n",nrow(data.out),"_",Sys.Date(),".rds"))
```

PRS

PRS variant file and score file

```
library(formattable)
PATH = "/Volumes/research/HG/HG7/Private/Datasets/UK Biobank/project/PRS/"
weights <- read.csv(paste0(PATH,"313_SNPS_NM.csv")) # Weights are available in Supple
mentary Ttable 1, columns: "Effective allele frequency in BCAC study" and "Beta from
 BCAC"
bim <- read.table(paste0("/Volumes/research/HG/HG7/Private/Datasets/UK Biobank/geno/P
RS/ukb22828-308snps.bim"),header=F)
colnames(bim) <- c("CHR","SNP","V3","POS","A1","A2")

weights$Beta_overall <- log(weights$OR_overall)
score.file <- merge(bim,weights,by.x=c("CHR","POS"),by.y=c("CHR","Position"),all = T)
dim(score.file)

id.notfound <- which(is.na(score.file$SNP))

score.file. <- score.file[,c("SNP","CHR","POS","SNPS","A1","A2","a1","EAF","Beta_over
all","OR_overall")]
colnames(score.file.)[1] <- "SNP_bim"
colnames(score.file.)[colnames(score.file.)=="a1"] <- "EA"
colnames(score.file.)[colnames(score.file.)=="EAF"] <- "EAF_BCAC"
score.file.[!is.na(score.file.$SNP_bim) & score.file.$SNP_bim=="rs10764337_",c("EA",
"EAF_BCAC","Beta_overall","OR_overall")] <- NA

score.file[id.notfound,]
score.file.[id.notfound,]

formattable(score.file.[id.notfound,])

summary(abs(score.file.$Beta_overall))
formattable(as.data.frame(t(as.matrix(summary(abs(score.file.$Beta_overall))))))

temp = NULL
for(i in 1:nrow(score.file)){
  temp1 <- as.character(score.file[i,c("a0","a1")])
  temp2 <- as.character(score.file[i,c("A1","A2")])
  if(length(setdiff(temp1,temp2))>=1){
    print(setdiff(temp1,temp2))
    print(score.file[i,])
    temp <- rbind(temp, score.file[i,])
  }
}
formattable(as.data.frame(temp[2,c("SNP","CHR","POS","A1","A2","a1","a0","Beta_overal
l")]))
score.file[!is.na(score.file$SNP) & score.file$SNP=="rs774021038","Beta_overall"] <-
 score.file[!is.na(score.file$SNP) & score.file$SNP=="rs774021038","Beta_overall"]*
 (-1)

score <- score.file[,c("SNP","a1","Beta_overall")]

write.table(score[!is.na(score$SNP),],paste0(PATH,"313_SNPS_NM_overall.score"),row.na
mes = F,quote = F,sep="\t",col.names = F)
write.table(score.file.,paste0(PATH,"variants314.txt"),row.names = F,quote = F,sep=
"\t") # Supplementary Table 1
```

Phenotype for frequency

```
library(stringr)
setwd("/Volumes/research/HG/HG7/Private/Datasets/UK Biobank/project/Overlap")

data.select <- readRDS("data/gail_females_n264741_selected_2022-10-06.rds")

pheno <- data.frame(data.select$FID)
colnames(pheno) <- "FID"
pheno$IID <- pheno$FID
pheno$Incidence <- 1
pheno$Incidence[data.select$bca.incidence=="Incident"] <- 2
pheno$Incidence[data.select$bca.incidence=="Prevalent"] <- -9
pheno$Prevalence <- 1
pheno$Prevalence[data.select$bca.incidence=="Prevalent"] <- 2
pheno$Prevalence[data.select$bca.incidence=="Incident"] <- -9

write.table(pheno,"data/pheno_264741_2022-10-06.txt",row.names = F,sep="\t",quote = F
)
```

Profile and frequency

```
Downloads/plink_mac_20220402/plink --bfile "/Volumes/research/HG/HG7/Private/Dataset
s/UK Biobank/geno/PRS/ukb22828-308snps" --score "/Volumes/research/HG/HG7/Private/Dat
asets/UK Biobank/project/PRS/313_SNPS_NM_overall.score" sum --freq --make-bed --out
"/Volumes/research/HG/HG7/Private/Datasets/UK Biobank/project/PRS/PRS308_n264246_2022
-10-06"

Downloads/plink_mac_20220402/plink --bfile "/Volumes/research/HG/HG7/Private/Dataset
s/UK Biobank/geno/PRS/ukb22828-308snps" --pheno "/Volumes/research/HG/HG7/Private/Dat
asets/UK Biobank/project/Overlap/data/pheno_264741_2022-10-06.txt" --freq case-contro
l --mpheno 1 --out "/Volumes/research/HG/HG7/Private/Datasets/UK Biobank/project/PRS/
PRS308_n264246_incidence_n7944_2022-10-06"

Downloads/plink_mac_20220402/plink --bfile "/Volumes/research/HG/HG7/Private/Dataset
s/UK Biobank/geno/PRS/ukb22828-308snps" --pheno "/Volumes/research/HG/HG7/Private/Dat
asets/UK Biobank/project/Overlap/data/pheno_264741_2022-10-06.txt" --freq case-contro
l --mpheno 2 --out "/Volumes/research/HG/HG7/Private/Datasets/UK Biobank/project/PRS/
PRS308_n264246_prevalence_n8131_2022-10-06"
```

```
library(stringr)
setwd("/Volumes/research/HG/HG7/Private/Datasets/UK Biobank/project/Overlap")

PATH = "/Volumes/research/HG/HG7/Private/Datasets/UK Biobank/project/PRS/"
variant <- read.table(paste0(PATH,"variants314.txt"),header = T)
frq.all <- read.table(paste0(PATH,"PRS308_n264246_2022-10-06.frq"),header = T)
frq.inc <- read.table(paste0(PATH,"PRS308_n264246_incidence_n7944_2022-10-06.frq.cc"
),header = T)
frq.pre <- read.table(paste0(PATH,"PRS308_n264246_prevalence_n8131_2022-10-06.frq.cc"
),header = T)

frq.inc. <- frq.inc[,c("SNP","MAF_U","NCHROBS_U","MAF_A","NCHROBS_A")]
colnames(frq.inc.) <- c("SNP","MAF_U","NCHROBS_U",paste0(c("MAF_A","NCHROBS_A"),"_Inc
ident"))
frq.pre. <- frq.pre[,c("SNP","MAF_A","NCHROBS_A")]
colnames(frq.pre.) <- c("SNP",paste0(c("MAF_A","NCHROBS_A"),"_Prevalent"))

frq <- merge(frq.all,frq.inc.,by="SNP")
frq <- merge(frq,frq.pre.,by="SNP")

output <- merge(variant[,!(colnames(variant)%in%c("MAF"))],frq,by.x=c("SNP_bim","CHR"
,"A1","A2"),by.y=c("SNP","CHR","A1","A2"),all=T)
output <- output[,c("SNP_bim","SNPS","CHR","POS","A1","A2","MAF","NCHROBS","MAF_U","N
CHROBS_U",paste0(c("MAF_A","NCHROBS_A"),"_Incident"),paste0(c("MAF_A","NCHROBS_A"),"_
Prevalent"),"EA","EAF_BCAC","Beta_overall","OR_overall")]
output <- output[order(output$POS),]
output <- output[order(output$CHR),]

write.csv(output,"output/PRS308_variants314_2022-10-06.csv",row.names = F)
```

PRS absolute risk

```
library(stringr)
library(formattable)
library(dplyr)
setwd("/Volumes/research/HG/HG7/Private/Datasets/UK Biobank/project/Overlap")

data.select <- readRDS("data/gail_females_n264741_selected_2022-10-06.rds")

prs <- read.table("/Volumes/research/HG/HG7/Private/Datasets/UK Biobank/project/PRS/P
RS308_n264246_2022-10-06.profile",header = T)

noPRS <- setdiff(data.select$f.eid,prs$FID)
write.table(noPRS, "output/IDs_noPRS.tab",sep="\t",row.names = F)

data.select <- merge(prs,data.select,by="FID")

data.mean <- data.select %>% group_by(bca.incidence) %>% dplyr::summarise(prs.mean =
 mean(SCORESUM),prs.sd = sd(SCORESUM))
formattable(data.mean)
```

```
library(numDeriv)
output.absolute.risk <- function(SD=0.616,MAX.AGE=80,INTERVAL=5,
                                 input.incidence.file.name,
                                 out.file.name,
                                 LIFE=F,
                                 PATH=NULL){
  incidence.read0 <- read.table(input.incidence.file.name, header = T, sep = ",")

  sd = SD
  d  = 0.6-0.4
  or    =   NULL
  seqx = seq(0,100,by=0.1)
  for (i in 1:1000){
    u = seqx[i]/100
    v = seqx[i+1]/100
    nu=d*(pnorm(qnorm(1-u)+sd)-pnorm(qnorm(1-v)+sd))
    de=(v-u)*(pnorm(qnorm(0.6)+sd)-pnorm(qnorm(0.4)+sd))
    or= c(or,nu/de)
  }

#The area under the curve (i.e the normal distribution curve, hence the use of pnorm)
gives the proportion of the population in any risk group.

  lower = seq(0,99.9,by=0.1)
  upper = lower+0.1
  name=paste(lower, "-", upper, "%", sep = "")
  prop = (upper - lower)/100

  or = cbind(name, round(or,4),prop)
  colnames(or)=c("PCT", "OR", "Nprob")

  beta.read = or

  # Input for
  incidence.read <- incidence.read0[,c("Age","BC_INCIDENCE_2011_2015","DEATH_2016")]
  colnames(incidence.read) <- c("t","BC_INCIDENCE","DEATH_INCIDENCE")

  n.prs       = dim(beta.read)[1]
  tau               = as.numeric(beta.read[,3])
  beta.g      = log(as.numeric(beta.read[,2]))
  prs.g       = beta.read[,"PCT"]
  incidence   = incidence.read[,"BC_INCIDENCE"]/100000
  mortality   = incidence.read[,"DEATH_INCIDENCE"]/100000

  Sg = lambda_g = AR_g = AR5_g          = matrix(NA, nrow = MAX.AGE, ncol = n.prs)
  lambda_0 = Sm                         = rep(NA, length = MAX.AGE)
  Sg0                                   = rep(1, length = n.prs)
  beta.g.mat                            = matrix(rep(beta.g, MAX.AGE), nrow = MAX.AG
E, ncol = n.prs, byrow = T)
  beta.g.mat[1:20,]                     = 0

  for (t in 1:MAX.AGE){

    numerator     = incidence[t]*sum(tau*Sg0)
    denominator   = sum(tau*exp(beta.g.mat[t,])*Sg0)
```

```
    lambda_0[t]   = numerator/denominator

    lambda_g[t,]  = lambda_0[t]*exp(beta.g.mat[t,])
    if (t == 1) {Sg[t,] = exp(-lambda_g[1,])} else {Sg[t,] = exp(-apply(lambda_g[1:
t,], 2, "sum"))}

    Sm[t]   = exp(-sum(mortality[1:t]))
    Sg0     = Sg[t,]

    if (t==1) {AR_g[t,] = lambda_0[t]*exp(beta.g.mat[t,])*Sg[t,]*Sm[t]}
    else {AR_g[t,] = apply(as.matrix(lambda_0[1:t]*Sm[1:t])%*%exp(beta.g.mat[t,])*Sg[
1:t,],2, "sum")}

    if (t >= 30){
      AR5_g[t-INTERVAL,]    =   (AR_g[t,]- AR_g[t-INTERVAL,])/(Sg[t-INTERVAL,]*Sm[t-I
NTERVAL])
    }

  }

  if(!LIFE){
    AR5_g   =   cbind(c(1:MAX.AGE), AR5_g)
    colnames(AR5_g) = c("AGE", as.character(prs.g))

    AR5_g[is.na(AR5_g)]=0

    OUT <- AR5_g
    colnames(OUT) <- c("AGE",lower)
    saveRDS(OUT, paste0(PATH,"interval_",str_pad(INTERVAL,width=2,side="left",pad=0),
".rds"))
  }
  if(LIFE){
    AR_g    =   cbind(c(1:MAX.AGE), AR_g)
    colnames(AR_g) = c("AGE", as.character(prs.g))

    LIFETIME <- AR_g
    colnames(LIFETIME) <- c("AGE",lower)
    saveRDS(LIFETIME,paste0(PATH,"lifetime_",MAX.AGE,".rds"))
  }
}

path = "/Volumes/research/HG/HG7/Private/Datasets/UK Biobank/project/PRS/absolute_ris
k/"

for(interval in c(2,5,10,15)){
  output.absolute.risk(
    SD=0.616,MAX.AGE=80,INTERVAL=interval,
    input.incidence.file.name="/Volumes/research/HG/HG7/Private/Datasets/UK Biobank/p
roject/Incident_mortality/breast cancer incident and mortality rates.csv",
    PATH=path)
}
for(max.age in c(60,70,80)){
  output.absolute.risk(
    SD=0.616,MAX.AGE=max.age,LIFE=T,
    input.incidence.file.name="/Volumes/research/HG/HG7/Private/Datasets/UK Biobank/p
roject/Incident_mortality/breast cancer incident and mortality rates.csv",
```

```
    PATH=path)
}
```

Individual's PRS absolute risk

```
library(dplyr)
path = "/Volumes/research/HG/HG7/Private/Datasets/UK Biobank/project/PRS/absolute_risk/"
filename <- list.files(path = path,pattern="*.rds")
filename <- cbind(filename,filename)
filename[,1] <- str_replace(filename[,1],".rds","")
rownames(filename) <- filename[,1]
absolute.risk.table = list()
for(i in filename[,1]){
  absolute.risk.table[[i]] <- readRDS(paste0(path,filename[i,2]))
}

mean(data.select$SCORESUM)
sd(data.select$SCORESUM)
data.select$prs.standardized <- (data.select$SCORESUM - (-0.306))/0.616
data.select$prs.percentile <- floor(pnorm(data.select$prs.standardized,mean=0,sd=1)*1000)/10


prs.abs <- list()
for(TYPE in filename[,1]){
  temp <- rep(NA,nrow(data.select))
  for(i in 1:nrow(data.select)){
    tempx <- absolute.risk.table[[TYPE]][absolute.risk.table[[TYPE]][,"AGE"]==data.select$age.recruitment[i],colnames(absolute.risk.table[[TYPE]])==data.select$prs.percentile[i]]
    if(length(tempx)==1) temp[i] <- tempx*100
    if(i%%10000==0) print(paste0(i," of ",nrow(data.select)))
  }
  prs.abs[[TYPE]] <- temp
  print(paste0(TYPE," complete"))
}

prs.abs. <- do.call(cbind,prs.abs)
prs.abs. <- as.data.frame(prs.abs.)
colnames(prs.abs.) <- paste0("prs.",colnames(prs.abs.))
prs.abs.$FID <- data.select$FID

output <- merge(data.select[,c("FID","age.recruitment","SCORESUM","prs.standardized","prs.percentile")],prs.abs.,by="FID")

temp <- output
temp$prs.interval_15[temp$age.recruitment+15>80] <- NA
temp$prs.interval_10[temp$age.recruitment+10>80] <- NA
output <- temp[,!colnames(temp)%in%"age.recruitment"]

saveRDS(output, paste0("data/prs308_AbsoluteRisk_n",nrow(data.select),"_",Sys.Date(),".rds"))
```

Outcome

```
data.select$bca.incidence.b480 <- data.select$bca.incidence
data.select$bca.incidence.b480[!is.na(as.numeric(data.select$bca.dx.age)) & as.numeri
c(data.select$bca.dx.age)>80] <- "No"

bca <- list()
bca.dx.age <- as.numeric(data.select$bca.dx.age)
recruitment.age <- as.numeric(data.select$age.recruitment)
for(i in c(2,5,10,15)){
  temp <- data.select$bca.incidence.b480
  temp[bca.dx.age>(recruitment.age +i)] <- "No"
  bca[[paste0("bca.interval_",str_pad(i,2,"left","0"))]] <- temp
}

for(i in c(60,70,80)){
  temp <- data.select$bca.incidence.b480
  temp[bca.dx.age>i] <- "No"
  bca[[paste0("bca.lifetime_",str_pad(i,2,"left","0"))]] <- temp
}

bca. <- as.data.frame(do.call(cbind,bca))
bca.$FID <- data.select$FID

saveRDS(bca., paste0("data/BCaOutput_n",nrow(data.select),"_",Sys.Date(),".rds"))
```

LoF

```
data.ptv <- read.csv("/Volumes/research/HG/HG7/Private/Datasets/UK Biobank/WES/9gene
s/ukb23158-9genes-254635females-508var.csv")
data.ptv <- data.ptv[,!str_starts(colnames(data.ptv),"X")]
saveRDS(data.ptv,paste0("data/ptv_n",nrow(data.ptv),".rds"))
```

Merging dataset

```
library(stringr)
setwd("/Volumes/research/HG/HG7/Private/Datasets/UK Biobank/project/Overlap")

data.select <- readRDS("data/gail_females_n264741_selected_2023-03-06.rds")

data.bca <- readRDS("data/BCaOutput_n264741_2022-10-06.rds")

data.gail <- readRDS("data/gail_females_n264741_2022-10-06.rds") # need to be edited
 to re-run

data.prs <- readRDS("data/prs308_AbsoluteRisk_n264246_2022-10-06.rds")

data.ptv <- readRDS("data/ptv_n254635.rds")

data.gail.bca <-  merge(data.gail,data.bca,by="FID")
data.prs.gail.bca <-  merge(data.prs,data.gail.bca,by="FID")
data.genetic <-  merge(data.ptv,data.prs.gail.bca,by="FID")


# checking numbers for Supplementary Figure - flowchart
setdiff(data.prs$FID,data.select$FID) # not in flowchart
ptv.no.questionnaire <- setdiff(data.ptv$FID,data.select$FID) # not in flowchart
prs.no.ptv <- setdiff(data.prs$FID,data.ptv$FID) # not in flowchart

questionnaire.no.prs <- setdiff(data.select$FID,data.prs$FID)
questionnaire.prs.no.ptv <- setdiff(data.prs.gail.bca$FID,data.ptv$FID)

length(questionnaire.no.prs)
length(questionnaire.prs.no.ptv)
noPTV <- setdiff(data.prs.gail.bca$FID,data.ptv$FID)
write.table(noPTV, "output/IDs_noPTV.tab",sep="\t",row.names = F)

data <- merge(data.genetic,data.select,by.x="FID")

saveRDS(data, paste0("data/females_n",nrow(data),"_gail_prs308_AbsoluteRisk_",Sys.Dat
e(),".rds"))
```

ANALYSIS

Setup

```
library(stringr)
library(ggplot2)
library(ggpubr)
library(gridExtra)
library(reshape2)
library(dplyr)
library(formattable)
library(VennDiagram)
library(brew)
library(Unicode)
library(ggrepel)
library(pROC)
setwd("/Volumes/research/HG/HG7/Private/Datasets/UK Biobank/project/Overlap")

data <- readRDS("data/females_n253953_gail_prs308_AbsoluteRisk_2023-03-07.rds")

### Functions

form <- function(x){
  format(x,big.mark = ",",big.interval = 3L)
}
form.2 <- function(var,dp=2){
  str_trim(format(round(as.numeric(var),dp),nsmall=dp))
}
form.ci <- function(var,ci,dp=2){
  paste0(form.2(var)," (",form.2(ci[,1])," \226 ",form.2(ci[,2]),")")
}

form.ci1 <- function(var,ci,dp=2){
  paste0(form.2(var)," (",form.2(ci[1])," \226 ",form.2(ci[2]),")")
}

form.e<- function(var,dp=2){
  str_trim(formatC(var,digits=dp,format="E"))
}

tab1.s <- function(VAR,DATA=data,dp=2){
  temp <- table(DATA[,VAR])
  temp1 <- temp/nrow(DATA)*100
  output <- paste0(form(temp)," [",form.2(temp1,dp=dp),"%]")
  return(as.data.frame(cbind(output,form(temp))))
}

tab2.s <- function(VAR, GROUP, DATA=data, dp=2){
  temp. <- table(DATA[,VAR],DATA[,GROUP])
  temp <- t(temp.)
  temp1 <- temp/rowSums(temp)*100
  output = NULL
  for(i in 1:ncol(temp1)){
    output = cbind(output,paste0(form(temp[,i])," (",form.2(temp1[,i],dp=dp),"%)"))
  }
  output <- t(output)
  colnames(output) <- colnames(temp.)
  return(as.data.frame(output))
}
```

```
tab2 <- function(VAR, GROUP, DATA=data, dp=2, MISSING.VALUE.CODE="missing"){
  if(class(DATA[,GROUP])!="factor") DATA[,GROUP] <- as.factor(DATA[,GROUP])
  lvl = levels(DATA[,GROUP])
  output1 = tab1.s(VAR=VAR,DATA=DATA,dp=dp)
  output2 = tab2.s(VAR=VAR, GROUP=GROUP, DATA=DATA, dp=dp)
  id <- which(!(is.na(DATA[,VAR]) | DATA[,VAR]%in%c(MISSING.VALUE.CODE)))
  p <- chisq.test(DATA[id,VAR],DATA[id,GROUP])$p.value
  output <- as.data.frame(cbind("","",output1,output2,""))
  output[1,ncol(output)] <- ifelse(p<0.001,"<0.001",form.2(p,3))
  colnames(output)[ncol(output)] <- "P"
  output[1,1] <- VAR
  output[,2] <- levels(as.factor(DATA[,VAR]))
  colnames(output)[1:4] <- c("Variable","Levels","REMOVE","All")
  formattable(output)
  return(output)
}

sum1.s <- function(VAR,DATA=data,dp=2){
  temp <- summary(DATA[,VAR])
  output <- paste0(form.2(temp[4],dp)," (",form.2(temp[2],dp)," to ",form.2(temp[5],d
p),")")
  return(cbind(length(which(!is.na(DATA[,VAR]))),output))
}

sum2.s <- function(VAR,GROUP,DATA=data,dp=2){ # must be factor
  if(class(DATA[,GROUP])!="factor") DATA[,GROUP] <- as.factor(DATA[,GROUP])
  lvl = levels(DATA[,GROUP])
  temp.o = NULL
  for(group in lvl){
    temp <- summary(DATA[DATA[,GROUP]==group,VAR])
    temp.o <- c(temp.o,paste0(form.2(temp[4],dp)," (",form.2(temp[2],dp)," to ",form.
2(temp[5],dp),")"))
  }
  output <- as.data.frame(t(as.matrix(temp.o)))
  colnames(output) <- lvl
  return(output)
}

sum2 <- function(VAR,GROUP,DATA=data,dp=2){
  if(class(DATA[,GROUP])!="factor") DATA[,GROUP] <- as.factor(DATA[,GROUP])
  lvl = levels(DATA[,GROUP])
  output1 <- sum1.s(VAR=VAR,DATA=DATA,dp=dp)
  output2 <- sum2.s(VAR=VAR,GROUP=GROUP,DATA=DATA,dp=dp)
  p <- kruskal.test(DATA[,VAR],DATA[,GROUP])$p.value
  output <- as.data.frame(cbind("","",output1,output2,""))
  output[1,ncol(output)] <- ifelse(p<0.001,"<0.001",form.2(p,3))
  colnames(output)[ncol(output)] <- "P"
  output[1,1] <- VAR
  output[1,2] <- VAR
  colnames(output)[1:4] <- c("Variable","Levels","REMOVE","All")
  formattable(output)
  return(output)
}

### Additional recoding
```

```
ptv9.list <- c("ATM","BARD1","BRCA1","BRCA2","CHEK2","PALB2","RAD51D","RAD51C","TP53"
)

type.list = c("interval_05","interval_10","interval_15","interval_02","lifetime_80",
"lifetime_70","lifetime_60")

ar.list <- c("gail.","prs.")

for(TYPE in type.list){
  VAR2 = paste0("bca.",TYPE)
  data[,VAR2] <- factor(data[,VAR2],c("No","Incident","Prevalent"))
}

data$ptv9 <- rowSums(data[,ptv9.list])
data$ptv9[data$ptv9>0] <- 1

data$FH.BCa <- data$N_Rels
data$FH.BCa[data$FH.BCa>0] <- 1

temp <- data$AgeMen
temp1 <- rep(NA,nrow(data))
temp1[temp!=99] <- data$AgeMen[temp!=99]
data$AgeMen. <- temp1
temp1 <- rep("missing",nrow(data))
temp1[temp!=99] <- data$AM_Cat[temp!=99]
data$AM_Cat. <- temp1

temp <- data$parity
temp1 <- rep("missing",nrow(data))
temp1[temp%in%c(0,1,2)] <- data$parity[temp%in%c(0,1,2)]
temp1[temp>=3] <- "3+"
data$parity. <- temp1

temp <- data$Age1st
temp1 <- rep(NA,nrow(data))
temp1[temp!=99] <- data$Age1st[temp!=99]
data$Age1st. <- temp1

temp1 <- rep("missing",nrow(data))
temp1[temp!=99] <- data$AF_Cat[temp!=99]
temp1[data$parity==0] <- "No child"
data$AF_Cat. <- temp1


temp <- data$age.recruitment
temp1 <- rep("missing",nrow(data))
temp1[temp<50] <- "1.39 to 49"
temp1[temp>=50 & temp<60] <- "2.50 to 59"
temp1[temp>=60 & temp<70] <- "3.60 to 69"
temp1[temp>=70] <- "4>=70"
data$age.recruitment.cat <- temp1

temp1 <- as.numeric(data$year.birth)
temp2 <- as.numeric(data$age.recruitment)
data$year.recruitment <- temp1 + temp2
```

```
temp <- data$menopause
temp1 <- rep("missing",nrow(data))
temp1[temp%in%c("No","Yes")] <- as.character(temp[temp%in%c("No","Yes")])
data$menopause. <- temp1


data$Race <- factor(data$Race,c(1,2,6,11,4))
data$AF_Cat. <- factor(data$AF_Cat.,c("No child",0,1,2,3,"missing"))
data$menopause. <- factor(data$menopause.,c("Yes","No","missing"))
data$BCaScreeningEver <- factor(data$BCaScreeningEver,c("Yes","No","missing"))

data$bca.dx.age. <- as.numeric(data$bca.dx.age)
data$bca.dx.age.[data$bca.lifetime_80=="No"] <- NA

data$followup = as.numeric(2020)
data$followup[data$bca.incidence!="No"] <- as.numeric(data$bca.year[data$bca.incidenc
e!="No"])

data$followup.time <- data$followup - data$year.recruitment
```

TABLE AND FIGURES

T1 Characteristics

```
TYPE = "lifetime_80"
dataset0 <- data[data[,paste0("bca.",TYPE)]!="Prevalent",]
dataset0[,paste0("bca.",TYPE)]  <- factor(as.character(dataset0[,paste0("bca.",TYPE)]
),c("No","Incident"))

#------------------------------------------------------------------------------#

var.cat.list <- c("Race","N_Rels","AM_Cat.","parity.","AF_Cat.","menopause.","BCaScre
eningEver","ptv9")

var.con.list <- c("age.recruitment","SCORESUM")
df.con.list <- list("age.recruitment"=0, "SCORESUM"=3)

order.list <- c("age.recruitment","Race","N_Rels","AM_Cat.","parity.","AF_Cat.","meno
pause.","BCaScreeningEver","SCORESUM","ptv9")

# c("bca.dx.age","bca.year")
#------------------------------------------------------------------------------#

GROUP=paste0("bca.",TYPE)

table1 <- list()
for(VAR in var.cat.list){
print(VAR)
    table1[[VAR]] <- tab2(VAR=VAR,GROUP=GROUP,DATA=dataset0,dp=0)
}
for(VAR in var.con.list){
  table1[[VAR]] <- sum2(VAR=VAR,GROUP=GROUP,dp=df.con.list[[VAR]],DATA=dataset0)
}

table1. = NULL

for(VAR in order.list){
  print(table1[[VAR]])
  table1. <- rbind(table1.,table1[[VAR]],"")
}

data$bca.dx.age. <- as.numeric(data$bca.dx.age)
data$bca.dx.age.[data$bca.lifetime_80=="No"] <- NA
VAR = "bca.dx.age."
temp <- sum1.s(VAR=VAR,DATA = dataset0,dp=0)
table1.[2,"Variable"] <- VAR
table1.[2,"REMOVE"] <- temp[1]
table1.[2,"Incident"] <- temp[2]

write.csv(table1.,"output/table1_incident_healthy.csv")
```

sF3 AUCs

```
dataset0 <- data[data$bca.lifetime_80!="Prevalent",]

type.list = c("interval_02","interval_05","interval_10")

type.order.list <- c(paste0("interval_", c("02","05","10")))

year.list <- list("interval_02"=2,"interval_05"=5,"interval_10"=10)

for(TYPE in type.list){
  dataset0[,paste0("bca.",TYPE)]  <- factor(as.character(dataset0[,paste0("bca.",TYP
E)] ),c("No","Incident"))
}

#------------------------------------------------------------------------------#

AR="prs"
TYPE = type.order.list[1]

measure.list <- list("interval_02"="2-year","interval_05"="5-year","interval_10"="10-
year")

threshold.list = plot.list = list()
for(TYPE in type.order.list){
  for(AR in c("prs","gail")){
    YEAR = year.list[[TYPE]]
    temp.data <- dataset0[(dataset0$age.recruitment+ YEAR) <= 80,]
    fit <- glm(family = "binomial",data=temp.data,
               formula=formula(paste0("bca.",TYPE,"~",AR,".",TYPE)))
    outcome <- temp.data[,paste0("bca.",TYPE)]
    response <- predict(fit,type = "response")
    fit.roc <- roc(outcome~response)
    temp <- ci(fit.roc)
    AUC <- temp[2]
    AUC.lb <- temp[1]
    AUC.ub <- temp[3]

    temp <- plot.roc(fit.roc,print.thres=T,print.thres.best.method = "youden")
    youdenJ <- temp$sensitivities + temp$specificities -1
    p <- temp$thresholds[which.max(youdenJ)]
    Threshold <- (log(p/(1-p)) - fit$coefficients[1]) / fit$coefficients[2]
    Sensitivity <- temp$sensitivities[which.max(youdenJ)]
    Specificity <- temp$specificities[which.max(youdenJ)]
    youdenJ <- youdenJ[which.max(youdenJ)]
    threshold.list[[TYPE]][[AR]] <- c(AR,TYPE,Threshold,youdenJ,Sensitivity,Specifici
ty,AUC,AUC.lb,AUC.ub)


    label.threshold <- paste0(toupper(AR)," ",measure.list[[TYPE]]," absolute risk th
reshold = ", form.2(Threshold,1),"% \nSensitivity = ",form.2(Sensitivity,2)," \nSpeci
ficity = ",form.2(Specificity,2))
    label.auc <- paste0("AUC (95%CI): ",form.2(AUC,3)," (",form.2(AUC.lb,3)," - ",for
m.2(AUC.ub,3),")")
    label.x = -1 #Specificity*(-1) + .2
    label.y = 1 #Sensitivity
    plot.list[[paste0(AR,".",TYPE)]] <-
```

```
      ggroc(fit.roc) +
      geom_abline(intercept = 1,slope=1,alpha=.5,color="darkslategrey",linetype="dash
ed") +
      geom_point(x=Specificity*(-1),y=Sensitivity) +
      geom_text(x=as.numeric(label.x),y=as.numeric(label.y),label=label.threshold,hju
st=0,vjust=1,size=3,position="identity",check_overlap = T) +
      geom_text(x=0,y=0,label=label.auc,hjust=1,vjust=0,size=3,position="identity",ch
eck_overlap = T) +
      labs(x="Specificity",y="Sensitivity") +
      theme_bw()

  }
}

png("plot/sF Getting best threshold_AUC_continuousMeasure.png",width=2500,height = 22
50,res=300)
ggarrange(plotlist=plot.list,ncol = 2,nrow=3,labels = "AUTO")
dev.off()

threshold.list. = list()
for(TYPE in type.order.list){
  threshold.list.[[TYPE]] <- do.call(rbind,threshold.list[[TYPE]])
}
threshold.list.. <- do.call(rbind,threshold.list.)
colnames(threshold.list..) <- c("Measure","TYPE","Threshold","youdenJ","Sensitivity",
"Specificity","AUC","AUC.lb","AUC.ub")

write.table(threshold.list..,"output/sF source data - AUC and threshold.txt",sep="\t"
,row.names = F)
```

Function

```
RiskRatio.AUC <- function(THRESHOLD,cases,high.risk){
  nAtrisk <- length(high.risk)
  nHighrisk <- length(which(high.risk==1))
  nLowrisk <- length(which(high.risk==0))
  ncases <- length(which(cases=="Incident"))
  nidentified <- length(which(cases=="Incident" & high.risk==1))
  nmissed <- length(which(cases=="Incident" & high.risk==0))
  pidentified.cases <- nidentified/ncases
  pcases.nAtrisk <- R0 <-ncases/nAtrisk
  pHighrisk.Atrisk <- nHighrisk/nAtrisk
  pidentifiedHighrisk <- R1 <- nidentified/nHighrisk
  pmissedLowrisk <-  nmissed/nLowrisk
  RiskRatio <- R1 / R0

  fit.roc <- roc(cases~high.risk)
  temp <- ci(fit.roc)
  AUC <- temp[2]
  AUC.lb <- temp[1]
  AUC.ub <- temp[3]

  temp <- table(high.risk,cases)
  FPR = temp["1","No"] / sum(temp["1",]) #false positive rate
  FNR = temp["0","Incident"] / sum(temp["0",]) #false negative rate
  TPR = temp["1","Incident"] / sum(temp["1",]) #true positive rate
  TNR = temp["0","No"] / sum(temp["0",]) # true negative rate
  sensitivity = temp["1","Incident"] / sum(temp[,"Incident"])
  specificity = temp["0","No"] / sum(temp[,"No"])

  diff.pidentified.casespHighrisk.Atrisk <- pidentified.cases - pHighrisk.Atrisk
  output <- as.data.frame(cbind(TYPE,THRESHOLD,
                                nAtrisk,nHighrisk,nLowrisk,pHighrisk.Atrisk,
                                ncases,nidentified,nmissed,
                                pcases.nAtrisk,pidentified.cases,
                                pidentifiedHighrisk,pmissedLowrisk,RiskRatio,
                                sensitivity,specificity,
                                AUC,AUC.lb,AUC.ub,
                                FPR,FNR,TPR,TNR))
  return(output)
}
```

```
highest.. <- read.table("output/sF source data - AUC and threshold.txt",header=T,sep=
"\t")

dataset0 <- data[data$bca.lifetime_80!="Prevalent",]

#-------------------------------------------------------------------------------#

type.list = c("interval_02","interval_05","interval_10")

type.order.list <- c(paste0("interval_", c("02","05","10")))

year.list <- list("interval_02"=2,"interval_05"=5,"interval_10"=10)

allocate.risk.var <- function(x,THRESHOLD){
  output <- rep("NA",length(x))
  temp <- x
  output[!is.na(temp) & temp>THRESHOLD] <- 1
  output[!is.na(temp) & temp<=THRESHOLD] <- 0
  return(output)
}

threshold.highest.list <- list()
for(TYPE in type.order.list){
  for(AR in c("prs","gail")){
    threshold.highest.list[[TYPE]][[AR]] = highest..[highest..$TYPE==TYPE & highes
t..$Measure==AR,"Threshold"]
  }
}

output = as.data.frame(dataset0[,c("FID")])
colnames(output) <- "FID"
for(TYPE in type.order.list){
  for(AR in c("prs","gail")){
    VAR = paste0(AR,".",TYPE)
    threshold = threshold.highest.list[[TYPE]][[AR]]
    DATA = dataset0[,c(VAR)]
    temp <- as.data.frame(allocate.risk.var(DATA,threshold))
    colnames(temp) <- paste0("high_",form.2(threshold,1),"_",VAR)
    output <- cbind(output,temp)
  }
  print(TYPE)
  print(threshold)
}
write.csv(output,paste0("data/Risk_bestThreshold_AUCcontinuous_n",nrow(output),"_",Sy
s.Date(),".csv"),row.names = F)
```

F1 Scatterplot - choose combination

```
data.risk <- read.csv("data/Risk_bestThreshold_AUCcontinuous_n246142_2023-05-08.csv")
dataset0 <- data[data$bca.lifetime_80!="Prevalent",]
dataset0 <- merge(dataset0,data.risk,by="FID")

for(TYPE in type.list){
  dataset0[,paste0("bca.",TYPE)]  <- factor(as.character(dataset0[,paste0("bca.",TYP
E)] ),c("No","Incident"))
}

#-------------------------------------------------------------------------------#

type.list = c("interval_02","interval_05","interval_10")

type.order.list <- c(paste0("interval_", c("02","05","10")))

year.list <- list("interval_02"=2,"interval_05"=5,"interval_10"=10)

key.list <- list("prs" = c(1,0,0,0),
                 "gail" = c(0,1,0,0),
                 "fh" = c(0,0,1,0),
                 "lof" = c(0,0,0,1),
                 "prs.gail" = c(1,1,0,0),
                 "prs.fh" = c(1,0,1,0),
                 "prs.lof" = c(1,0,0,1),
                 "gail.fh" = c(0,1,1,0),
                 "gail.lof" = c(0,1,0,1),
                 "fh.lof" = c(0,0,1,1),
                 "prs.gail.fh" = c(1,1,1,0),
                 "prs.gail.lof" = c(1,1,0,1),
                 "prs.fh.lof" = c(1,0,1,1),
                 "gail.fh.lof" = c(0,1,1,1),
                 "prs.gail.fh.lof" = c(1,1,1,1)
)

#-------------------------------------------------------------------------------#


temp <- colnames(dataset0)[str_starts(colnames(dataset0),"high")]
high_risk = list()
for(TYPE in type.order.list){
  for(AR in c("prs","gail")){
    high_risk[[TYPE]][[AR]] <- temp[str_detect(temp,TYPE) & str_detect(temp,AR) ]
  }
}

output = NULL
for(TYPE in type.order.list){
  for(KEY in names(key.list)){
    YEAR = year.list[[TYPE]]
    temp <- dataset0[(dataset0$age.recruitment+ YEAR) <= 80,]

    high.risk.list <- list(
      "prs" = high_risk[[TYPE]][["prs"]],
      "gail" = high_risk[[TYPE]][["gail"]],
      "fh" = "FH.BCa",
```

```
      "lof" = "ptv9"
    )

    # get columns to make high.risk variable
    key.use <- key.list[[KEY]]
    temp.c = NULL
    for(k in 1:4){
      if(key.use[k]==1){
        print(high.risk.list[[k]])
        temp.c = cbind(temp.c,as.numeric(temp[,high.risk.list[[k]]]))
      }
    }
    temp.c. <- rowSums(temp.c)
    temp.c.[temp.c.>1] <- 1
    high.risk <- temp.c.

    cases <- temp[,paste0("bca.",TYPE)]
    output <- rbind(output,RiskRatio.AUC(THRESHOLD=KEY,cases,high.risk))
  }
}
output

dataplot <- output
colnames(dataplot)[str_detect(colnames(dataplot),"THRESHOLD")] <- "KEY"
dataplot$KEY <- toupper(dataplot$KEY)
dataplot$KEY <- str_replace_all(dataplot$KEY,"\\.","\u00B7")
dataplot$KEY <- str_replace_all(dataplot$KEY,"LOF","LoF")

dataplot$highestpoint <- "no"
for(TYPE in type.order.list){
  temp <- max(as.numeric(dataplot$AUC)[dataplot$TYPE==TYPE])
  id <- which(dataplot$TYPE==TYPE & as.numeric(dataplot$AUC)==temp)
  dataplot$highestpoint[id] <- "yes"
}

dataplot$KEY. <- NA
dataplot$KEY.. <- NA
dataplot$KEY.[dataplot$highestpoint=="no"] <- as.character(dataplot$KEY[dataplot$high
estpoint=="no"])
dataplot$KEY..[dataplot$highestpoint=="yes"] <- as.character(dataplot$KEY[dataplot$hi
ghestpoint=="yes"])

MARGIN = margin(15,5,5,5)
FONT.SIZE = 10
x.breaks = seq(0,1,.1)
png("plot/F Getting best combination.png",width=2000,height = 2000,res=300)
ggplot(data=dataplot,aes(x=as.numeric(pHighrisk.Atrisk),y=as.numeric(pidentified.case
s))) +
  geom_point(aes(color = TYPE),alpha=.75) +
  geom_abline(intercept = 0,slope=1,colour="grey",linetype= "dashed") +
  geom_text_repel(aes(label = KEY.,  color = TYPE),
                  size = 2, direction="x", force=0.5, nudge_x=0.075,
                  alpha=1, segment.size=0.1, max.overlaps=20,show.legend = F,na.rm =
T)+
  geom_label_repel(aes(label = KEY..,  color = TYPE), size = 3,
                   direction="both",nudge_y = 0.1, show.legend = F,na.rm=T) +
```

```
  scale_color_manual(breaks=c(paste0("interval",c("_02","_05","_10"))),
                     values = c("black","darkred","azure4"),
                     labels=c(paste0(c(2,5,10),"-year"))) +
  scale_x_continuous(breaks = x.breaks,label = x.breaks) +
  scale_y_continuous(breaks = x.breaks,label = x.breaks) +
  coord_cartesian(xlim=c(x.breaks[1],x.breaks[length(x.breaks)]),
                  ylim=c(x.breaks[1],x.breaks[length(x.breaks)]),expand=F) +
  labs(x="Proportion of individuals flagged as high-risk",
       y="Proportion of cases diagnosed within x years identified as high risk",
       color="x-year absolute risk") +
  theme_bw() +
  theme(legend.position = c(.95,0.05),
      legend.direction = "vertical",
      legend.justification = c(1,0),
      legend.background = element_rect(fill=rgb(1,1,1,0.4)),
      legend.text = element_text(size = FONT.SIZE),
      legend.title = element_text(size = FONT.SIZE),
      axis.title = element_text(size = FONT.SIZE),
      plot.margin = MARGIN)
dev.off()


write.csv(dataplot,"output/Supporting data 3-F Getting best combination.csv",row.name
s = F)
```

F2 Venn diagram

```
source("quad.venn.change.base_1.R")

#-------------------------------------------------------------------------------#

temp <- colnames(dataset0)[str_starts(colnames(dataset0),"high")]
high_risk = list()
for(TYPE in type.order.list){
  for(AR in c("prs","gail")){
    high_risk[[TYPE]][[AR]] <- temp[str_detect(temp,TYPE) & str_detect(temp,AR) ]
  }
}

fig.lab.list <- list("interval_02"="2-year absolute risk",
                     "interval_05"="A. 5-year absolute risk",
                     "interval_10"="B. 10-year absolute risk")

plot.venn = list()
for(TYPE in type.order.list){
  OUTCOME = paste0("bca.",TYPE)
  ID = "FID"
  VAR1 = high_risk[[TYPE]][["prs"]]
  VAR2 = high_risk[[TYPE]][["gail"]]
  VAR3 = "ptv9"
  VAR4 = "FH.BCa"

  DATA <- dataset0[(dataset0$age.recruitment+ year.list[[TYPE]]) <= 80,]
  table(DATA[,OUTCOME],useNA = "ifany")

  base.x <- list("PRS"=unique(DATA[DATA[,VAR1]==1,ID]),
                 "LoF"=unique(DATA[DATA[,VAR3]==1,ID]),
                 "GAIL"=unique(DATA[DATA[,VAR2]==1,ID]),
                 "FH"=unique(DATA[DATA[,VAR4]==1,ID]))

  print(paste0(TYPE,": High-risk"))
  temp = length(unique(do.call("c",base.x)))
  print(paste0("ALL: ",temp ," (",
               round(temp/nrow(DATA)*100),"%)"))
  var ="GAIL"
  print(paste0(var,": ",length(base.x[[var]])," (",
               round(length(base.x[[var]])/nrow(DATA)*100),"%)"))
  var ="PRS"
  print(paste0(var,": ",length(base.x[[var]])," (",
               round(length(base.x[[var]])/nrow(DATA)*100),"%)"))
  var ="FH"
  print(paste0(var,": ",length(base.x[[var]])," (",
               round(length(base.x[[var]])/nrow(DATA)*100),"%)"))
  var ="LoF"
  print(paste0(var,": ",length(base.x[[var]])," (",
               round(length(base.x[[var]])/nrow(DATA)*100),"%)"))

  DATA1 <- DATA[DATA[,OUTCOME]=="Incident",]
  x <- list("PRS"=unique(DATA1[DATA1[,VAR1]==1,ID]),
            "LoF"=unique(DATA1[DATA1[,VAR3]==1,ID]),
            "GAIL"=unique(DATA1[DATA1[,VAR2]==1,ID]),
            "FH"=unique(DATA1[DATA1[,VAR4]==1,ID]))
```

```
  print(paste0(TYPE,": case - ", nrow(DATA1)))
  temp = length(unique(do.call("c",x)))
  print(paste0("ALL: ",temp ," (",
               round(temp/nrow(DATA1)*100),"%)"))
  var ="GAIL"
  print(paste0(var,": ",length(x[[var]])," (",
               round(length(x[[var]])/nrow(DATA1)*100),"%)"))
  var ="PRS"
  print(paste0(var,": ",length(x[[var]])," (",
               round(length(x[[var]])/nrow(DATA1)*100),"%)"))
  var ="FH"
  print(paste0(var,": ",length(x[[var]])," (",
               round(length(x[[var]])/nrow(DATA1)*100),"%)"))
  var ="LoF"
  print(paste0(var,": ",length(x[[var]])," (",
               round(length(x[[var]])/nrow(DATA1)*100),"%)"))

  grid.newpage()
  list.names <- names(x)


    p <- as_ggplot(quad.venn.change.base(x=x,
                                         base.x=base.x,
                                         category = list.names,
                                         print.mode="both",
                                         sigdigs=0))
   plot.venn[[TYPE]] <- annotate_figure(p,fig.lab = fig.lab.list[[TYPE]],fig.lab.pos=
"top.left")

}


png("plot/F Venn diagram_2.png",width=2200,height = 1800,res=300,pointsize = 10)
plot.venn$interval_02
dev.off()


png("plot/F Venn diagram_5_10.png",width=2200,height = 2400,res=300,pointsize = 10)
grid.arrange(grobs=list(plot.venn$interval_05,plot.venn$interval_10),ncol=1)
dev.off()
```

sF2 Density plots - distribution

```
dataplot.o = NULL
tab.o = NULL
for(TYPE in type.order.list){
  temp <- dataset0[,c(paste0("bca.",TYPE),paste0("prs.",TYPE),paste0("gail.",TYPE))]
  dataplot <- melt(temp,id=paste0("bca.",TYPE))
  colnames(dataplot) <- c("Outcome","Type","Value")

  temp.tab <- dataplot %>%  group_by(Outcome,Type) %>%
    summarise(median=summary(Value)[3],
              IQR1=summary(Value)[2],
              IQR2=summary(Value)[5],
              max=summary(Value)[6]) %>%  as.data.frame()
  tab.o <- rbind(tab.o,temp.tab)
  dataplot.o <- rbind(dataplot.o,dataplot)
}
unique(dataplot.o$Type)
dataplot.o$Measure <- toupper(str_split_fixed(dataplot.o$Type,"\\.",2)[,1])
dataplot.o$xyear <- paste0(as.numeric(str_split_fixed(dataplot.o$Type,"_",2)[,2]),"-y
ear")
dataplot.o$xyear <- factor(dataplot.o$xyear,c(paste0(c(2,5,10),"-year")))
dataplot.o$Outcome <- factor(dataplot.o$Outcome,c("No","Incident"))

tab.o$Measure <- toupper(str_split_fixed(tab.o$Type,"\\.",2)[,1])
tab.o$xyear <- paste0(as.numeric(str_split_fixed(tab.o$Type,"_",2)[,2]),"-year")
tab.o$xyear <- factor(tab.o$xyear,c(paste0(c(2,5,10),"-year")))
tab.o$Outcome <- factor(tab.o$Outcome,c("No","Incident"))
tab.o$labels <- paste0(tab.o$Outcome,": ",
                       form.2(tab.o$median,1)," (",
                       form.2(tab.o$IQR1,1)," \u2012 ",
                       form.2(tab.o$IQR2,1),")",
                       "\nMax=",form.2(tab.o$max,1))
x.max <- floor(max(tab.o$max))
x.max.5 <- x.max/2
tab.o$x <- x.max
tab.o$x[tab.o$Outcome=="No"] <- x.max.5


tab.o$y <-1
tab.o$y[tab.o$xyear=="2-year"] <- 2
tab.o$y[tab.o$xyear=="5-year"] <- 1
tab.o$y[tab.o$xyear=="10-year"] <- .5

png("plot/sF distribution of risk.png",width=2500,height =2000,res=300)
ggplot(data=dataplot.o,aes(x=Value,color=Outcome,linetype=Outcome)) +
  geom_density(adjust=2) +
  scale_linetype_manual(breaks=c("Incident","No"),values=c("solid","dashed")) +
  scale_color_manual(breaks=c("Incident","No"),values=c("black","azure4")) +
  geom_text(data=tab.o,aes(x=x,y=y,color=Outcome,label=labels),
            vjust=1,hjust=1,size=3.5,show.legend = F) +
  facet_grid(xyear~Measure,scales="free_y") +
  labs(x="x-year absolute risk",y="Density",color="Breast cancer events",linetype="Br
east cancer events") +
  theme_bw() +
  theme(legend.position = "none")
dev.off()
```

Changing weights

```
fig.lab.list <- list("interval_02"="A. 2-year absolute risk",
                     "interval_05"="B. 5-year absolute risk",
                     "interval_10"="C. 10-year absolute risk")

#-------------------------------------------------------------------------------#

youden.stat <- function(OUTCOME,RESPONSE,MODEL="Not defined"){
  fit.roc <- roc(OUTCOME~RESPONSE)

  youdenJ.index <- which.max(fit.roc$sensitivities + fit.roc$specificities - 1)
  Threshold <- fit.roc$thresholds[youdenJ.index]
  sensitivity <- fit.roc$sensitivities[youdenJ.index]
  specificity <- fit.roc$specificities[youdenJ.index]
  youdenJ <- sensitivity + specificity -1
  output.youden.statistics <- cbind.data.frame(Threshold,sensitivity,specificity,youd
enJ)

  response. <- ifelse(RESPONSE>Threshold,1,0)
  new_data <- cbind.data.frame(response.,OUTCOME)
  temp.tab2 <- tab2(VAR="response.",GROUP="OUTCOME",DATA=new_data)
  temp.tab2[1,1] <- MODEL
  temp.tab2[2,1] <- paste0("YoudenJ threshold: >",form.2(Threshold*100,dp=1),"%")

  temp <- ci(fit.roc)
  AUC <- temp[2]
  AUC.lb <- temp[1]
  AUC.ub <- temp[3]

  temp <- table(response.,OUTCOME)
  FPR = temp["1","No"] / sum(temp["1",]) #false positive rate
  FNR = temp["0","Incident"] / sum(temp["0",]) #false negative rate
  TPR = temp["1","Incident"] / sum(temp["1",]) #true positive rate
  TNR = temp["0","No"] / sum(temp["0",]) # true negative rate


  output.AUC <- cbind.data.frame(AUC,AUC.lb,AUC.ub)
  output.rates <- cbind.data.frame(FPR,FNR,TPR,TNR)

  return(list(tab.cartisen=temp.tab2,
              youden.statistics=output.youden.statistics,
              AUC=output.AUC,
              rates=output.rates))
}

model.glm <- function(DATA,outcome.col.id=1,STEPWISE=F,DIRECTION="backward",DP=2){

  colnames(DATA)[outcome.col.id] <- "Outcome"
  outcome <- DATA$Outcome
  n.all <- nrow(DATA)
  n.cases <- length(which(outcome=="Incident"))
  n.no <- length(which(outcome!="Incident"))

  tab.cartisen = fit = output.model = list()

  fit[["full"]] <- glm(Outcome ~ .,data = DATA, family="binomial")
```

```
  if(STEPWISE==T){
    fit.step0 <- glm(Outcome ~ 1,data = DATA, family="binomial")
    fit[["full"]] <- glm(Outcome ~ .,data = DATA, family="binomial")
    if(DIRECTION == "forward"){
      fit[["best"]] <- step(fit.step0,direction=DIRECTION,scope=formula(fit[["full"
]]))
    }
    if(DIRECTION != "forward"){
      fit[["best"]] <- step(fit[["full"]],direction=DIRECTION)
    }
  }
  for(i in names(fit)){
    response <- predict(fit[[i]],type = "response",newdata = DATA)

    temp <- fit[[i]]$coefficients
    temp <- cbind(form.2(temp,dp=DP),names(temp))
    temp[1,2] <- ""
    temp = paste0(temp[,1]," ",temp[,2])
    temp <- paste0(temp,collapse = " + ")
    temp <- paste0("Model: ",temp)
    output.model = model <- str_replace_all(temp,pattern="  "," ")

    tab.cartisen[[i]] <- youden.stat(OUTCOME = outcome, RESPONSE = response, MODEL=mo
del)
  }
  output <- list(output.cartisen = tab.cartisen,output.model=output.model)
  return(output)
}


format.model.data <- function(TABLE,DP=1){

  col.list <- c("Threshold","sensitivity","specificity","youdenJ",
                "AUC","AUC.lb","AUC.ub",
                "FPR","FNR","TPR","TNR")
  col.format <- colnames(TABLE)[colnames(TABLE)%in%col.list]
  for(COL in col.format){
    TABLE[,COL] <- form.2(as.numeric(TABLE[,COL])*100,DP)
  }

  return(TABLE)
}
```

```
# All ethnicities
output <- list()
for(TYPE in type.order.list){
  YEAR=year.list[[TYPE]]
  DATA = dataset0[(dataset0$age.recruitment+ YEAR) <= 80,
                  c(paste0("bca.",TYPE),paste0("prs.",TYPE),paste0("gail.",TYPE),"FH.
BCa","ptv9")]
  colnames(DATA) <- c("Outcome","PRS","GAIL","FH","LoF")
  output[[TYPE]] <- model.glm(DATA=DATA)
}

out.tab.cartisen = out.model.statistics= NULL
for(TYPE in type.order.list){
  Type <- c(TYPE,"")
  temp <- cbind.data.frame(Type,output[[TYPE]]$output.cartisen$full$tab.cartisen)
  out.tab.cartisen <- rbind(out.tab.cartisen,temp,"")
  prs.threshold = colnames(dataset0)[str_starts(colnames(dataset0),"high") & str_dete
ct(colnames(dataset0),paste0("prs.",TYPE))]
  prs.threshold = str_split_fixed(prs.threshold,"_",3)[2]

  gail.threshold = colnames(dataset0)[str_starts(colnames(dataset0),"high") & str_det
ect(colnames(dataset0),paste0("gail.",TYPE))]
  gail.threshold = str_split_fixed(gail.threshold,"_",3)[2]

  temp2 <- cbind.data.frame(TYPE,prs.threshold,gail.threshold,
                            str_replace(output[[TYPE]]$output.model,"Model: ","Logit
(p): "),
                            output[[TYPE]]$output.cartisen$full$youden.statistics,
                            output[[TYPE]]$output.cartisen$full$AUC,
                            output[[TYPE]]$output.cartisen$full$rates)
  out.model.statistics <- rbind(out.model.statistics,temp2)

}
colnames(out.model.statistics)[str_detect(colnames(out.model.statistics),"model")] <-
"Model"

format.model.data(TABLE=out.model.statistics,DP=1)

write.table(out.tab.cartisen,paste0("output/Models",Sys.Date(),".csv"),row.names=F,se
p=",",quote=F)
write.table(format.model.data(TABLE=out.model.statistics,DP=1),paste0("output/Models
 statistics",Sys.Date(),".csv"),row.names=F,sep=",",quote=F)

## Same variables as uniform weights
output <- list()
for(TYPE in type.order.list){
  YEAR=year.list[[TYPE]]
  DATA = dataset0[(dataset0$age.recruitment+ YEAR) <= 80,
                  c(paste0("bca.",TYPE),paste0("prs.",TYPE),"FH.BCa","ptv9")]
  colnames(DATA) <- c("Outcome","PRS","FH","LoF")
  output[[TYPE]] <- model.glm(DATA=DATA)
}

out.tab.cartisen = out.model.statistics= NULL
for(TYPE in type.order.list){
```

```
  Type <- c(TYPE,"")
  temp <- cbind.data.frame(Type,output[[TYPE]]$output.cartisen$full$tab.cartisen)
  out.tab.cartisen <- rbind(out.tab.cartisen,temp,"")
  prs.threshold = colnames(dataset0)[str_starts(colnames(dataset0),"high") & str_dete
ct(colnames(dataset0),paste0("prs.",TYPE))]
  prs.threshold = str_split_fixed(prs.threshold,"_",3)[2]

  gail.threshold = ""

  temp2 <- cbind.data.frame(TYPE,prs.threshold,gail.threshold,
                            str_replace(output[[TYPE]]$output.model,"Model: ","Logit
(p): "),
                            output[[TYPE]]$output.cartisen$full$youden.statistics,
                            output[[TYPE]]$output.cartisen$full$AUC,
                            output[[TYPE]]$output.cartisen$full$rates)
  out.model.statistics <- rbind(out.model.statistics,temp2)

}
colnames(out.model.statistics)[str_detect(colnames(out.model.statistics),"model")] <-
"Model"
format.model.data(TABLE=out.model.statistics,DP=1)

write.table(out.tab.cartisen,paste0("output/Models",Sys.Date(),".csv"),row.names=F,se
p=",",append=T,quote=F)
write.table(format.model.data(TABLE=out.model.statistics,DP=1),paste0("output/Models
 statistics",Sys.Date(),".csv"),row.names=F,sep=",",append=T,quote=F)

temp <- read.csv("output/Supporting data 3-F Getting best combination.csv")
temp <- temp[!is.na(temp$KEY..),]
temp$Model <- str_replace_all(temp$KEY,"\u00B7"," + ")
temp.out <- temp[,intersect(colnames(out.model.statistics),colnames(temp))]

write.table(format.model.data(TABLE=temp.out,DP=1),paste0("output/T2 Models  statisti
cs_union",Sys.Date(),".csv"),row.names=F,sep=",",quote=F)
```

REVISION

Survival analysis

```
library(survival)
library(survAUC)
library(survminer)
library(pec)

TYPE = "lifetime_80"
data.risk <- read.csv("data/Risk_bestThreshold_AUCcontinuous_n246142_2023-05-08.csv")
dataset0 <- data[data$bca.lifetime_80!="Prevalent",]
dataset0 <- merge(dataset0,data.risk,by="FID")

summary(dataset0$followup.time)
table(dataset0$bca.incidence)
length(which(dataset0$followup.time==0))
summary(dataset0$followup.time[dataset0$bca.incidence!="No"])

type.list = c("interval_02","interval_05","interval_10",
              "lifetime_80","lifetime_70","lifetime_60")

type.order.list <- c(paste0("interval_", c("02","05","10")),"lifetime_80")

year.list <- list("interval_02"=2,"interval_05"=5,"interval_10"=10,"lifetime_80"=0)

out.list = NULL
for(TYPE in type.order.list[1:3]){
  temp.plot = NULL

  for(AR in c("prs","gail")){
    YEAR = year.list[[TYPE]]
    temp.data <- dataset0[(dataset0$age.recruitment+ YEAR) <= 80,]
    temp.data$event = ifelse(temp.data[,paste0("bca.",TYPE)]=="No",0,1)
    temp.data$event[temp.data$followup.time>YEAR] <- 0
    temp.data$followup.time[temp.data$followup.time>YEAR] <- YEAR

    temp <- colnames(temp.data)[str_detect(colnames(temp.data),AR)]
    temp <- temp[str_detect(temp,TYPE)]
    temp <- temp[str_detect(temp,"high")]

    temp.plot.0 <- cbind(temp.data[,c("event","followup.time",temp)],toupper(AR),
                         paste0(as.numeric(str_replace(TYPE,"interval_","")),"-year a
bsolute risk"))
    colnames(temp.plot.0) <- c("event","followup.time","group","Tool","Period")
    temp.plot = rbind.data.frame(temp.plot,temp.plot.0)

  }
  fit <- survfit(formula(paste0("Surv(followup.time,event)~group")),data=temp.plot)
  png(paste0("plot/survival/Survival_",TYPE,".png"),width=1500,height = 750,res=180)
  print(ggsurvplot_facet(fit,data=temp.plot, facet.by = c("Tool"),
                   palette = c("grey","black"), pval = F, ylim=c(.9,1),legend.labs=c(
"Low", "High"),
                  legend.title="",
                   ylab="Proportion of women \nnot diagnosed with breast cancer")
  )
  dev.off()

  cox <- coxph(formula(paste0("Surv(followup.time,event)~",temp)),data=temp.data,x=T,
```

```
y=T,method="breslow")
  out.list <- rbind(out.list,cbind.data.frame(temp,summary(cox)$coef))
}

temp <- out.list
tmp <- temp[,1]
tmp <- str_replace(tmp,"high_","")
tmp <- str_replace(tmp,"\\.interval","")
tmp <- str_replace(tmp,"\\.lifetime","")

tmp1 <- temp[,"se(coef)"]
LB = exp(temp$coef - qnorm(.975)*tmp1)
UB = exp(temp$coef + qnorm(.975)*tmp1)
HR.CI <- paste0(form.2(temp[,"exp(coef)"],1)," (",form.2(LB,1)," to ",form.2(UB,1),
")")

output <- cbind.data.frame(str_split_fixed(tmp,"_",3),HR.CI,form.e(temp$`Pr(>|z|)`,2
))
write.csv(output,"output/survival/hazards ratio.csv",row.names = F)
```